CONSISTENT ESTIMATION OF THE ORDER FOR HIDDEN MARKOV MODELS

Berlian Setiawaty

Thesis submitted for the degree of
Doctor of Philosophy
in the Department of Applied Mathematics
The University of Adelaide

July 2, 1999

Contents

Abstract

The order of a hidden Markov model cannot be estimated using a classical maximum likelihood method, since increasing the size parameter will increase the likelihood.

In this thesis, a maximum compensated log-likelihood method is proposed for estimating the order of general hidden Markov models. This method is based on the compensation of the log-likelihood function. A compensator, which is decreasing in size parameter K, is added to the maximum log-likelihood, and the resulting compensated log-likelihood is maximized with respect to K. The problem is then to find a proper compensator, which allows the strongly consistent estimation of the order.

Following the method of Baras and Finesso [3], sufficient conditions for compensators avoiding under estimation and compensators avoiding over estimation are obtained.

Sufficient condition on the compensator avoiding under estimation requires consistent estimation of parameters of general hidden Markov models which is obtained by generalizing [34].

Sufficient conditions on the compensator avoiding over estimation require precise information on the rate of growth of the maximized log-likelihood ratio

and the upper bound to the rate. The rate and the upper bound are obtained using Csiszar lemma [13], from information theory.

To conclude the thesis, an example of a compensator which generate a strongly consistent estimator of the order is given.

Signed Statement

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

Adelaide, July 2, 1999

Berlian Setiawaty

Acknowledgements

I wish to express my sincerest appreciation to my supervisor, Dr. John van der Hoek, for his continued guidance and support during the course of this research. His direction and encouragement were crucial and valuable to the completion of this thesis.

I would also like to thank the Australian Agency for International Development (AusAID) for providing the scholarship, and Bogor Agricultural University for giving me the opportunity to pursue my PhD.

In addition, I am thankful to Dr. Peter Gill, the head of the Department of Applied Mathematics, for his support and encouragement.

Finally, I wish to express my gratitude to my parents and my parents in-law, for their continued encouragement. Special thanks go to my husband Djarot and my son Aldy, without their support, patient and understanding, the completion of this thesis would not have been possible.

List of Notation

$\boldsymbol{N}$	set of positive integers
$\boldsymbol{R}$	set of real numbers
$\boldsymbol{Z}$	set of integers
$\boldsymbol{R}^n$	n-dimensional Euclidean space
$\boldsymbol{R}^\infty$	set of real sequences
$\mathcal{R}^n$	Borel σ-field of $\boldsymbol{R}^n$
$\mathcal{R}^\infty$	Borel σ-field of $\boldsymbol{R}^\infty$
$A_{K,L}$	$K \times L$ real matrix
$0_{K,L}$	$K \times L$ zero matrix
I_K	$K \times K$ identity matrix
$\bar{A}$	closure of A
A^*	complement of A
$\emptyset$	empty set

Chapter 1

Introduction

A *hidden Markov model* (HMM) is a discrete time stochastic process $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ such that $\{X_t\}$ is a finite state *Markov chain*, and given $\{X_t\}$, $\{Y_t\}$ is a sequence of conditionally independent random variables, with the conditional distribution of Y_n depends on $\{X_t\}$ only through X_n. The name hidden Markov model comes from the assumption that the Markov chain $\{X_t\}$ is *not observable*, that is, *hidden* in the observations of $\{Y_t\}$.

The states of a hidden Markov model could be associated with *regimes*. In econometrics and finance, hidden Markov models are also known as Markov switching models or regime switching models.

Hidden Markov models have during the last decade become widespread for modelling sequences of dependent random variables with application in areas, such as: speech processing [40], [41], [42], [43]; biology [35]; finance [19]; and econometrics [26], [27], [28], [29], [30].

Theoretical work on hidden Markov models was initially started in 1950s. The

first contribution was made by Blackwell and Kopman [12] who studied a special class of hidden Markov models, when the observed process takes values in a finite set. This class is referred to as a probabilistic function of a Markov chain. Since then, a long line of research has been undertaken which has enriched the theory of hidden Markov models.

Inference for hidden Markov model was first considered by Baum and Petrie [5] who also treated the case when the observed process $\{Y_t\}$ takes values in a finite set. In [5], results on consistency and asymptotic normality of the maximum likelihood estimate are given. Petrie [38] weakened the conditions for consistency in [5]. In [38], the identifiability problem is also discussed: under what conditions are there no other parameters that induce the same law for the observed process $\{Y_t\}$ as the true parameter, with exception for permutation of states.

For general hidden Markov models with Y_n conditioned on X_n having density $f(\cdot, \theta_{X_n})$, Leroux [34] proved the consistency of the maximum likelihood estimate under mild conditions. An interesting fact to be noticed in [34], is that the consistency of the maximum likelihood estimate does not depend on the initial probability distributions. Assymptotic normality of the maximum likelihood estimator for general hidden Markov models has been proved by Bickel et. al. [10].

Estimation of the parameters of a hidden Markov model has most often been performed using maximum likelihood estimation. Baum and Eagon [4] gave an algorithm for locating a local maximum of the likelihood function for a probabilistic function of a Markov chain. In 1970, Baum et. al. [6] developed the *expectation maximization* (EM) algorithm and applied it to general hidden Markov models. Dempster, Laird and Rubin [14] further developed the EM algorithm and made it popular for applications. Examples of the application

of the EM algorithm to speech processing can be found in [40] and [41].

Recent work on parameter estimation of hidden Markov models includes those of Ryden [47] and Elliott et.al. [18]. Ryden in [47] considered a recursive estimator for general hidden Markov models based on the m-dimensional distribution of the observation process and proved that this estimator converges to the set of stationary points of the corresponding Kullback-Leibler information. Reference [18] contains extensive work in estimation and control for a wide range of hidden Markov models. In [18], the *reference probability* method which involves a *probability measure change* and the expectation maximization algorithm was used to produce recursive estimator for parameter of hidden Markov models. This reference also includes an extensive bibliography.

Although much work has been dedicated to parameter estimation for hidden Markov models, only recently has the *order estimation* problem received some attention. To estimate the order of a hidden Markov model, the classical maximum likelihood cannot be used, since increasing the size parameter will automatically increase the likelihood. This is the typical behaviour of the likelihood function when the parameter is *structural*, that is, the parameter (usually integer valued) indexes the complexity of the model.

So far, the only technique that has been used to estimate the order of hidden Markov models is the *compensated likelihood* estimation. This technique is based on a compensation of the likelihood function. A *compensator*, decreasing in size parameter K, is added to the maximum likelihood and the resulting compensated likelihood is maximized with respect to K. Proper choice of the compensator allows the strongly consistent estimation of the order.

The first contribution to the order estimation along these lines was made in 1991 by Baras and Finesso [3]. Using the compensated likelihood technique,

they proved the consistent estimation of the order for hidden Markov models in which the observation process takes values on a finite set.

The second and also the last contribution was given by Ryden [46] in 1995. In [46], the compensated likelihood was also used to estimate the order of general hidden Markov models. Ryden proved that in the limit, the estimator which is based on m-dimensional distribution does not under estimate the order.

Inspiring by the work of Finesso and Baras, this thesis is dedicated to solve the problem of order estimation for general hidden Markov models by adapting the procedure and techniques used in [3] and [34]. Under weaker conditions than in [46], we will show that in the limit the estimator does not under estimate, but also does not over estimate the order.

We conclude the introduction with a brief summary of the thesis. Chapter 1 contains literature research, the aim and a brief summary of the thesis. Chapter 2 presents definitions, notations and basic results concerning general hidden Markov models. In Chapter 3, the identifiability of general hidden Markov models is derived from the identifiability of finite mixtures. Chapter 4 proves the consistent estimation of parameters of general hidden Markov models, which generalizes the result of Leroux [34]. Finally, in Chapter 5, using the compensated likelihood, we prove the consistent estimation of the order for general hidden Markov models. The results of this chapter can be seen as the extension of the results of Baras and Finesso [3], which hold for hidden Markov models in which the observed process takes values in a finite set, to general hidden Markov models.

Chapter 2

Hidden Markov Model Fundamentals

The purpose of this chapter is to introduce hidden Markov models and to present some definitions and basic results that will be used in the sequel.

In the first section, some definitions and standard properties of Markov chains are presented. Even though, most of these definitions can be found in many places, such as [25], [32] and [8], this section is necessary for completeness and to make the thesis self-contained.

A hidden Markov model is formally defined in section 2.2 and an example is given. In section 2.3, the nature of dependencies between the random variables in a hidden Markov model is discussed. Using the results of section 2.3, the finite dimensional joint distributions of the observed process are derived. So the parameters which characterize a hidden Markov model can be analysed in section 2.4. Such parameters will be referred to as a *representation* of the model. Based on the laws of the observation processes, an equivalence relation

for representations of hidden Markov models is defined in section 2.5. For a hidden Markov model, our main interest is to identify the *simplest* representation which is equivalent to the model's representation. Such representation will be called a *true parameter* of the hidden Markov model. Section 2.6 presents the characteristics of this true parameter.

In the last section, a (strictly) stationary hidden Markov model is discussed. Here, we build a past history and give sufficient conditions for the ergodicity of the observed process.

2.1 Markov Chains

Let $\{X_t : t \in \boldsymbol{N}\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values in a finite set $\boldsymbol{S} = \{1, \ldots, K\}$. $\{X_t\}$ is said to be a *Markov chain* if it satisfies

$$P(X_{n+1} = i_{n+1} | X_1 = i_1, \ldots, X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n), \qquad (2.1)$$

for all $i_1, \ldots, i_{n+1} \in \boldsymbol{S}$ and $n \in \boldsymbol{N}$. Property (2.1) is called the *Markov property.*

Let $m \leq n$, then by (2.1),

$$\begin{aligned}
&P(X_{n+1} = i_{n+1} | X_1 = i_1, \ldots, X_m = i_m) \\
&= \sum_{i_{m+1}=1}^{K} \cdots \sum_{i_n=1}^{K} P(X_{m+1} = i_{m+1}, \ldots, X_{n+1} = i_{n+1} | X_1 = i_1, \ldots, X_m = i_m) \\
&= \sum_{i_{m+1}=1}^{K} \cdots \sum_{i_n=1}^{K} \Big\{ P(X_{m+1} = i_{m+1} | X_1 = i_1, \ldots, X_m = i_m) \\
&\qquad \times P(X_{m+2} = i_{m+2} | X_1 = i_1, \ldots, X_{m+1} = i_{m+1}) \\
&\qquad \times \cdots \times P(X_{n+1} = i_{n+1} | X_1 = i_1, \ldots, X_n = i_n) \Big\} \\
&= \sum_{i_{m+1}=1}^{K} \cdots \sum_{i_n=1}^{K} \Big\{ P(X_{m+1} = i_{m+1} | X_m = i_m) \cdot P(X_{m+2} = i_{m+2} | X_{m+1} = i_{m+1})
\end{aligned}$$

$$\times \cdots \times P(X_{n+1} = i_{n+1} | X_n = i_n) \Big\}$$

$$= \sum_{i_{m+1}=1}^{K} \cdots \sum_{i_n=1}^{K} P(X_{m+1} = i_{m+1}, \ldots, X_{n+1} = i_{n+1} | X_m = i_m)$$

$$= P(X_{n+1} = i_{n+1} | X_m = i_m). \tag{2.2}$$

So the Markov property (2.1) is equivalent with (2.2).

Assume that $P(X_{n+1} = j | X_n = i)$ depends only on (i, j) and not on n. Let

$$\alpha_{ij} = P(X_{n+1} = j | X_n = i), \qquad i, j = 1, 2, \ldots, K, \tag{2.3}$$

then α_{ij} are called the *transition probabilities* from state i to state j and the $K \times K$ matrix A defined by

$$A = (\alpha_{ij}), \tag{2.4}$$

is called the *transition probability matrix* of the Markov chain $\{X_t\}$. Notice that A satisfies

$$0 \leq \alpha_{ij} \leq 1, \qquad i, j = 1, \ldots, K$$

$$\textstyle\sum_{j=1}^{K} \alpha_{ij} = 1, \qquad i = 1, \ldots, K.$$

Thus A is a *stochastic matrix.*

Let

$$\pi_i = P(X_1 = i), \qquad i = 1, \ldots, K \tag{2.5}$$

and

$$\pi = (\pi_i). \tag{2.6}$$

The $1 \times K$-matrix π is called the *initial probability distribution* of the Markov chain $\{X_t\}$. Notice that π satisfies

$$0 \leq \pi_i \leq 1, \qquad i = 1, \ldots, K \qquad \text{and} \qquad \sum_{i=1}^{K} \pi_i = 1.$$

By (2.1), (2.3), (2.4), (2.5) and (2.6),

$$P(X_1 = i_1, \ldots, X_n = i_n) \;=\; P(X_1 = i_1) \cdot P(X_2 = i_2 | X_1 = i_1)$$

$$\times \cdots \times P(X_n = i_n | X_{n-1} = i_{n-1})$$

$$= \pi_{i_1} \cdot \alpha_{i_1,i_2} \cdots \alpha_{i_{n-1},i_n}. \tag{2.7}$$

Then by (2.7),

$$\begin{aligned} P(X_n = i) &= \sum_{i_1=1}^{K} \cdots \sum_{i_{n-1}=1}^{K} P(X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n) \\ &= \sum_{i_1=1}^{K} \cdots \sum_{i_{n-1}=1}^{K} \pi_{i_1} \cdot \alpha_{i_1,i_2} \cdots \alpha_{i_{n-1},i} \\ &= \pi A^{n-1} e_i, \end{aligned} \tag{2.8}$$

where $A^n = AA \cdots A$ and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$. Hence, it can be concluded that the probability distribution of the Markov chain $\{X_t\}$ is completely determined by the initial probability π and the transition probability matrix A.

If the initial probability distribution π satisfies

$$\pi A = \pi, \tag{2.9}$$

then π is called a *stationary probability distribution* with respect to A. By (2.8) and (2.9), for every $n \in \boldsymbol{N}$,

$$\begin{aligned} P(X_n = i) &= \pi A^{n-1} e_i \\ &= \pi e_i \\ &= \pi_i, \end{aligned}$$

implying

$$\begin{aligned} P(X_{m+1} = i_1, \ldots, X_{m+n} = i_n) &= P(X_{m+1} = i_1) \cdot P(X_{m+2} = i_2 | X_{m+1} = i_1) \\ &\quad \times \cdots \times P(X_{m+n} = i_n | X_{m+n-1} = i_{n-1}) \\ &= \pi_{i_1} \cdot \alpha_{i_1,i_2} \cdots \alpha_{i_{n-1},i_n} \\ &= P(X_1 = i_1, \ldots, X_n = i_n), \end{aligned}$$

for $m \in \boldsymbol{N}$ and $i_1, \ldots, i_n \in \boldsymbol{S}$. So in this case, the Markov chain $\{X_t\}$ is *(strictly) stationary.*

To classify the states of the Markov chain $\{X_t\}$, define a *communication* relation "$\leftrightarrow$" as follows. A state j is said to be *accessible* or reachable from a state i, denoted as $i \rightarrow j$, if there is an integer n, $0 \leq n < K$, such that the (i, j) entry of A^n is positive. If $i \rightarrow j$ and $j \rightarrow i$, then i and j are said to *communicate* with each other, denoted as $i \leftrightarrow j$.

For each state i, define a *communicating class*

$$C(i) = \{j \in \boldsymbol{S} : i \leftrightarrow j\}.$$

Since relation $\leftrightarrow$ is an equivalence relation, then the communicating classes satisfy :

(a). For every state i, $i \in C(i)$.

(b). If $j \in C(i)$, then $i \in C(j)$.

(c). For any state i and j, either $C(i) = C(j)$ or $C(i) \cap C(j) = \emptyset$

Thus the state space $\boldsymbol{S}$ can be *partitioned* into these classes.

A Markov chain is said to be *irreducible*, if all states communicate with each other. So in this case, the Markov chain has only one communicating class.

A communicating class C is called *ergodic* if

$$\sum_{j \in C} \alpha_{ij} = 1, \qquad \forall i \in C. \tag{2.10}$$

The individual states in an ergodic class are also called *ergodic.*

A communicating class C is called *transient*, if there is $i \in C$, such that

$$\sum_{j \in C} \alpha_{ij} < 1. \tag{2.11}$$

The individual states in a transient class are also called *transient.*

To identify the transition probability matrix within a communicating class, the irreducibility of square matrix is introduced. An $n \times n$-matrix $B = (\beta_{ij})$ is said to be *irreducible*, if there is a permutation of indices σ, such that the matrix $\tilde{B} = (\tilde{\beta}_{ij})$, with $\tilde{\beta}_{ij} = \beta_{\sigma(i),\sigma(j)}$, has form

$$\tilde{B} = \begin{pmatrix} C & 0 \\ D & E \end{pmatrix}$$

where C and E are $l \times l$ and $m \times m$ matrices respectively, and $l + m = n$.

Let C_e be an ergodic class and n_e be the number of ergodic states in C_e. Let A_e be the $n_e \times n_e$ transition probability matrix within C_e. Then by (2.10) A_e is a *stochastic* matrix. Moreover, A_e is *irreducible*, since if A_e is *reducible*, then by some permutation σ, A_e can be reduced to the form

$$\tilde{A}_e = \begin{pmatrix} B_e & 0 \\ C_e & D_e \end{pmatrix}, \tag{2.12}$$

where B_e and D_e are $k_e \times k_e$ and $l_e \times l_e$ matrices respectively, with $k_e + l_e = n_e$. But from (2.12), it can be seen that every state in $\{\sigma(1), \ldots, \sigma(k_e)\}$ does not communicate with every state in $\{\sigma(k_e+1), \ldots, \sigma(n_e)\}$, contradicting with the fact that C_e is a communicating class. Therefore, A_e must not be reducible.

Let C_t be a transient class and n_t be the number of transient states in C_t. Let A_t be the $n_t \times n_t$ transition probability matrix within C_t. Then by (2.11), A_t is a *substochastic* matrix, that is, its individual row sums are ≤ 1.

The next lemma shows the relation between irreducible Markov chains and irreducible transition probability matrices.

Lemma 2.1.1 *Let $\{X_t\}$ be a Markov chain with a $K \times K$ transition probability matrix A. Then $\{X_t\}$ is irreducible if and only if A is irreducible.*

Proof :

Let $\{X_t\}$ be a Markov chain with a $K \times K$ transition probability matrix A. If $\{X_t\}$ is irreducible, then it consists of a single communicating class C and the transition probablity matrix within C is A. Since A is a stochastic matrix, then C is an ergodic class. From the ergodicity of C, the irreducibility of A follows.

On the otherhand, if A is irreducible, then from [24], page 63, for every $1 \leq i, j \leq K$, there is an integer n, $0 \leq n \leq K$, such that the (i, j) entry of A^n is positive. This means that every state communicates with each other. So the chain $\{X_t\}$ is irreducible. ■

Let $\{X_t\}$ be a Markov chain with a $K \times K$ transition probability matrix A. Let K_e and K_t be the number of ergodic states and transient states respectively. In general, after a suitable permutation of indices, the transition probability matrix A can be written in the block form as

$$\tilde{A} = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B is a $K_e \times K_e$-stochastic matrix and D is a $K_t \times K_t$-substochastic matrix.

The block D describes the transient $\rightarrow$ transient movements in the chain. For each class of transient states, at least one row in D will have sum < 1.

The $K_t \times K_e$-block C describes the transient $\rightarrow$ ergodic movements in the chain. For each class of transient states, at least one row in C will have a non-zero entry.

Finally, The $K_e \times K_e$-block B describes the movements within each ergodic class in the chain. Suppose that the chain has e ergodic classes. Since it is

impossible to leave an ergodic class, B has the form,

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_e \end{pmatrix},$$

where B_i is the transition matrix within the i-th ergodic class. For each i, B_i is an irreducible stochastic matrix.

The following lemma shows the relation between the communicating classes and the stationary probability distributions.

Lemma 2.1.2 *Let $\{X_t\}$ be a Markov chain with a $K \times K$ transition probability matrix A. Let*

$$\begin{aligned} S &= \{\pi = (\pi_i) : \pi_i \geq 0,\ i = 1, \ldots, K,\ \sum_{i=1}^{K} \pi_i = 1,\ \pi A = \pi\} \\ S^+ &= \{\pi \in S : \pi_i > 0,\ i = 1, \ldots, K\} \\ S^o &= S - S^+. \end{aligned}$$

(a). If $\{X_t\}$ is irreducible, then $S = S^+$ and $S^o = \emptyset$.

(b). If $\{X_t\}$ has e communicating classes, with $2 \leq e \leq K$ which all are ergodic, then $S^+ \neq \emptyset$ and $S^o \neq \emptyset$.

(c). If $\{X_t\}$ has k communicating classes, with $2 \leq k \leq K$, where e of them are ergodic, $1 \leq e < k$, and t of them are transient, $e + t = k$, then $S = S^o$ and $S^+ = \emptyset$.

Proof :

To prove (a), let $\{X_t\}$ be an irreducible Markov chain. Suppose there is $\pi \in S^o$.

Let k be the number of non-zero π_i. Without loss of generality, suppose that

$$\begin{aligned} \pi_i &> 0, \qquad \text{for} \quad i = 1, \ldots, k \\ \pi_i &= 0, \qquad \text{for} \quad i = k+1, \ldots, K. \end{aligned}$$

As $\pi A = \pi$, then

$$\alpha_{ij} = 0, \qquad \text{for} \quad i = 1, \ldots, K, \qquad j = k+1, \ldots, K.$$

Thus A has form

$$A = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B is a $k \times k$-matrix and D is a $(K-k) \times (K-k)$-matrix. So A is reducible, contradicting with the fact that A is irreducible by Lemma 2.1.1. Therefore, it must be $S^o = \emptyset$ and hence $S = S^+$.

To prove (b), let $\{X_t\}$ be a Markov chain having e communicating classes with $2 \leq e \leq K$, which all are ergodic. Then without loss of generality, A is of the form

$$A = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_e \end{pmatrix},$$

where B_i is the transition matrix within the i-th ergodic class and it is an irreducible stochastic matrix.

Let π^i be an $1 \times e_i$-matrix, where e_i is the number of ergodic states in B_i, such that for $i = 1, \ldots, e$,

$$\pi^i_j \geq 0, \qquad j = 1, \ldots, e_i, \qquad \text{and} \qquad \sum_{j=1}^{e_i} \pi^i_j = 1$$

and

$$\pi^i B_i = \pi^i.$$

By (a),

$$\pi_j^i > 0, \qquad \text{for} \quad j = 1, \ldots, e_i \qquad \text{and} \quad i = 1, \ldots, e.$$

Let

$$\hat{\pi} = (\pi^1, 0, \ldots, 0),$$

then

$$\begin{aligned} \hat{\pi}A &= (\pi^1 B_1, 0, \ldots, 0) \\ &= (\pi^1, 0, \ldots, 0) \\ &= \hat{\pi}. \end{aligned}$$

So $\hat{\pi} \in S^o$ and hence $S^o \neq \emptyset$.

Let a_i, $i = 1, \ldots, e$ be any real numbers such that

$$a_i > 0, \qquad i = 1, \ldots, e \qquad \text{and} \qquad \sum_{i=1}^{e} a_i = 1.$$

Let

$$\tilde{\pi} = (a_1\pi^1, a_2\pi^2, \ldots, a_e\pi^e),$$

then

$$\tilde{\pi}_i > 0, \qquad \text{for} \quad i = 1, \ldots, K$$

and

$$\begin{aligned} \sum_{i=1}^{K} \tilde{\pi}_i &= \sum_{i=1}^{e} \sum_{j=1}^{e_i} a_i \pi_j^i \\ &= \sum_{i=1}^{e} a_i \\ &= 1. \end{aligned}$$

Moreover,

$$\begin{aligned} \tilde{\pi}A &= (a_1\pi^1 B_1, a_2\pi^2 B_2, \ldots, a_e\pi^e B_e) \\ &= (a_1\pi^1, a_2\pi^2, \ldots, a_e\pi^e) \\ &= \tilde{\pi}, \end{aligned}$$

then $\tilde{\pi} \in S^+$ and hence $S^+ \neq \emptyset$.

To prove (c), let $\{X_t\}$ be a Markov chain having k communicating classes, $2 \leq k \leq K$, where e of them are ergodic, $1 \leq e < k$, and t of them are transient, $e + t = k$. Let K_e and K_t be the number of ergodic states and transient states of $\{X_t\}$ respectively. Without loss of generality, assume that the matrix transition A is of the form

$$A = \begin{pmatrix} B & 0 \\ C & D \end{pmatrix},$$

where B is a $K_e \times K_e$ stochastic matrix, D is a $K_t \times K_t$ substochastic matrix and C is a $K_t \times K_e$ matrix, $C \neq 0$.

Let $\pi = (\pi_1, \ldots, \pi_K) \in S$, since

$$\pi A = \pi,$$

then

$$\begin{aligned} A^T \pi^T &= \pi^T \\ (A^T - I_K)\pi^T &= 0 \end{aligned}$$

or

$$\begin{pmatrix} B^T - I_{K_e} & C^T \\ 0 & D^T - I_{K_t} \end{pmatrix} \begin{pmatrix} \pi^1 \\ \pi^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{2.13}$$

where $\pi^1 = (\pi_1, \ldots, \pi_{K_e})$ and $\pi^2 = (\pi_{K_e+1}, \ldots, \pi_K)$. By (2.13), π^1 and π^2 satisfy

$$(B^T - I_{K_e})\pi^1 + C^T \pi^2 = 0 \tag{2.14}$$

$$(D^T - I_{K_t})\pi^2 = 0. \tag{2.15}$$

By [32], page 44, $D^T - I_{K_t}$ is invertible, so the only solution for (2.15) is $\pi^2 = 0$. So π must have form $\pi = (\pi^1, 0)$, where π^1 satisfy (2.14). This means that $\pi \in S^o$. Therefore $S \subset S^o$, implying $S = S^o$ and $S^+ = \emptyset$. ■

2.2 Hidden Markov Models

Let $\{X_t : t \in \boldsymbol{N}\}$ be a finite state Markov chain defined on a probability space $(\Omega, \mathcal{F}, P)$. Suppose that $\{X_t\}$ is not observed directly, but rather there is an *observation* process $\{Y_t : t \in \boldsymbol{N}\}$ defined on $(\Omega, \mathcal{F}, P)$. Then consequently, the Markov chain is said to be *hidden* in the observations. A pair of stochastic processes $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ is called a hidden Markov model. Precisely, according to [47], a hidden Markov model is formally defined as follows.

Definition 2.2.1 *A pair of discrete time stochastic processes $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in a set $\boldsymbol{S} \times \mathcal{Y}$, is said to be a* ***hidden Markov model*** *(HMM), if it satisfies the following conditions:*

(a). $\{X_t\}$ is a finite state Markov chain.

(b). Given $\{X_t\}$, $\{Y_t\}$ is a sequence of conditionally independent random variables.

(c). The conditional distribution of Y_n depends on $\{X_t\}$ only through X_n.

Assume that the Markov chain $\{X_t\}$ ***is not observable****. The cardinality K of $\boldsymbol{S}$, will be called the* ***size*** *of the hidden Markov model.*

The following is an example of a hidden Markov model which is adapted from [18].

Example 2.2.2 Let $\{X_t\}$ be a Markov chain defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values on $\boldsymbol{S} = \{1, \ldots, K\}$. The observed process $\{Y_t\}$ is defined by

$$Y_t = c(X_t) + \sigma(X_t)\omega_t, \qquad t \in \boldsymbol{N}, \tag{2.16}$$

where c and σ are real valued functions and positive real valued function on $\boldsymbol{S}$ respectively, and $\{\omega_t\}$ is a sequence of $N(0,1)$ independent, identically distributed (i.i.d.) random variables.

Since $\{\omega_t\}$ is a sequence of $N(0,1)$ i.i.d. random variables, then given $\{X_t\}$, $\{Y_t\}$ is a sequence of independent random variables. From (2.16), it is clear that Y_t is a function of X_t only, then by Definition 2.2.1, $\{(X_t, Y_t)\}$ is a hidden Markov model.

Notice that for $y \in \mathcal{Y}$ and $i \in \boldsymbol{S}$,

$$\begin{aligned} P(Y_t \leq y | X_t = i) &= P(c_i + \sigma_i \omega_t \leq y) \\ &= P(\sigma_i \omega_t \leq y - c_i) \\ &= \int_{-\infty}^{y-c_i} \phi_i(z)\, dz, \end{aligned} \tag{2.17}$$

where $c_i = c(i)$, $\sigma_i = \sigma(i)$ and

$$\phi_i(z) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z}{\sigma_i}\right)^2}. \tag{2.18}$$

From (2.17) and (2.18), the conditional density of Y_t given $X_t = i$ is $\phi_i(\cdot - c_i)$, which does not depend on t.

2.3 Dependencies between Random Variables

This section shows the nature of dependencies between the random variables in a hidden Markov model.

Let $\{(X_t, Y_t)\}$ be a hidden Markov model defined on a probability space $(\Omega, \mathcal{F}, P)$, where the Markov chain $\{X_t\}$ taking values in a set $\boldsymbol{S} = \{1, \ldots, K\}$ and the observed process $\{Y_t\}$ taking values on $\mathcal{Y}$. Throughout this thesis,

we will assume that Y_t is scalar valued and without loss of generality, we will suppose that $\mathcal{Y} = \boldsymbol{R}$. The generalization to vector case is straight forward.

Assume that the conditional density of Y_t given $X_t = i$, for all $t \in \boldsymbol{N}$ and $i = 1, \ldots, K$ do not depend on t, and are dominated by a σ-finite measure μ. The conditional density of Y_t given $X_t = i$, with respect to μ, will be denoted by $p(\cdot|i)$. This means that for all $t \in \boldsymbol{N}$ and $i = 1, \ldots, K$,

$$P(Y_t \leq y | X_t = i) = \int_{-\infty}^{y} p(z|i)\, d\mu(z).$$

Notation 2.3.1 Here and in the sequel, p will be used as a generic symbol for a probability density function. If there is no confusion, for random variables U and V defined on $(\Omega, \mathcal{F}, P)$, the joint density function of U and V, $p_{U,V}(\cdot,\cdot)$ will be denoted by $p(\cdot,\cdot)$ and the conditional density function of U given V, $p_{U|V}(\cdot|\cdot)$ will simply be denoted by $p(\cdot|\cdot)$.

Let U and V be any random variables defined on $(\Omega, \mathcal{F}, P)$. Notice that the joint density function of U and V and the conditional density function of U given V can be expressed as

$$\begin{aligned} p(u,v) &= p(U(\omega), V(\omega)) \\ &= p(U,V)(\omega) \\ p(u|v) &= p(U(\omega)|V(\omega)) \\ &= p(U|V)(\omega), \end{aligned}$$

where $U(\omega) = u$ and $V(\omega) = v$, for some $\omega \in \Omega$.

First we prove some general rules for conditional densities.

Lemma 2.3.2 *Let U, V and W be any random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, then*

(a). $p(U|V,W) = \dfrac{p(U|V)\cdot p(W|U,V)}{p(W|V)}$.

(b). $p(U|V,W) = \dfrac{p(U,V|W)}{p(V|W)}$.

(c). $p(U,V|W) = p(V|W)\cdot p(U|V,W)$.

Proof :

The conditional probability density function of U given V is defined by

$$p(u|v) = \frac{p(u,v)}{p(v)}, \tag{2.19}$$

for all u and for all v such that $p(v) > 0$. By equation (2.19), we have

$$p(U|V) = \frac{p(U,V)}{p(V)}. \tag{2.20}$$

Analog with (2.20),

$$p(W|U,V) = \frac{p(U,V,W)}{p(U,V)} \tag{2.21}$$

$$p(U|V,W) = \frac{p(U,V,W)}{p(V,W)}. \tag{2.22}$$

By equations (2.20), (2.21) and (2.22),

$$\begin{aligned}
\frac{p(U|V)\cdot p(W|U,V)}{p(W|V)} &= \frac{\dfrac{p(U,V)}{p(V)}\cdot\dfrac{p(U,V,W)}{p(U,V)}}{\dfrac{p(V,W)}{p(V)}} \\
&= \frac{p(U,V,W)}{p(V,W)} \\
&= p(U|V,W),
\end{aligned}$$

$$\begin{aligned}
\frac{p(U,V|W)}{p(V|W)} &= \frac{\dfrac{p(U,V,W)}{p(W)}}{\dfrac{p(V,W)}{p(W)}} \\
&= \frac{p(U,V,W)}{p(V,W)} \\
&= p(U|V,W),
\end{aligned}$$

and

$$\begin{aligned} p(V|W) \cdot p(U|V,W) &= \frac{p(V,W)}{p(W)} \cdot \frac{p(U,V,W)}{p(V,W)} \\ &= \frac{p(U,V,W)}{p(W)} \\ &= p(U,V|W). \end{aligned}$$

So the lemma is proved. ■

Using the general rules from Lemma 2.3.2, we prove the following lemmas which describe the nature of dependencies between random variables in the hidden Markov model.

Notation 2.3.3 For convenience, sometimes $X_m, \ldots, X_n$ and its realizations $x_m, \ldots, x_n$ will be abbreviated X_m^n and x_m^n respectively. Similar notations are also applied for the $\{Y_t\}$ process and its realizations.

Lemma 2.3.4 *Let* $1 \leq m \leq t < n$.
(a). $p(X_{t+1}, Y_{t+1}|X_m^t, Y_m^t) = p(X_{t+1}, Y_{t+1}|X_t)$.
(b). $p(X_t, Y_t|X_{t+1}^n, Y_{t+1}^n) = p(X_t, Y_t|X_{t+1})$.

Proof :
By the third part of Lemma 2.3.2,

$$p(X_{t+1}, Y_{t+1}|X_m^t, Y_m^t) = p(X_{t+1}|X_m^t, Y_m^t) \cdot p(Y_{t+1}|X_m^{t+1}, Y_m^t). \tag{2.23}$$

By the first part of Lemma 2.3.2 and the Markov property,

$$\begin{aligned} p(X_{t+1}|X_m^t, Y_m^t) &= \frac{p(X_{t+1}|X_m^t) \cdot p(Y_m^t|X_m^{t+1})}{p(Y_m^t|X_m^t)} \\ &= \frac{p(X_{t+1}|X_t) \cdot p(Y_m^t|X_m^{t+1})}{p(Y_m^t|X_m^t)}. \end{aligned} \tag{2.24}$$

Also by the first part of Lemma 2.3.2 and condition (c) of Definition 2.2.1,

$$\begin{aligned} p(Y_{t+1}|X_m^{t+1}, Y_m^t) &= \frac{p(Y_{t+1}|X_m^{t+1}) \cdot p(Y_m^t|X_m^{t+1}, Y_{t+1})}{p(Y_m^t|X_m^{t+1})} \\ &= \frac{p(Y_{t+1}|X_{t+1}) \cdot p(X_m^{t+1}, Y_m^{t+1})}{p(Y_m^t|X_m^{t+1}) \cdot p(X_m^{t+1}, Y_{t+1})} \\ &= \frac{p(Y_{t+1}|X_{t+1}) \cdot p(Y_m^{t+1}|X_m^{t+1})}{p(Y_m^t|X_m^{t+1}) \cdot p(Y_{t+1}|X_m^{t+1})} \\ &= \frac{p(Y_{t+1}|X_{t+1}) \cdot p(Y_m^{t+1}|X_m^{t+1})}{p(Y_m^t|X_m^{t+1}) \cdot p(Y_{t+1}|X_{t+1})} \\ &= \frac{p(Y_m^{t+1}|X_m^{t+1})}{p(Y_m^t|X_m^{t+1})}. \end{aligned} \tag{2.25}$$

From (2.23), (2.24), (2.25) and conditions (b) and (c) of Definition 2.2.1,

$$\begin{aligned} p(X_{t+1}, Y_{t+1}|X_m^t, Y_m^t) &= \frac{p(X_{t+1}|X_t) \cdot p(Y_m^{t+1}|X_m^{t+1})}{p(Y_m^t|X_m^t)} \\ &= p(X_{t+1}|X_t) \cdot p(Y_{t+1}|X_{t+1}) \\ &= p(X_{t+1}, Y_{t+1}|X_t). \end{aligned}$$

The proof for (b) is similar using the first part of Lemma 2.3.2, the Markov property and conditions (b) and (c) of Definition 2.2.1. ■

Corollary 2.3.5 *Let* $1 \leq m < t < n$.

(a). $p(X_{t+1}|X_m^t, Y_m^t) = p(X_{t+1}|X_t)$.

(b). $p(Y_{t+1}|X_m^t, Y_m^t) = p(Y_{t+1}|X_t)$.

(c). $p(X_t|X_{t+1}^n, Y_{t+1}^n) = p(X_t|X_{t+1})$.

(d). $p(Y_t|X_{t+1}^n, Y_{t+1}^n) = p(Y_t|X_{t+1})$.

Proof :

For (a), using the first part of Lemma 2.3.4,

$$\begin{aligned} p(x_{t+1}|x_m^t, y_m^t) &= \int_{-\infty}^{\infty} p(x_{t+1}, y_{t+1}|x_m^t, y_m^t)\, d\mu(y_{t+1}) \\ &= \int_{-\infty}^{\infty} p(x_{t+1}, y_{t+1}|x_t)\, d\mu(y_{t+1}) \\ &= p(x_{t+1}|x_t), \end{aligned} \tag{2.26}$$

which gives

$$p(X_{t+1}|X_m^t, Y_m^t) = p(X_{t+1}|X_t).$$

The proofs for (b), (c) and (d) are similar using Lemma 2.3.4. ■

Lemma 2.3.6 *Let* $1 \leq m < t < n$.

(a). $p(X_{t+1}, Y_{t+1}|X_m^t, Y_{m+1}^t) = p(X_{t+1}, Y_{t+1}|X_t)$.

(b). $p(X_t, Y_t|X_{t+1}^n, Y_{t+1}^{n-1}) = p(X_t, Y_t|X_{t+1})$.

Proof :

For (a), by the first part of Lemma 2.3.2, Lemma 2.3.4 and the third part of Corollary 2.3.5,

$$\begin{aligned} p(X_{t+1}, Y_{t+1}|X_m^t, Y_{m+1}^t) &= p(X_{t+1}, Y_{t+1}|X_m, X_{m+1}^t, Y_{m+1}^t) \\ &= \frac{p(X_{t+1}, Y_{t+1}|X_{m+1}^t, Y_{m+1}^t) \cdot p(X_m|X_{m+1}^{t+1}, Y_{m+1}^{t+1})}{p(X_m|X_{m+1}^t, Y_{m+1}^t)} \\ &= \frac{p(X_{t+1}, Y_{t+1}|X_t) \cdot p(X_m|X_{m+1})}{p(X_m|X_{m+1})} \\ &= p(X_{t+1}, Y_{t+1}|X_t). \end{aligned}$$

The proof for (b) is similar using the first part of Lemma 2.3.2, Lemma 2.3.4 and Corollary 2.3.5. ■

Lemma 2.3.7 *Let* $1 \leq m \leq t < n$.

(a). $p(X_{t+1}^n, Y_{t+1}^n|X_m^t, Y_m^t) = p(X_{t+1}^n, Y_{t+1}^n|X_t)$.

(b). $p(X_m^t, Y_m^t|X_{t+1}^n, Y_{t+1}^n) = p(X_m^t, Y_m^t|X_{t+1})$.

Proof :

For (a), using the third part of Lemma 2.3.2 and the first parts of Lemma 2.3.4

and Lemma 2.3.6,

$$\begin{aligned}
&p(X_{t+1}^n, Y_{t+1}^n | X_m^t, Y_m^t) \\
&= p(X_{t+1}, Y_{t+1} | X_m^t, Y_m^t) p(X_{t+2}, Y_{t+2} | X_m^{t+1}, Y_m^{t+1}) \cdots p(X_n, Y_n | X_m^{n-1}, Y_m^{n-1}) \\
&= p(X_{t+1}, Y_{t+1} | X_t) p(X_{t+2}, Y_{t+2} | X_{t+1}) \cdots p(X_n, Y_n | X_{n-1}) \\
&= p(X_{t+1}, Y_{t+1} | X_t) p(X_{t+2}, Y_{t+2} | X_t^{t+1}, Y_{t+1}) \cdots p(X_n, Y_n | X_t^{n-1}, Y_{t+1}^{n-1}) \\
&= p(X_{t+1}^n, Y_{t+1}^n | X_t).
\end{aligned}$$

The proof for (b) is similar using the third part of Lemma 2.3.2 and the second parts of Lemma 2.3.4 and Lemma 2.3.6. ■

Lemma 2.3.8 *Let* $1 \leq k, l \leq t < m, n$.

(a). $p(X_{t+1}^m, Y_{t+1}^n | X_k^t, Y_l^t) = p(X_{t+1}^m, Y_{t+1}^n | X_t)$.

(b). $p(X_k^t, Y_l^t | X_{t+1}^m, Y_{t+1}^n) = p(X_k^t, Y_l^t | X_{t+1})$.

Proof :

For (a), let $1 \leq k, l, \leq t < m, n$ and suppose that $k < l$ and $m < n$, then by the first part of Lemma 2.3.7

$$\begin{aligned}
&p(x_{t+1}^m, y_{t+1}^n | x_k^t, y_l^t) \\
&= \frac{p(x_k^t, x_{t+1}^m, y_l^t, y_{t+1}^n)}{p(x_k^t, y_l^t)} \\
&= \frac{\sum_{x_{m+1}=1}^K \cdots \sum_{x_n=1}^K \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_k^t, x_{t+1}^n, y_k^t, y_{t+1}^n) d\mu(y_k) \cdots d\mu(y_{l-1})}{p(x_k^t, y_l^t)} \\
&= \frac{\sum_{x_{m+1}=1}^K \cdots \sum_{x_n=1}^K \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_{t+1}^n, y_{t+1}^n | x_k^t, y_k^t) p(x_k^t, y_k^t) d\mu(y_k) \cdots d\mu(y_{l-1})}{p(x_k^t, y_l^t)} \\
&= \frac{\sum_{x_{m+1}=1}^K \cdots \sum_{x_n=1}^K \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_{t+1}^n, y_{t+1}^n | x_t) p(x_k^t, y_k^t) d\mu(y_k) \cdots d\mu(y_{l-1})}{p(x_k^t, y_l^t)} \\
&= \frac{p(x_{t+1}^m, y_{t+1}^n | x_t) p(x_k^t, y_l^t)}{p(x_k^t, y_l^t)} \\
&= p(x_{t+1}^m, y_{t+1}^n | x_t). \qquad (2.27)
\end{aligned}$$

The proofs for the other possibilities of k and l are similar. So from (2.27), (a) follows.

The proof for (b) is similar using the second part of Lemma 2.3.7. ■

Corollary 2.3.9 *Let* $1 \leq k, l \leq t < m, n$.

(a). $p(X_{t+1}^m | X_k^t, Y_l^t) = p(X_{t+1}^m | X_t)$.

(b). $p(Y_{t+1}^n | X_k^t, Y_l^t) = p(Y_{t+1}^n | X_t)$.

(c). $p(X_k^t | X_{t+1}^m, Y_{t+1}^n) = p(X_k^t | X_{t+1})$.

(d). $p(Y_l^t | X_{t+1}^m, Y_{t+1}^n) = p(Y_l^t | X_{t+1})$.

Proof :

This lemma is a direct consequence of Lemma 2.3.8 which is obtained by integrating part (a) and (b) of Lemma 2.3.8 with respect to x and y. ■

Corollary 2.3.10 *Let* $1 \leq k < t$, *then*

$$p(Y_t | X_t, Y_k^{t-1}) = p(Y_t | X_t).$$

Proof :

By the first part of Lemma 2.3.2 and the third part of Corollary 2.3.9,

$$\begin{aligned} p(Y_t | X_t, Y_k^{t-1}) &= \frac{p(Y_t | X_t) \cdot p(Y_k^{t-1} | X_t, Y_t)}{p(Y_k^{t-1} | X_t)} \\ &= \frac{p(Y_t | X_t) \cdot p(Y_k^{t-1} | X_t)}{p(Y_k^{t-1} | X_t)} \\ &= p(Y_t | X_t). \end{aligned}$$

■

Lemma 2.3.11

(a). If $1 \leq k \leq l < t \leq m, n$, *then* $p(X_t^m | X_k^l, Y_1^n) = p(X_t^m | X_l, Y_{l+1}^n)$.

(b). If $1 \leq l \leq t < m \leq n_1, n_2$, *then* $p(X_l^t | X_m^{n_1}, Y_1^{n_2}) = p(X_l^t | X_m, Y_1^{m-1})$.

Proof :

For(a), by the first part of Lemma 2.3.8, the second parts of Corollary 2.3.9 and Lemma 2.3.2,

$$\begin{aligned}
p(x_t^m | x_k^l, y_1^n) &= \frac{p(x_k^l, x_t^m, y_1^n)}{p(x_k^l, y_1^n)} \\
&= \frac{p(x_k^l, x_t^m, y_1^l, y_{l+1}^n)}{p(x_k^l, y_1^l, y_{l+1}^n)} \\
&= \frac{\sum_{x_{l+1}=1}^K \cdots \sum_{x_{t-1}=1}^K p(x_k^l, x_{l+1}^m, y_1^l, y_{l+1}^n)}{p(x_k^l, y_1^l, y_{l+1}^n)} \\
&= \frac{\sum_{x_{l+1}=1}^K \cdots \sum_{x_{t-1}=1}^K p(x_{l+1}^m, y_{l+1}^n | x_k^l, y_1^l)}{p(y_{l+1}^n | x_k^l, y_1^l)} \\
&= \frac{\sum_{x_{l+1}=1}^K \cdots \sum_{x_{t-1}=1}^K p(x_{l+1}^m, y_{l+1}^n | x_l)}{p(y_{l+1}^n | x_l)} \\
&= \frac{p(x_t^m, y_{l+1}^n | x_l)}{p(y_{l+1}^n | x_l)} \\
&= p(x_t^m | x_l, y_{l+1}^n).
\end{aligned}$$

Thus (a) follows.

The proof for (b) is similar, using the second part of Lemma 2.3.8, the last part of Corollary 2.3.9 and the second part of Lemma 2.3.2. ■

2.4 Representations of Hidden Markov Models

The aim of this section is to find parameters which determine the characteristics of a hidden Markov model.

Since the Markov chain $\{X_t\}$ in a hidden Markov model $\{(X_t, Y_t)\}$ is not observable, then inference concerning the hidden Markov model has to be based on the information of $\{Y_t\}$ alone. By knowing the finite dimensional joint distributions of $\{Y_t\}$, parameters which characterize the hidden Markov model can then be analysed.

Let $\{(X_t, Y_t)\}$ be a hidden Markov model defined on the probability space $(\Omega, \mathcal{F}, P)$, taking values on $\boldsymbol{S} \times \mathcal{Y}$, where $\boldsymbol{S} = \{1, \ldots, K\}$ and $\mathcal{Y} = \boldsymbol{R}$. Let $A = (\alpha_{ij})$ be the transition probability matrix and $\pi = (\pi_i)$ be the initial probability distribution of the Markov chain $\{X_t\}$. Assume for $i = i, \ldots, K$, the conditional densities of Y_t given $X_t = i$ with respect to the measure μ, $p(\cdot|i)$, belong to the same family $\boldsymbol{\mathcal{F}}$, where $\boldsymbol{\mathcal{F}} = \{f(\cdot|\theta) : \theta \in \Theta\}$ is a family of densities on a Euclidean space with respect to the measure μ, indexed by $\theta \in \Theta$. This means that for each $i = 1, \ldots, K$,

$$p(\cdot|i) = f(\cdot, \theta_i),$$

for some $\theta_i \in \Theta$.

For $y \in \mathcal{Y}$ and $i, j = 1, \ldots, K$, define

$$m_{ij}(y) = \alpha_{ij} \cdot f(y, \theta_j).$$

For every $y \in \mathcal{Y}$, let $M(y)$ be the $K \times K$-matrix defined by

$$M(y) = (m_{ij}(y)).$$

Then

$$M(y) = A \cdot B(y), \qquad y \in \mathcal{Y}, \tag{2.28}$$

where

$$B(y) = \begin{pmatrix} f(y,\theta_1) & 0 & 0 & \cdots & 0 \\ 0 & f(y,\theta_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & f(y,\theta_K) \end{pmatrix}.$$

Observe that

$$\begin{aligned}\int_{-\infty}^{\infty} M(y)\,d\mu(y) &= \left(\int_{-\infty}^{\infty} m_{ij}(y)\,d\mu(y)\right)\\ &= \left(\int_{-\infty}^{\infty} \alpha_{ij} f(y,\theta_j)\,d\mu(y)\right)\\ &= (\alpha_{ij})\\ &= A. \end{aligned} \tag{2.29}$$

Lemma 2.4.1 *For each $n \in \boldsymbol{N}$, the n-dimensional joint density fuction of $Y_1, Y_2, \ldots, Y_n$ is*

$$p(Y_1, Y_2, \ldots, Y_n) = \pi B(Y_1) M(Y_2) \cdots M(Y_n) e, \tag{2.30}$$

where $e = (1,1,1,\ldots,1)^T$.

Proof :

By Lemma 2.3.2, Corollary 2.3.5, Lemma 2.3.6 and Lemma 2.3.7, the joint density function of $Y_1, Y_2, \ldots, Y_n$ can be expressed as,

$$\begin{aligned} p(y_1, y_2, \ldots, y_n) &= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} p(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)\\ &= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \Big\{ p(x_1) \cdot p(y_1|x_1)\\ &\qquad \times p(x_2|x_1, y_1) \cdot p(y_2|x_1^2, y_1)\\ &\qquad \times \cdots \times p(x_n|x_1^{n-1}, y_1^{n-1}) \cdot p(y_n|x_1^n, y_1^{n-1}) \Big\}\\ &= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \Big\{ p(x_1) \cdot p(y_1|x_1)\\ &\qquad \times p(x_2|x_1) \cdot p(y_2|x_2)\\ &\qquad \times \cdots \times p(x_n|x_{n-1}) \cdot p(y_n|x_n) \Big\}\\ &= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \Big\{ P(X_1 = x_1) \cdot f(y_1, \theta_{x_1})\\ &\qquad \times P(X_2 = x_2|X_1 = x_1) \cdot f(y_2, \theta_{x_2})\\ &\qquad \times \cdots \times P(X_n = x_n|X_{n-1} = x_{n-1}) \cdot f(y_n, \theta_{x_n}) \Big\} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} \cdot f(y_1, \theta_{x_1}) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t} \cdot f(y_t, \theta_{x_t}) \\
&= \pi B(y_1) M(y_2) \cdots M(y_n) e,
\end{aligned}
$$

so the conclusion of the lemma follows. ■

Corollary 2.4.2 *If $\{X_t\}$ is a stationary Markov chain, then for each $n \in \boldsymbol{N}$, the n-dimensional joint density function of $Y_1, Y_2, \ldots, Y_n$ is*

$$p(Y_1, Y_2, \ldots, Y_n) = \pi M(Y_1) M(Y_2) \cdots M(y_n) e.$$

Proof :

Since $\{X_t\}$ is a stationary Markov chain, then the initial probability distribution π satisfies

$$\pi A = A. \tag{2.31}$$

By Lemma 2.4.1 and equation (2.31), for any $n \in \boldsymbol{N}$, the n-dimensional joint density function of $Y_1, Y_2, \ldots, Y_n$ is

$$
\begin{aligned}
p(Y_1, Y_2, \ldots, Y_n) &= \pi B(Y_1) M(Y_2) \cdots M(Y_n) e \\
&= \pi A B(Y_1) M(Y_2) \cdots M(Y_n) e \\
&= \pi M(Y_1) M(Y_2) \cdots M(Y_n) e.
\end{aligned}
$$

■

Since for $i = 1, \ldots, K$,

$$P(X_n = i) = \pi_i \qquad \forall n \in \boldsymbol{N},$$

when $\{X_t\}$ is a stationary Markov chain, then using a similar proof as in the proofs of Lemma 2.4.1 and Corollary 2.4.2, for any $m, n \in \boldsymbol{N}$, the n-dimensional joint density function of $Y_m, Y_{m+1}, \ldots, Y_{m+n-1}$ has the form

$$p(Y_m, Y_{m+1}, \ldots, Y_{m+n-1}) = \pi M(Y_m) M(Y_{m+1}) \cdots M(Y_{m+n-1}) e. \tag{2.32}$$

Equation (2.32) shows that the observation process $\{Y_t\}$ is a (strictly) stationary process. This implies the pair of stochastic processes $\{(X_t, Y_t)\}$ is also (strictly) stationary. So we have the following corollary.

Corollary 2.4.3 *If $\{X_t\}$ is a stationary Markov chain, then the hidden Markov model $\{(X_t, Y_t)\}$ is also stationary.*

Lemma 2.4.4 *For each $n \in \boldsymbol{N}$, the conditional density function of $Y_1, Y_2, \ldots, Y_n$ given $X_1 = i$, for $i = 1, \ldots, K$, is*

$$p(Y_1, Y_2, \ldots, Y_n | X_1 = i) = e_i^T B(Y_1) M(Y_2) \cdots M(Y_n) e,$$

where $e_i^T = (0, \ldots, 0, 1, 0, \ldots, 0)$.

Proof :

Let $n \in \boldsymbol{N}$ and $i \in \{1, \ldots, K\}$, then by Lemma 2.3.2, Corollary 2.3.5, Lemma 2.3.6 and Lemma 2.3.7, the conditional density function of $Y_1, Y_2, \ldots, Y_n$ given $X_1 = i$ is

$$\begin{aligned}
p(y_1, y_2, \ldots, y_n | i) &= \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} p(y_1, x_2, y_2, \ldots, x_n, y_n | i) \\
&= \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} \Big\{ p(y_1|i) p(x_2|i, y_1) p(y_2|i, x_2, y_1) \\
&\qquad \times \cdots \times p(x_n|i, x_2^{n-1}, y_1^{n-1}) p(y_n|i, x_2^n, y_1^{n-1}) \Big\} \\
&= \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} \Big\{ p(y_1|i) p(x_2|i) p(y_2|x_2) \\
&\qquad \times \cdots \times p(x_n|x_{n-1}) p(y_n|x_n) \Big\} \\
&= \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} \Big\{ f(y_1, \theta_i) P(X_2 = x_2 | X_1 = i) f(y_2, \theta_{x_2}) \\
&\qquad \times \cdots \times P(X_n = x_n | X_{n-1} = x_{n-1}) f(y_n, \theta_{x_n}) \Big\} \\
&= \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} f(y_1, \theta_i) \alpha_{i,x_2} f(y_2, \theta_{x_2}) \prod_{t=3}^{n} \alpha_{x_{t-1}, x_t} f(y_t, \theta_{x_t}) \\
&= e_i^T B(y_1) M(y_2) \cdots M(y_n) e.
\end{aligned}$$

So the conclusion of the lemma follows. ■

From Lemma 2.4.1, it can be seen that the law of the hidden Markov model $\{(X_t, Y_t)\}$ is completely specified by :

(a). The size K.

(b). The transition probability matrix $A = (\alpha_{ij})$, satisfying

$$\alpha_{ij} \geq 0, \qquad \sum_{j=1}^{K} \alpha_{ij} = 1, \qquad i, j = 1, \ldots, K.$$

(c). The initial probability distribution $\pi = (\pi_i)$, satisfying

$$\pi_i \geq 0, \qquad i = 1, \ldots, K, \qquad \sum_{i=1}^{K} \pi_i = 1.$$

(d). The vector $\theta = (\theta_i)^T$, $\theta_i \in \Theta$, $i = 1, \ldots, K$, which describes the conditional densities of Y_t given $X_t = i$, $i = 1, \ldots, K$.

Definition 2.4.5 *Let*

$$\phi = (K, A, \pi, \theta).$$

The parameter ϕ is called a ***representation*** *of the hidden Markov model $\{(X_t, Y_t)\}$.*

Thus, the hidden Markov model $\{(X_t, Y_t)\}$ can be represented by a representation $\phi = (K, A, \pi, \theta)$.

On the otherhand, we can also generate a hidden Markov model $\{(X_t, Y_t)\}$ from a representation $\phi = (K, A, \pi, \theta)$, by choosing a Markov chain $\{X_t\}$ which takes values on $\{1, \ldots, K\}$ and its law is determined by the $K \times K$-transition probability matrix A and the initial probability π, and an observation process $\{Y_t\}$ taking values on $\mathcal{Y}$, where the density functions of Y_t given $X_t = i$, for $i = 1, \ldots, K$ are determined by θ.

2.5 Equivalent Representations

Let $\phi = (K, A, \pi, \theta)$ and $\widehat{\phi} = (\widehat{K}, \widehat{A}, \widehat{\pi}, \widehat{\theta})$ be two representations which respectively generate hidden Markov models $\{(X_t, Y_t)\}$ and $\{(\widehat{X}_t, Y_t)\}$. The $\{(X_t, Y_t)\}$ takes values on $\{1, \ldots, K\} \times \mathcal{Y}$ and $\{(\widehat{X}_t, Y_t)\}$ takes values on $\{1, \ldots, \widehat{K}\} \times \mathcal{Y}$. For any $n \in \boldsymbol{N}$, let $p_\phi(\cdot, \cdots, \cdot)$ and $p_{\widehat{\phi}}(\cdot, \cdots, \cdot)$ be the n-dimensional joint density function of $Y_1, \ldots Y_n$ with respect to ϕ and $\widehat{\phi}$. Suppose that for every $n \in \boldsymbol{N}$,

$$p_\phi(Y_1, \ldots, Y_n) = p_{\widehat{\phi}}(Y_1, \ldots, Y_n).$$

Then $\{Y_t\}$ has the same law under ϕ and $\widehat{\phi}$. Since in hidden Markov models $\{(X_t, Y_t)\}$ and $\{(\widehat{X}_t, Y_t)\}$, the Markov chains $\{X_t\}$ and $\{\widehat{X}_t\}$ are not observable and we only observed the values of $\{Y_t\}$, then theoretically, the hidden Markov models $\{(X_t, Y_t)\}$ and $\{(\widehat{X}_t, Y_t)\}$ are *indistinguishable*. In this case, it is said that $\{(X_t, Y_t)\}$ and $\{(\widehat{X}_t, Y_t)\}$ are *equivalent*. The representations ϕ and $\widehat{\phi}$ are also said to be *equivalent*, and will be denoted as $\phi \sim \widehat{\phi}$.

For each $K \in \boldsymbol{N}$, define

$$\begin{aligned} \Phi_K \;=\; \Big\{ \phi \,:\, & \phi = (K, A, \pi, \theta), \text{ where } A, \pi \text{ and } \theta \text{ satisfy :} \\ & A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K \\ & \pi = (\pi_i), \quad \pi_i \geq 0, \quad i = 1, \ldots, K, \quad \sum_{i=1}^{K} \pi_i = 1 \\ & \theta = (\theta_i)^T, \quad \theta_i \in \Theta, \quad i = 1, \ldots, K \Big\} \end{aligned} \tag{2.33}$$

and

$$\Phi = \bigcup_{K \in \boldsymbol{N}} \Phi_K. \tag{2.34}$$

The relation $\sim$ is now defined on Φ as follows.

Definition 2.5.1 *Let $\phi, \widehat{\phi} \in \Phi$. Representations ϕ and $\widehat{\phi}$ are said to be* ***equivalent****, denoted as*

$$\phi \sim \widehat{\phi}$$

if and only if for every $n \in \boldsymbol{N}$,

$$p_{\phi}(Y_1, Y_2, \ldots, Y_n) = p_{\widehat{\phi}}(Y_1, Y_2, \ldots, Y_n).$$

Remarks 2.5.2 It is clear that relation $\sim$ forms an equivalence relation on Φ.

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, then under ϕ, $Y_1, \ldots, Y_n$, for any n, has joint density function

$$p_{\phi}(y_1, \ldots, y_n) = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} f(y_1, \theta_{x_1}) \cdot \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t} f(y_t, \theta_{x_t}). \tag{2.35}$$

Let σ be any permutation of $\{1, 2, \ldots, K\}$. Define

$$\begin{aligned} \sigma(A) &= (\alpha_{\sigma(i),\sigma(j)}) \\ \sigma(\pi) &= (\pi_{\sigma(i)}) \\ \sigma(\theta) &= (\theta_{\sigma(i)})^T. \end{aligned}$$

Let

$$\sigma(\phi) = (K, \sigma(A), \sigma(\pi), \sigma(\theta)),$$

then $\sigma(\phi) \in \Phi_K$ and easy to see from (2.35) that

$$p_{\phi}(y_1, \ldots, y_n) = p_{\sigma(\phi)}(y_1, \ldots, y_n).$$

implying $\phi \sim \sigma(\phi)$. So we have the following lemma.

Lemma 2.5.3 *Let $\phi \in \Phi_K$, then for every permutation σ of $\{1, 2, \ldots, K\}$,*

$$\sigma(\phi) \sim \phi.$$

Lemma 2.5.4 *Let $\phi = (K, A, \pi, \theta) \in \Phi_K$. If $\theta_i = \gamma$, $i = 1, \ldots, K$, for some $\gamma \in \Theta$, then*

$$\phi \sim \phi(\gamma),$$

where $\phi(\gamma) = (1, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_1$, with $\widehat{A} = (1)$, $\widehat{\pi} = (1)$ and $\widehat{\theta} = (\gamma)$.

Proof :

For any $n \in \boldsymbol{N}$,

$$\begin{aligned} p_\phi(y_1, \ldots, y_n) &= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} f(y_1, \gamma) \prod_{t=1}^{n} \alpha_{x_{t-1}, x_t} f(y_t, \gamma) \\ &= \prod_{t=1}^{n} f(y_t, \gamma) \\ &= p_{\phi(\gamma)}(y_1, \ldots, y_n), \end{aligned}$$

hence $\phi \sim \phi(\gamma)$. ■

Lemma 2.5.5 *Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, where π is a stationary probability distribution of A. Let N be the number of non-zero π_i. Then there is $\widehat{\phi} = (N, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_N$, such that :*

(a). $\widehat{\pi}_i > 0$, for $i = 1, \ldots, N$.

(b). $\widehat{\pi}$ is a stationary probability distribution of $\widehat{A}$.

(c). $\phi \sim \widehat{\phi}$.

Proof :

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, where π is a stationary probability distribution of A. Let N be the number of non-zero π_i. Without lost of generality, suppose that

$$\begin{aligned} \pi_i &> 0, \qquad \text{for } i = 1, \ldots, N \\ \pi_i &= 0, \qquad \text{for } i = N+1, \ldots, K. \end{aligned}$$

Since π satisfies

$$\pi A = \pi,$$

then

$$\alpha_{ij} = 0, \qquad \text{for} \qquad i = 1, \ldots, N \qquad \text{and} \qquad j = N+1, \ldots, K.$$

Thus A has form

$$A = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,N} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \cdots & \alpha_{N,N} & 0 & \cdots & 0 \\ \alpha_{N+1,1} & \cdots & \alpha_{N+1,N} & \alpha_{N+1,N+1} & \cdots & \alpha_{N+1,K} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{K,1} & \cdots & \alpha_{K,N} & \alpha_{K,N+1} & \cdots & \alpha_{K,K} \end{pmatrix}$$

Set

$$\begin{aligned} \widehat{\alpha}_{ij} &= \alpha_{ij}, \qquad i,j = 1, \ldots, N \\ \widehat{\pi}_i &= \pi_i, \qquad i = 1, \ldots, N \\ \widehat{\theta}_i &= \theta_i, \qquad i = 1, \ldots, N. \end{aligned}$$

Let

$$\widehat{A} = (\widehat{\alpha}_{ij}), \qquad \widehat{\pi} = (\widehat{\pi}_i), \qquad \widehat{\theta} = (\widehat{\theta}_i)^T$$

and $1 \times \widehat{K}$-matrix

$$\widehat{e} = (1, 1, \ldots, 1).$$

Then it is clear that

$$\widehat{\pi}_i > 0, \qquad \text{for} \qquad i = 1, \ldots, N$$

and

$$\widehat{\pi}\,\widehat{A} = \widehat{\pi}.$$

Let

$$\widehat{B}(y) = \begin{pmatrix} f(y,\theta_1) & 0 & \cdots & 0 \\ 0 & f(y,\theta_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(y,\theta_N) \end{pmatrix}, \qquad y \in \mathcal{Y}$$

and

$$\widehat{M}(y) = \hat{A}\hat{B}(y).$$

Take $\hat{\phi} = (N, \hat{A}, \hat{\pi}, \hat{\theta})$, then $\hat{\phi} \in \Phi_N$ and it is clear that

$$\begin{aligned} p_{\phi}(y_1, \ldots, y_n) &= \pi M(y_1)M(y_2)\cdots M(y_n)e \\ &= \hat{\pi}\widehat{M}(y_1)\widehat{M}(y_2)\cdots\widehat{M}(y_n)\hat{e} \\ &= p_{\hat{\phi}}(y_1, \ldots, y_n), \end{aligned}$$

implying $\phi \sim \hat{\phi}$. ■

Next lemma gives sufficient conditions for representations to be equivalent. The idea of this lemma comes from [22].

Lemma 2.5.6 *Let $\phi = (K, A, \pi, \theta) \in \Phi_K$ and $\hat{\phi} = (\widehat{K}, \hat{A}, \hat{\pi}, \hat{\theta}) \in \Phi_{\widehat{K}}$. If there are $K \times \widehat{K}$-matrix X and $\widehat{K} \times K$-matrix Y, such that*

$$\hat{A} = YAX \tag{2.36}$$

$$X\hat{B}(y) = B(y)X, \qquad \forall y \in \mathcal{Y} \tag{2.37}$$

$$\hat{\pi} = \pi X$$

$$\hat{e} = Ye$$

$$XY = I_K,$$

then $\phi \sim \hat{\phi}$.

Proof :

From (2.36) and (2.37), for every $y \in \mathcal{Y}$,

$$\widehat{M}(y) = \hat{A}\hat{B}(y)$$

$$
\begin{aligned}
&= YAX\hat{B}(y) \\
&= YAB(y)X \\
&= YM(y)X.
\end{aligned}
$$

For any $n \in \boldsymbol{N}$,

$$
\begin{aligned}
p_{\hat{\phi}}(y_1, \ldots, y_n) &= \hat{\pi}\hat{B}(y_1)\widehat{M}(y_2)\cdots\widehat{M}(y_n)\hat{e} \\
&= \pi X\hat{B}(y_1)YM(y_2)X\cdots YM(y_n)XYe \\
&= \pi B(y_1)XYM(y_2)X\cdots YM(y_n)XYe \\
&= \pi B(y_1)I_K M(y_2)\cdots I_K M(y_n)I_K e \\
&= \pi B(y_1)M(y_2)\cdots M(y_n)e \\
&= p_\phi(y_1, \ldots, y_n).
\end{aligned}
$$

Hence $\phi \sim \hat{\phi}$. ■

Lemma 2.5.7 *Let $\phi = (K, A, \pi, \theta) \in \Phi_K$ and $\hat{\phi} = (\widehat{K}, \hat{A}, \hat{\pi}, \hat{\theta}) \in \Phi_{\widehat{K}}$, where π and $\hat{\pi}$ are stationary probability distributions of A and $\hat{A}$ respectively. If there are $K \times \widehat{K}$-matrix X and $\widehat{K} \times K$-matrix Y, such that*

$$
\begin{aligned}
\widehat{M}(y) &= YM(y)X, \qquad \forall y \in \mathcal{Y} \qquad (2.38) \\
\hat{\pi} &= \pi X \\
\hat{e} &= Ye \\
XY &= I_K,
\end{aligned}
$$

then $\phi \sim \hat{\phi}$.

Remarks 2.5.8 Equation (2.38) implies $\hat{A} = YAX$.

Proof :

For any $n \in \boldsymbol{N}$,

$$
p_{\hat{\phi}}(y_1, \ldots, y_n) = \hat{\pi}\widehat{M}(y_1)\widehat{M}(y_2)\cdots\widehat{M}(y_n)\hat{e}
$$

$$
\begin{aligned}
&= \pi XYM(y_1)XYM(y_2)X\cdots YM(y_n)XYe \\
&= \pi I_K M(y_1)I_K M(y_2)I_K \cdots I_K M(y_n)I_K e \\
&= \pi M(y_1)M(y_2)\cdots M(y_n)e \\
&= p_\phi(y_1,\ldots,y_n).
\end{aligned}
$$

Thus $\phi \sim \widehat{\phi}$. ■

Based on Lemma 2.5.6 and Lemma 2.5.7, we derive the following lemmas.

Lemma 2.5.9 *For any $K \in \boldsymbol{N}$ and $\phi \in \Phi_K$, there is $\widehat{\phi} \in \Phi_{K+1}$, such that $\phi \sim \widehat{\phi}$.*

Proof :

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$. Define a $K \times (K+1)$-matrix X and a $(K+1) \times K$-matrix Y respectively as follow

$$
X = \begin{pmatrix} I_{K-1} & O_{K-1,2} \\ O_{1,K-1} & a \quad b \end{pmatrix}, \qquad Y = \begin{pmatrix} I_{K-1} & O_{K-1,1} \\ & 1 \\ O_{2,K-1} & \\ & 1 \end{pmatrix} \tag{2.39}
$$

where a and b are any real numbers, such that $a, b \geq 0$ and $a + b = 1$. Notice that

$$XY = I_K$$

and

$$\widehat{e} = Ye.$$

Let $\widehat{A} = (\widehat{\alpha}_{ij})$ be a $(K+1) \times (K+1)$-matrix defined by

$$\widehat{A} = YAX$$

$$= \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,K-1} & a\cdot\alpha_{1,K} & b\cdot\alpha_{1,K} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha_{K-1,1} & \cdots & \alpha_{K-1,K-1} & a\cdot\alpha_{K-1,K} & b\cdot\alpha_{K-1,K} \\ \alpha_{K,1} & \cdots & \alpha_{K,K-1} & a\cdot\alpha_{K,K} & b\cdot\alpha_{K,K} \\ \alpha_{K,1} & \cdots & \alpha_{K,K-1} & a\cdot\alpha_{K,K} & b\cdot\alpha_{K,K} \end{pmatrix}. \quad (2.40)$$

It is clear that

$$\widehat{\alpha}_{ij} \geq 0, \qquad i,j = 1,\ldots,K+1$$
$$\sum_{j=1}^{K+1} \widehat{\alpha}_{ij} = 1, \qquad i = 1,\ldots,K+1.$$

Let $\widehat{\pi} = (\widehat{\pi}_i)$ be a $1 \times (K+1)$-matrix which is defined by

$$\begin{aligned} \widehat{\pi} &= \pi X \\ &= (\pi_1,\ldots,\pi_{K-1}, a\cdot\pi_K, b\cdot\pi_K), \end{aligned} \quad (2.41)$$

then it is obvious that

$$\widehat{\pi}_i \geq 0, \qquad i = 1,\ldots,K+1 \qquad \text{and} \qquad \sum_{i=1}^{K} \widehat{\pi}_i = 1.$$

Let $\widehat{\theta} = (\widehat{\theta}_i)$ be a $(K+1) \times 1$-matrix which is defined by

$$\begin{aligned} \widehat{\theta} &= Y\theta \\ &= (\theta_1,\ldots,\theta_{K-1},\theta_K,\theta_K)^T \end{aligned} \quad (2.42)$$

and for $y \in \mathcal{Y}$, $\widehat{B}(y)$ be a $(K+1) \times (K+1)$-diagonal matrix defined by

$$\widehat{B}(y) = \begin{pmatrix} f(y,\theta_1) & 0 & \cdots & 0 & 0 \\ 0 & f(y,\theta_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & f(y,\theta_K) & 0 \\ 0 & 0 & \cdots & 0 & f(y,\theta_K) \end{pmatrix}. \quad (2.43)$$

Notice that

$$X\widehat{B}(y) = B(y)X, \qquad \forall y \in \mathcal{Y}.$$

Let $\widehat{\phi} = (K+1, \widehat{A}, \widehat{\pi}, \widehat{\theta})$, then $\widehat{\phi} \in \Phi_{K+1}$ and by Lemma 2.5.6, $\phi \sim \widehat{\phi}$. ■

From the proof of Lemma 2.5.9, for $\phi \in \Phi_K$, there are infinitely many $\widehat{\phi} \in \Phi_{K+1}$ depending on how a and b were chosen, such that $\phi \sim \widehat{\phi}$. So we have the following corollary.

Corollary 2.5.10 *For $\phi \in \Phi_K$, there are infinitely many $\widehat{\phi} \in \Phi_{K+1}$ such that $\phi \sim \widehat{\phi}$.*

Lemma 2.5.11 *For any $K \in \mathbf{N}$ and $\phi = (K, A, \pi, \theta) \in \Phi_K$, where π is a stationary probability distribution of A, then there is $\phi = (K+1, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K+1}$ such that :*

(a). $\widehat{\pi}$ is a stationary probability distribution of $\widehat{A}$.

(b). $\widehat{\phi} \sim \phi$.

Proof :

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, where π is a stationary probability distribution of A. Let $\widehat{\phi} = (K+1, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K+1}$ as in the proof of Lemma 2.5.9, hence $\phi \sim \widehat{\phi}$. Since π is a stationary probability distribution of A, then

$$\pi A = \pi,$$

implying

$$\begin{aligned} \widehat{\pi}\widehat{A} &= \pi XYAX \\ &= \pi I_K AX \\ &= \pi AX \\ &= \pi X \\ &= \widehat{\pi}. \end{aligned}$$

So $\widehat{\pi}$ is a stationary probability distribution of $\widehat{A}$. ■

Remarks 2.5.12 In Lemma 2.5.11, if $\pi_i > 0$, for $i = 1, \ldots, K$, then by choosing $a, b > 0$ in matrix X, we have $\widehat{\pi}_i > 0$, for $i = 1, \ldots, K+1$.

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, then by Lemma 2.4.4, the conditional density function of $Y_1, \ldots, Y_n$, given $X_1 = i$, under ϕ is,

$$p_\phi(Y_1, \ldots, Y_n | X_1 = i) = e_i^T B(Y_1) M(Y_2) \cdots M(Y_n) e.$$

Define,

$$q_\phi(Y_1, \ldots, Y_n) = \max_{1 \le i \le K} p_\phi(Y_1, \ldots, Y_n | X_1 = i).$$

Lemma 2.5.13 *For any $K \in \boldsymbol{N}$ and $\phi \in \Phi_K$, there is $\widehat{\phi} \in \Phi_{K+1}$, such that :*

(a). $\phi \sim \widehat{\phi}$.

(b). $q_\phi(Y_1, \ldots, Y_n) = q_{\widehat{\phi}}(Y_1, \ldots, Y_n)$, *for every* $n \in \boldsymbol{N}$.

Proof :

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$. Let $\widehat{\phi} = (K+1, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K+1}$, as in the proof of Lemma 2.5.9, then $\widehat{\phi} \sim \phi$. Notice that from definition of X in (2.39),

$$\widehat{e}_i^T = e_i^T X, \qquad \text{for} \quad i = 1, \ldots, K-1. \tag{2.44}$$

Therefore by (2.44) and Lemma 2.4.4, for $i = 1, \ldots, K-1$,

$$\begin{aligned}
p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = i) &= \widehat{e}_i^T \widehat{B}(Y_1) \widehat{M}(Y_2) \cdots \widehat{M}(Y_n) \widehat{e} \\
&= e_i^T X \widehat{B}(Y_1) Y M(Y_2) X \cdots Y M(Y_n) X Y e \\
&= e_i^T B(Y_1) X Y M(Y_2) X \cdots Y M(Y_n) X Y e \\
&= e_i^T B(Y_1) I_K M(Y_2) I_K \cdots I_K M(Y_n) I_K e \\
&= e_i^T B(Y_1) M(Y_2) \cdots M(Y_n) e \\
&= p_\phi(Y_1, \ldots, Y_n | X_1 = i)
\end{aligned} \tag{2.45}$$

Since by (2.40), the K-th and $K+1$-th rows of $\widehat{A}$ are the same and $\widehat{\theta}_K = \widehat{\theta}_{K+1}$, then by Lemma 2.4.4,

$$\begin{aligned}
& p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = K) \\
&= f(Y_1, \widehat{\theta}_K) \sum_{x_2=1}^{K+1} \cdots \sum_{x_n=1}^{K+1} \widehat{\alpha}_{K,x_2} f(Y_2, \widehat{\theta}_{x_2}) \prod_{t=3}^{n} \widehat{\alpha}_{x_{t-1},x_t} f(Y_t, \widehat{\theta}_{x_t}) \\
&= f(Y_1, \widehat{\theta}_{K+1}) \sum_{x_2=1}^{K+1} \cdots \sum_{x_n=1}^{K+1} \widehat{\alpha}_{K+1,x_2} f(Y_2, \widehat{\theta}_{x_2}) \prod_{t=3}^{n} \widehat{\alpha}_{x_{t-1},x_t} f(Y_t, \widehat{\theta}_{x_t}) \\
&= p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = K+1). \qquad (2.46)
\end{aligned}$$

Also notice that for $a, b \geq 0$ and $a + b = 1$,

$$a\widehat{e}_K^T + b\widehat{e}_{K+1}^T = e_K^T X, \qquad (2.47)$$

then by (2.47) and Lemma 2.4.4,

$$\begin{aligned}
& a p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = K) + b p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = K+1) \\
&= a\widehat{e}_K^T \widehat{B}(Y_1) \widehat{M}(Y_2) \cdots \widehat{M}(Y_n) \widehat{e} + b\widehat{e}_K^T \widehat{B}(Y_1) \widehat{M}(Y_2) \cdots \widehat{M}(Y_n) \widehat{e} \\
&= (a\widehat{e}_K^T + b\widehat{e}_K^T) \widehat{B}(Y_1) \widehat{M}(Y_2) \cdots \widehat{M}(Y_n) \widehat{e} \\
&= e_K^T X \widehat{B}(Y_1) Y M(Y_2) X \cdots Y M(Y_n) X Y e \\
&= e_K^T B(Y_1) X Y M(Y_2) X \cdots Y M(Y_n) X Y e \\
&= e_K^T B(Y_1) I_K M(Y_2) I_K \cdots I_K M(Y_n) I_K e \\
&= e_K^T B(Y_1) M(Y_2) \cdots M(Y_n) e \\
&= p_\phi(Y_1, \ldots, Y_n | X_1 = K). \qquad (2.48)
\end{aligned}$$

Since $a, b \geq 0$ and $a + b = 1$, then from (2.46) and (2.48),

$$p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = i) = p_\phi(Y_1, \ldots, Y_n | X_1 = K), \qquad \text{for } i = K, K+1. \quad (2.49)$$

From (2.45) and (2.49),

$$\begin{aligned}
q_\phi(Y_1, \ldots, Y_n) &= \max_{1 \leq i \leq K} p_\phi(Y_1, \ldots, Y_n | X_1 = i) \\
&= \max_{1 \leq i \leq K+1} p_{\widehat{\phi}}(Y_1, \ldots, Y_n | X_1 = i) \\
&= q_{\widehat{\phi}}(Y_1, \ldots, Y_n)
\end{aligned}$$

■

By Lemma 2.5.9, we can define an order $\prec$ in $\{\Phi_K\}$.

Definition 2.5.14 *Define an* ***order*** $\prec$ *on* $\{\Phi_K\}$ *by*

$$\Phi_K \prec \Phi_L, \qquad K, L \in \boldsymbol{N},$$

if and only if for every $\phi \in \Phi_K$, *there is* $\widehat{\phi} \in \Phi_L$ *such that* $\phi \sim \widehat{\phi}$.

As a consequence of Lemma 2.5.9, Lemma 2.5.15 follows.

Lemma 2.5.15 *For every* $K \in \boldsymbol{N}$,

$$\Phi_K \prec \Phi_{K+1}.$$

From Lemma 2.5.15, the families of hidden Markov models represented by $\{\Phi_K\}$ are ***nested families*** as shown in Figure 1.

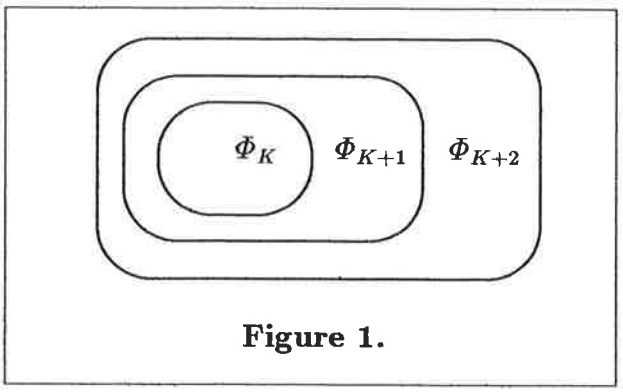

Figure 1.

2.6 A True Parameter

From section 2.5, it is known that representation, which generates the observed process of a hidden Markov model, is not unique. Our main interest is to find

the *simplest* one, that is, the one with *minimum size*. Such representation will be called a *true parameter* . One of our task is to identify a true parameter and its size. Therefore, the main aim of this section is to collect facts regarding the true parameter for our purpose.

We begin this section with a formal definition of a true parameter.

Definition 2.6.1 *Let* $\{(X_t, Y_t)\}$ *be a hidden Markov model with representation* $\phi \in \Phi$. *Let* $\phi^o = (K^o, A^o, \pi^o, \theta^o) \in \Phi$. *The hidden Markov model* $\{(X_t, Y_t)\}$ *is called to have a* ***true parameter*** ϕ^o *and an* ***order*** K^o, *if and only if*

(a). $\phi^o \sim \phi$.

(b). K^o *is* ***minimum***, *that is, there is no* $\widehat{\phi} \in \Phi_K$, *with* $K < K^o$, *such that* $\widehat{\phi} \sim \phi$.

A true parameter of a hidden Markov model $\{(X_t, Y_t)\}$ is not unique. By Lemma 2.5.3, for every permutation σ of $\{1, \ldots, K^o\}$,

$$\sigma(\phi^o) \sim \phi^o.$$

So $\sigma(\phi^o)$ is also a true parameter of the hidden Markov model $\{(X_t, Y_t)\}$.

As a straight consequence of Definition 2.6.1, we have the following lemma.

Lemma 2.6.2 *Let* $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ *be a true parameter of a hidden Markov model* $\{(X_t, Y_t)\}$. *Then there is no* $\phi \in \Phi_K$, *with* $K < K^o$ *such that* $\phi \sim \phi^o$.

The next two lemmas show some properties of true parameter which generates a stationary hidden Markov model.

Lemma 2.6.3 *Let $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$. If π^o is a stationary probability distribution of A^o, then*

$$\pi_i^o > 0, \qquad \text{for } i = 1, \ldots, K^o.$$

Proof :

Let N^o be the number of non-zero π_i^o, then $1 \leq N^o \leq K$. If $N^o < K^o$, then by Lemma 2.5.5, there is $\phi = (N^o, A, \pi, \theta) \in \Phi_{N^o}$, such that $\phi \sim \phi^o$, contradicting with Lemma 2.6.2. Thus, it must be $N^o = K^o$. ■

Lemma 2.6.4 *Let $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$, where π^o is a stationary probability distribution of A^o. Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, where $\phi \sim \phi^o$ and N be the number of non-zero π_i.*

(a). If $K = K^o$, then $N = K^o$.

(b). If $K > K^o$, then $N \geq K^o$.

Proof :

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, where $\phi \sim \phi^o$. By Lemma 2.6.2,

$$K \geq K^o.$$

Let N be the number of non-zero π_i, then

$$1 \leq N \leq K.$$

Suppose that $N < K^o$, since $\phi \sim \phi^o$, then π is a stationary probability distribution of A. By Lemma 2.5.5, there is $\widehat{\phi} = (N, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_N$, such that $\phi \sim \widehat{\phi}$, implying $\widehat{\phi} \sim \phi^o$, contradicting with Lemma 2.6.2. Thus, it must be

$$K^o \leq N \leq K. \tag{2.50}$$

If $K = K^o$, then by (2.50), $N = K^o$. If $K > K^o$, then $N \geq K^o$. ■

Corollary 2.6.5 *let $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$, where π^o is a stationary probability distribution of A^o. Let $\phi = (K^o, A, \pi, \theta) \in \Phi_{K^o}$. If $\phi \sim \phi^o$, then*

$$\pi_i > 0, \qquad for \quad i = 1, \ldots, K^o.$$

Proof :

This is part (a) of Lemma 2.6.4.

2.7 Stationary Hidden Markov Models

From Corollary 2.4.3, if the Markov chain of a hidden Markov model is stationary, then the observed process is also stationary. As a stationary process, the observed process has several properties, the most important is ergodicity . The ergodicity is essential for limit theorems which will be used later in Chapter 4. Therefore, finding sufficient conditions for the ergodicity of the observed process will be the focus of this section.

Let $\{(X_t, Y_t)\}$ be a hidden Markov model defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values on $\boldsymbol{S} \times \mathcal{Y}$, where $\boldsymbol{S} = \{1, \ldots, K\}$ and $\mathcal{Y} = \boldsymbol{R}$.

Let Λ be the set of all realizations $\{(x_t, y_t)\}$ of the hidden Markov model $\{(X_t, Y_t)\}$. Let $\mathcal{B}_\Lambda$ be the Borel σ-field of Λ. For each $t \in \boldsymbol{N}$, define mappings

$$\widetilde{X}_t : \Lambda \longrightarrow \boldsymbol{S},$$

by

$$\widetilde{X}_t(\lambda) = x_t$$

and

$$\widetilde{Y}_t : \Lambda \longrightarrow \mathcal{Y},$$

by

$$\widetilde{Y}_t(\lambda) = y_t,$$

for $\lambda = \{(x_t, y_t)\} \in \Lambda$. For $t \in \boldsymbol{N}$, $\widetilde{X}_t$, $\widetilde{Y}_t$ are *coordinate projections* on Λ.

The next lemma shows that there is a probability measure $\widetilde{P}$ defined on $\mathcal{B}_\Lambda$ such that the hidden Markov model $\{(X_t, Y_t)\}$ and the pair of processes $\{(\widetilde{X}_t, \widetilde{Y}_t)\}$ have the same law.

Lemma 2.7.1 *There exists a probability measure $\widetilde{P}$ defined on $\mathcal{B}_\Lambda$ such that the pair of coordinate projections $\{(\widetilde{X}_t, \widetilde{Y}_t)\}$ and the hidden Markov model $\{(X_t, Y_t)\}$ have the same law.*

Proof :

The idea of the proof comes from [11], page 511.

The hidden Markov model $\{(X_t, Y_t)\}$ is defined on the probability space $(\Omega, \mathcal{F}, P)$. For each $\omega \in \Omega$, let

$$\begin{aligned} X_t(\omega) &= x_t, \qquad t \in \boldsymbol{N} \\ Y_t(\omega) &= y_t, \qquad t \in \boldsymbol{N}. \end{aligned}$$

For each $k \in \boldsymbol{N}$ and distinct $t_1, \ldots, t_k \in \boldsymbol{N}$, let $\nu_{t_1,\ldots,t_k}$ be the joint distribution of $X_{t_1}, \ldots, X_{t_k}; Y_{t_1}, \ldots, Y_{t_k}$,

$$\nu_{t_1,\ldots,t_k}(A \times B) = P\Big\{(X_{t_1}, \ldots, X_{t_k}) \in A, (Y_{t_1}, \ldots, Y_{t_k}) \in B\Big\}, \tag{2.51}$$

for $A \in \mathcal{S}_k$ and $B \in \mathcal{B}_k$, where $\mathcal{S}_k$ and $\mathcal{B}_k$ are the Borel σ-field of $\boldsymbol{S}^k$ and $\mathcal{Y}^k$ respectively.

Define a mapping

$$\zeta : \Omega \to \Lambda,$$

by the requirement

$$\begin{aligned}\widetilde{X}_t(\zeta(\omega)) &= X_t(\omega) = x_t \\ \widetilde{Y}_t(\zeta(\omega)) &= Y_t(\omega) = y_t,\end{aligned}$$

for $\omega \in \Omega$ and $t \in \boldsymbol{N}$. Clearly,

$$\begin{aligned}&\zeta^{-1}\Big\{\lambda \in \Lambda : (\widetilde{X}_{t_1}(\lambda),\ldots,\widetilde{X}_{t_k}(\lambda)) \in A, (\widetilde{Y}_{t_1}(\lambda),\ldots,\widetilde{Y}_{t_k}(\lambda)) \in B\Big\} \\ &\quad = \Big\{\omega \in \Omega : (\widetilde{X}_{t_1}(\zeta(\omega)),\ldots,\widetilde{X}_{t_k}(\zeta(\omega))) \in A, (\widetilde{Y}_{t_1}(\zeta(\omega)),\ldots,\widetilde{Y}_{t_k}(\zeta(\omega))) \in B\Big\} \\ &\quad = \Big\{\omega \in \Omega : (X_{t_1}(\omega),\ldots,X_{t_k}(\omega)) \in A, (Y_{t_1}(\omega),\ldots,Y_{t_k}(\omega)) \in B\Big\} \\ &\quad \in \mathcal{F}, \qquad (2.52)\end{aligned}$$

if $A \in \mathcal{S}_k$ and $B \in \mathcal{B}_k$. Thus ζ is measurable.

Define probability measure $\tilde{P} = P\zeta^{-1}$ on $\mathcal{B}_\Lambda$, then from (2.51) and (2.52),

$$\begin{aligned}&\tilde{P}\Big\{\lambda \in \Lambda : (\widetilde{X}_{t_1}(\lambda),\ldots,\widetilde{X}_{t_k}(\lambda)) \in A, (\widetilde{Y}_{t_1}(\lambda),\ldots,\widetilde{Y}_{t_k}(\lambda)) \in B\Big\} \\ &\qquad = P\zeta^{-1}\Big\{\lambda \in \Lambda : (\widetilde{X}_{t_1}(\lambda),\ldots,\widetilde{X}_{t_k}(\lambda)) \in A, (\widetilde{Y}_{t_1}(\lambda),\ldots,\widetilde{Y}_{t_k}(\lambda)) \in B\Big\} \\ &\qquad = P\Big\{\omega \in \Omega : (X_{t_1}(\omega),\ldots,X_{t_k}(\omega)) \in A, (Y_{t_1}(\omega),\ldots,Y_{t_k}(\omega)) \in B\Big\} \\ &\qquad = \nu_{t_1,\ldots,t_k}(A \times B). \qquad (2.53)\end{aligned}$$

The equation (2.53) shows that $\{(\widetilde{X}_t, \widetilde{Y}_t)\}$, defined on $(\Lambda, \mathcal{B}_\Lambda, \tilde{P})$ also has finite dimensional distribution $\nu_{t_1,\ldots,t_k}$. Thus $\{(X_t, Y_t)\}$ and $\{(\widetilde{X}_t, \widetilde{Y}_t)\}$ have the same law. ■

Remarks 2.7.2 By Lemma 2.7.1, from now on, the hidden Markov model $\{(X_t, Y_t)\}$ may be considered as the pair of coordinate projection processes $\{(\widetilde{X}_t, \widetilde{Y}_t)\}$, defined on $(\Lambda, \mathcal{B}_\Lambda, \tilde{P})$. For convenience, we will drop the tilde.

Suppose that the Markov chain $\{X_t\}$ is stationary, then by Corollary 2.4.3, the hidden Markov model $\{(X_t, Y_t)\}$ is also stationary. We want to build a *past* for

the hidden Markov model $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ without loosing its stationarity. The problem is to find a pair of stochastic processes $\{(\bar{X}_t, \bar{Y}_t) : t \in \boldsymbol{Z}\}$ such that $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ and $\{(\bar{X}_t, \bar{Y}_t) : t \in \boldsymbol{N}\}$ have the same law.

Lemma 2.7.3 *There is a stationary process* $\{(\bar{X}_t, \bar{Y}_t)\}$ *indexed by* $t \in \boldsymbol{Z}$, *such that* $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ *and* $\{(\bar{X}_t, \bar{Y}_t) : t \in \boldsymbol{N}\}$ *have the same law.*

Proof :

The proof follows from [2], page 21.

Let $I = \{t_1, t_2, \ldots, t_k\} \in \boldsymbol{Z}$. For all r large enough, the integers $I_r = \{r + t_1, r + t_2, \ldots, r + t_k\} \subset \boldsymbol{N}$ and the joint law of $\{(X_t, Y_t) : t \in I_r\}$ is independent of r, since $\{(X_t, Y_t)\}$ is stationary. Let Π_I be this law. The family Π_I is consistent. Kolmogorov consistency theorem ([2], page 6) grants the existence of the process $\{(\bar{X}_t, \bar{Y}_t)\}$ indexed by $\boldsymbol{Z}$, such that for all I as above, Π_I is the joint law of $\{(\bar{X}_t, \bar{Y}_t) : t \in I\}$. Clearly $\{(X_t, Y_t) : t \in \boldsymbol{N}\}$ and $\{(\bar{X}_t, \bar{Y}_t) : t \in \boldsymbol{N}\}$ have the same law. ■

Remarks 2.7.4 Without loss of generality, by Lemma 2.7.3, now we have the stationary hidden Markov model $\{(X_t, Y_t) : t \in \boldsymbol{Z}\}$, defined on the probability space $(\Lambda, \mathcal{B}_\Lambda, P)$, where Λ is the set of realizations $\lambda = \{(x_t, y_t)\}$, $\mathcal{B}_\Lambda$ is the Borel σ-field of Λ and X_t, Y_t are coordinate projections defined on Λ.

If $z = \{z_t\}$ is a real sequence, let Tz denote the shifted sequence $\{z_{t+1}\}$. T is called the *shift operator*. A set $\mathcal{A}$ of real sequences is called *shift invariant*, when $Tz \in \mathcal{A}$ if and only if $z \in \mathcal{A}$. A stationary process $Z = \{Z_t\}$ is said to be *ergodic* if

$$P(Z \in \mathcal{A}) = 0 \qquad \text{or} \qquad 1,$$

whenever $\mathcal{A}$ is shift invariant.

From [53], page 33, a stationary and irreducible Markov chain is ergodic. Based on this, Leroux [34] derived the ergodicity of the observed process $\{Y_t\}$.

Lemma 2.7.5 (Leroux [34]) *If the Markov chain $\{X_t\}$ is stationary and irreducible, then the observed process $\{Y_t\}$ is stationary and ergodic.*

Proof :

Let $\mathcal{A}$ be a shift invariant set of sequences $y = \{y_t\}$ of possible realizations of $Y = \{Y_t\}$. It will be proved that

$$P(Y \in \mathcal{A}) = 0 \qquad \text{or} \qquad 1.$$

By the approximation theorem ([11], page 167), there is a subsequence $\{k'\}$ and cylinder set $\mathcal{A}_{k'}$ having form

$$\begin{aligned} \mathcal{A}_{k'} &= \left\{\lambda \in \Lambda : (Y_{-k'}(\lambda), \ldots, Y_{k'}(\lambda)) \in B_{2k'}\right\} \\ &= \left\{\lambda \in \Lambda : (y_{-k'}, \ldots, y_{k'}) \in B_{2k'}\right\}, \end{aligned}$$

where $B_{2k'} \in \mathcal{B}_{2k'}$, that is the Borel σ-field of $\mathcal{Y}^{2k'}$, such that $\forall k \in \boldsymbol{N}$,

$$P(Y \in \mathcal{A} \triangle \mathcal{A}_{k'}) < 2^{-k}. \tag{2.54}$$

Since Y is stationary and $\mathcal{A}$ is invariant,

$$\begin{aligned} P(Y \in \mathcal{A} \triangle \mathcal{A}_{k'}) &= P(T^{2k'}Y \in \mathcal{A} \triangle \mathcal{A}_{k'}) \\ &= P(Y \in \mathcal{A} \triangle T^{-2k'}\mathcal{A}_{k'}) \\ &= P(Y \in \mathcal{A} \triangle \tilde{\mathcal{A}}_{k'}), \end{aligned} \tag{2.55}$$

where

$$\begin{aligned} \tilde{\mathcal{A}}_{k'} &= T^{-2k'}\mathcal{A}_{k'} \\ &= \left\{\lambda \in \Lambda : (Y_{k'}(\lambda), \ldots, Y_{3k'}(\lambda)) \in B_{2k'}\right\} \\ &= \left\{\lambda \in \Lambda : (y_{k'}, \ldots, y_{3k'}) \in B_{2k'}\right\}. \end{aligned}$$

Let

$$\tilde{A} = \bigcap_{k\geq 1} \bigcup_{j\geq k} \tilde{A}_{j'},$$

then

$$\begin{aligned} A^c \bigcap \tilde{A} &= A^c \bigcap \left(\bigcap_{k\geq 1} \bigcup_{j\geq k} \tilde{A}_{j'} \right) \\ &= \bigcap_{k\geq 1} \bigcup_{j\geq k} \left(A^c \bigcap \tilde{A}_{j'} \right) \\ &= \limsup_{k'\to\infty} A^c \bigcap \tilde{A}_{k'} \end{aligned}$$

and

$$\begin{aligned} A \bigcap \tilde{A}^c &= A \bigcap \left(\bigcap_{k\geq 1} \bigcup_{j\geq k} \tilde{A}_{j'} \right)^c \\ &= A \bigcap \left(\bigcup_{k\geq 1} \bigcap_{j\geq k} \tilde{A}^c_{j'} \right) \\ &= \bigcup_{k\geq 1} \bigcap_{j\geq k} \left(A \bigcap \tilde{A}^c_{j'} \right) \\ &= \liminf_{k'\to\infty} A \bigcap \tilde{A}^c_{k'} . \end{aligned}$$

Hence,

$$\begin{aligned} A \,\Delta\, \tilde{A} &= \left(A \bigcap \tilde{A}^c \right) \bigcup \left(A^c \bigcap \tilde{A} \right) \\ &= \left(\liminf_{k'\to\infty} A \bigcap \tilde{A}^c_{k'} \right) \bigcup \left(\limsup_{k'\to\infty} A^c \bigcap \tilde{A}_{k'} \right) \\ &\subset \left(\limsup_{k'\to\infty} A \bigcap \tilde{A}^c_{k'} \right) \bigcup \left(\limsup_{k'\to\infty} A^c \bigcap \tilde{A}_{k'} \right) \\ &= \limsup_{k'\to\infty} \left(\left(A \bigcap \tilde{A}^c_{k'} \right) \bigcup \left(A^c \bigcap \tilde{A}_{k'} \right) \right) \\ &= \limsup_{k'\to\infty} \left(A \,\Delta\, \tilde{A}_{k'} \right) . \end{aligned} \tag{2.56}$$

From (2.54) and (2.55)

$$\begin{aligned} \sum_{k=1}^{\infty} P(Y \in A \,\Delta\, \tilde{A}_{k'}) &= \sum_{k=1}^{\infty} P(Y \in A \,\Delta\, A_{k'}) \\ &\leq \sum_{k=1}^{\infty} 2^{-k} \\ &= 1, \end{aligned}$$

so by Borrel Cantelli's Lemma,

$$0 \leq P(Y \in \mathcal{A} \triangle \tilde{\mathcal{A}}) \leq P(Y \in \limsup \mathcal{A} \triangle \tilde{\mathcal{A}}_{k'}) = 0,$$

implying

$$P(Y \in \mathcal{A} \triangle \tilde{\mathcal{A}}) = 0. \tag{2.57}$$

Since (2.57) holds, showing $P(Y \in \mathcal{A}) = 0$ or 1, is equivalent with proving that

$$P(Y \in \tilde{\mathcal{A}}) = 0 \quad \text{or} \quad 1.$$

By definition, $\tilde{\mathcal{A}} = \bigcap_{k \geq 1} \bigcup_{j \geq k} \tilde{\mathcal{A}}_{j'}$, so $\tilde{\mathcal{A}}$ is in the tail σ-field, that is, $\tilde{\mathcal{A}}$ is contained in the σ-field $\mathcal{F}_k = \sigma(Y_k, Y_{k+1}, \ldots)$, for all k. Since Y_t are conditionally independent given a realization $x = \{x_t\}$ of the underlying Markov chain $X = \{X_t\}$, then the zero-one law implies

$$P(Y \in \tilde{\mathcal{A}}|x) = 0 \quad \text{or} \quad 1.$$

Let

$$B = \left\{x : P(Y \in \tilde{\mathcal{A}}|x) = 1\right\},$$

so

$$\begin{aligned} P(Y \in \tilde{\mathcal{A}}) &= E\left[1_{Y \in \tilde{\mathcal{A}}}\right] \\ &= E\left[E[1_{Y \in \tilde{\mathcal{A}}}|x]\right] \\ &= E\left[P(Y \in \tilde{\mathcal{A}}|x)\right] \\ &= 0 + 1 \cdot P(X \in B) \\ &= P(X \in B). \end{aligned} \tag{2.58}$$

But, B is invariant, as

$$\begin{aligned} P(Y \in \tilde{\mathcal{A}}|x) &= P(TY \in \tilde{\mathcal{A}}|Tx) \\ &= P(Y \in \tilde{\mathcal{A}}|Tx). \end{aligned}$$

Since the Markov chain $\{X_t\}$ is stationary and irreducible, then $\{X_t\}$ is ergodic, implying

$$P(X \in B) = 0 \quad \text{or} \quad 1.$$

Hence, by (2.58),

$$P(Y \in \tilde{\mathcal{A}}) = 0 \quad \text{or} \quad 1.$$

■

Chapter 3

Identifiability

This chapter concentrates on studying the identifiability problem for hidden Markov models. The aim of this chapter is to find conditions on parameters A, π and θ of a representation $\phi = (K, A, \pi, \theta)$, which is equivalent to the true parameter $\phi^o = (K^o, A^o, \pi^o, \theta^o)$, such that parameters A, π and θ can be identified with parameters A^o, π^o and θ^o.

For convenience, this chapter will be divided into three sections. The first section, section 3.1, explains the identifiability problem for hidden Markov models and shows that this problem is *similar* to the identifiability problem for finite mixtures. In section 3.2, we collect results regarding the identifiability of finite mixtures. Finally, in the last section, using a slight modification, we derive the identifiability of hidden Markov models from the identifiability of finite mixtures. In this section, we also give the characteristics of the true parameter and parameters which are equivalent to it.

3.1 The Identifiability problem

Let $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$. By Lemma 2.6.2, if $\phi \in \Phi_K$ and $\phi \sim \phi^o$, then $K \geq K^o$. Moreover, by Lemma 2.5.9 and Lemma 2.5.3, there are infinitely many $\phi \in \Phi_K$, with $K > K_o$ and at least finitely many $\phi \in \Phi_K$, with $K = K^o$, such that $\phi \sim \phi^o$.

Let

$$\mathcal{T} = \Big\{\phi \in \bigcup_{K \geq K^o} \Phi_K \ : \ \phi \sim \phi^o\Big\}.$$

For parameter estimation purposes, every $\phi \in \mathcal{T}$ must be *identifiable*. This means that all parameters of ϕ can be identified with parameters of ϕ^o.

Let $\phi = (K^o, A, \pi, \theta) \in \mathcal{T}$. Since $\phi \sim \phi^o$, then by definition, for any $n \in \boldsymbol{N}$, the n-dimensional joint density functions of $Y_1, \ldots, Y_n$ under ϕ and ϕ^o are the same, that is,

$$p_{\phi^o}(y_1, \ldots, y_n) = p_\phi(y_1, \ldots, y_n), \tag{3.1}$$

for every $(y_1, \ldots, y_n) \in \mathcal{Y}^n$. Consider a special case of (3.1), when $n = 1$,

$$\begin{aligned} p_{\phi^o}(y_1) &= p_\phi(y_1) \\ \sum_{i=1}^{K^o} \pi_i^o f(y_1, \theta_i^o) &= \sum_{i=1}^{K^o} \pi_i f(y_1, \theta_i). \end{aligned} \tag{3.2}$$

From (3.2), we must be able to identify each (π_i, θ_i), with (π_j^o, θ_j^o). In other words, we must be able to show that for every $i = 1, \ldots, K^o$, there is j, $1 \leq j \leq K^o$, such that

$$\pi_i = \pi_j^o \qquad \text{and} \qquad \theta_i = \theta_j^o,$$

which can be written in the implication form,

$$\sum_{i=1}^{K^o} \pi_i f(y_1, \theta_i) = \sum_{i=1}^{K^o} \pi_i^o f(y_1, \theta_i^o) \quad \Longrightarrow \quad \begin{array}{l} \forall i = 1, \ldots K^o, \ \exists j, \ 1 \leq j \leq K^o \\ \text{such that } \pi_i = \pi_j^o \text{ and } \theta_i = \theta_j^o. \end{array} \tag{3.3}$$

Consider the following example.

Example 3.1.1 Suppose that from the observation, Y_1 has a density function as in Figure 2. Since we only observe the values of $\{Y_t\}$, then there is no way we can tell if the observation comes from

$$p(y_1) = \tfrac{1}{4}U(-1,1) + \tfrac{3}{4}U(-3,3)$$

or

$$p(y_1) = \tfrac{1}{2}U(-3,1) + \tfrac{1}{2}U(-1,3),$$

where $U(a,b)$ is a uniform distribution with range (a,b).

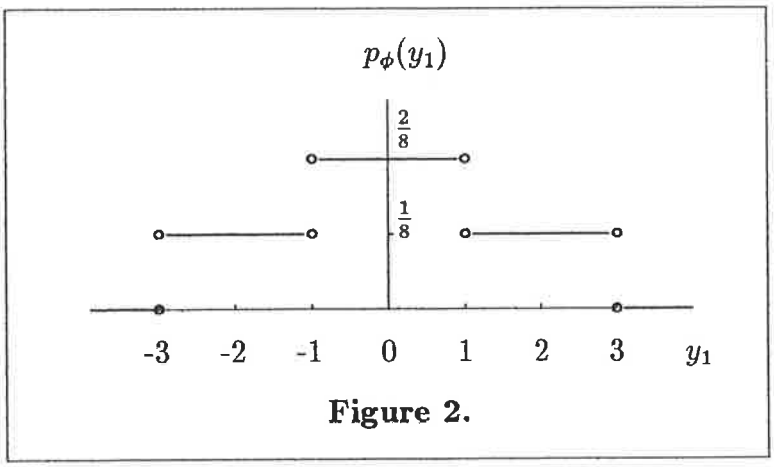

Figure 2.

The Example 3.1.1 above, shows that not every family of densities satisfies (3.3). Therefore, we have to find conditions on the family of densities $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta\}$, so that (3.3) holds.

Later, it can be shown, when π_i and π_i^o are positive for all $i = 1,\ldots,K^o$, (3.3) is a special case of identifiability criteria for finite mixtures. Using a slight modification, we can apply identifiability criteria for finite mixtures, which have already been established, to hidden Markov models, so it can be used to identify the true parameter ϕ^o.

3.2 Identifiability of Finite Mixtures

The purposes of this subsection are to collect results concerning identifiability of finite mixtures which are scattered in several journals and books, and to present them as coherent as possible for our use.

A good review of this subject can be found in [52], [20] and [37]. In particular, [52] provides extensive references.

This section begins with a formal definition of mixture distributions which is cited from [50].

Definition 3.2.1 *Let $\mathcal{F} = \{F(\cdot,\theta) : \theta \in \mathcal{B}\}$ be a family of one dimensional distribution functions, taking values on $\mathcal{Y}$, indexed by a point θ in a Borel subset $\mathcal{B}$ of Euclidean m-space $\boldsymbol{R}^m$, such that $F(\cdot,\cdot)$ is measurable in $\mathcal{Y} \times \mathcal{B}$. Let G be any distribution function such that the measure μ_G induced by G assigns measure 1 to $\mathcal{B}$. Then*

$$H(y) = \int_{\mathcal{B}} F(y,\theta)\, d\mu_G(\theta) = \int_{\mathcal{B}} F(y,\theta)\, dG(\theta), \qquad y \in \mathcal{Y} \tag{3.4}$$

*is called a **mixture distribution** and G is called the **mixing distribution**.*

Reference [20] gives a corresponding definition for a mixture density,

Definition 3.2.2 *Let $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \mathcal{B}\}$ be a family of one dimensional density functions, indexed by a point θ in a Borel subset $\mathcal{B}$ of Euclidean m-space $\boldsymbol{R}^m$, such that $f(\cdot,\cdot)$ is measurable in $\mathcal{Y} \times \mathcal{B}$. Let G be any distribution function such that the measure μ_G induced by G assigns measure 1 to $\mathcal{B}$. Then*

$$h(y) = \int_{\mathcal{B}} f(y,\theta)\, d\mu_G(\theta) = \int_{\mathcal{B}} f(y,\theta)\, dG(\theta), \qquad y \in \mathcal{Y} \tag{3.5}$$

*is called a **mixture density** and G is called the **mixing distribution**.*

Example 3.2.3 Let $\mathcal{F}$ be the family of uniform distribution functions $U(a,b)$ with range (a,b), where $a \leq b$. Let

$$G(a,b) = \tfrac{1}{2}\delta_{(-2,1)} + \tfrac{1}{2}\delta_{(-1,2)}, \qquad a,b \in \boldsymbol{R}, \qquad a \leq b,$$

where $\delta_{(a,b)}$ is Dirac distribution of a point mass at (a,b). Then,

$$H(y) = \int_{\{(a,b):a,b\in\mathbf{R},a\leq b\}} U(a,b)\, dG(a,b) = \tfrac{1}{2}U(-2,1) + \tfrac{1}{2}U(-1,2), \quad y \in \boldsymbol{R}$$

is a mixture distribution.

Example 3.2.4 Let $\mathcal{F}$ be the family of Poisson density function $f(\cdot,\theta)$, where

$$f(y,\theta) = \frac{e^{-\theta}\theta^y}{y!}, \qquad y = 0,1,\ldots \qquad \text{and} \qquad \theta \in (0,\infty).$$

Let

$$G(\theta) = e^{-\theta}, \qquad \theta \in (0,\infty).$$

Then from the simple recurrence $h(y) = \frac{1}{2}h(y-1)$, with $h(0) = \frac{1}{2}$,

$$h(y) = \int_0^\infty \frac{e^{-\theta}\theta^y}{y!} \cdot e^{-\theta}\, d\theta = 2^{-(y+1)}, \qquad y = 0,1,\ldots$$

is a mixture density

From Definitions 3.2.1 and 3.2.2, it can be seen that a mixture distribution and a mixture density can be derived from one another. Hence, it is enough to consider mixture distributions only. However, the results that we have for mixture distributions will also apply for mixture densities.

Let $\mathcal{G}$ denote the class of all such m-dimensional distribution functions G and $\mathcal{H}$ the induced class of mixtures H on $\mathcal{F}$. Teicher [50] defined the identifiability of $\mathcal{H}$ as follows.

Definition 3.2.5 *A class of mixture distributions $\mathcal{H}$ is said to be* ***identifiable*** *if and only if the equality of two representations*

$$\int_B F(y,\theta)\, dG(\theta) = \int_B F(y,\theta)\, d\hat{G}(\theta), \qquad \forall y \in \mathcal{Y}$$

implies $G = \hat{G}$.

Definition 3.2.1 given above is the general one, but most of our applications are concerned with a special type of mixture. This type is generated by the special case when G is discrete and assigns positive probability to only finite number of points, as in the Example 3.2.3.

Definition 3.2.6 *H is called a* ***finite mixture*** *if its mixing distribution G or rather the corresponding measure μ_G is discrete and assigns positive mass to only a finite number of points in $\mathcal{B}$. Thus the class $\widetilde{\mathcal{H}}$ of finite mixtures on $\mathcal{F}$ is defined by*

$$\widetilde{\mathcal{H}} = \left\{ H(\cdot) : H(\cdot) = \sum_{i=1}^{N} c_i F(\cdot,\theta_i),\ c_i > 0, \sum_{i=1}^{N} c_i = 1,\ F(\cdot,\theta_i) \in \mathcal{F}, N \in \boldsymbol{N} \right\}$$

that is, $\widetilde{\mathcal{H}}$ is the convex hull of $\mathcal{F}$.

Remarks 3.2.7 In every expression of finite mixture

$$H(\cdot) = \sum_{i=1}^{N} c_i F(\cdot,\theta_i),$$

$\theta_1, \ldots, \theta_N$ are assumed to be *distinct* members of Θ. The c_i and θ_i, $i = 1, \ldots, N$ will be called respectively the *coeffients* and *support points* of the finite mixture.

Applying Definition 3.2.5 to the class of finite mixtures, we have the identifiability criteria for finite mixtures. The following formal definition states that the class of finite mixtures $\widetilde{\mathcal{H}}$ is identifiable if and only if all members of $\widetilde{\mathcal{H}}$ are distinct.

Definition 3.2.8 *Let $\widetilde{\mathcal{H}}$ be the class of finite mixtures on $\mathcal{F}$. $\widetilde{\mathcal{H}}$ is* ***identifiable*** *if and only if*

$$\sum_{i=1}^{N} c_i F(\cdot, \theta_i) = \sum_{i=1}^{\widehat{N}} \widehat{c}_i F(\cdot, \widehat{\theta}_i)$$

implies $N = \widehat{N}$ and for each $i = 1, \ldots, N$, there is j, $1 \leq j \leq N$, such that $c_i = \widehat{c}_j$ and $\theta_i = \widehat{\theta}_j$.

Definition 3.2.8 can be stated in different way using Dirac distributions. To show this, the following lemma is needed.

Lemma 3.2.9 *Suppose that*

$$\sum_{i=1}^{N} c_i \delta_{\theta_i} = \sum_{i=1}^{\widehat{N}} \widehat{c}_i \delta_{\widehat{\theta}_i}, \tag{3.6}$$

for

$$\begin{aligned}
&c_i \geq 0, \qquad i = 1, \ldots, N, \qquad \sum_{i=1}^{N} c_i = 1 \\
&\widehat{c}_i \geq 0, \qquad i = 1, \ldots, \widehat{N}, \qquad \sum_{i=1}^{\widehat{N}} \widehat{c}_i = 1 \\
&\theta_i,\ \theta_j \in \Theta, \qquad , i = 1, \ldots, N, \qquad j = 1, \ldots, \widehat{N}
\end{aligned}$$

where δ_θ denotes the Dirac distribution of a point mass at θ.

(a). Suppose there are i, j, where $1 \leq i \leq N$ and $1 \leq j \leq \widehat{N}$, such that $\theta_i = \widehat{\theta}_j$. Let $D = \{k : \theta_k = \theta_i\}$ and $\widehat{D} = \{k : \widehat{\theta}_k = \widehat{\theta}_j\}$, then $\sum_{k \in D} c_k = \sum_{k \in \widehat{D}} \widehat{c}_k$.

(b). If $c_i > 0$, for some i, $1 \leq i \leq N$, then there is j, $1 \leq j \leq \widehat{N}$, such that $\theta_i = \widehat{\theta}_j$.

(c). If $c_i > 0$ and θ_i are distinct, for $i = 1, \ldots, N$, then $\widehat{N} \geq N$ and for every $i = 1, \ldots N$, there is j, $1 \leq j \leq \widehat{N}$ such that $\theta_i = \widehat{\theta}_j$.

(d). If $c_i > 0$ and θ_i are distinct for $i = 1, \ldots, N$ and $N = \widehat{N}$, then there is a permutation σ on $\{1, \ldots, N\}$ such that $c_i = \widehat{c}_{\sigma(i)}$ and $\theta_i = \widehat{\theta}_{\sigma(i)}$, for $i = 1, \ldots, N$.

Proof :

To prove (a), suppose that there are i, j, where $1 \leq i \leq N$ and $1 \leq j \leq \widehat{N}$, such that $\theta_i = \widehat{\theta}_j$. Let $D = \{k : \theta_k = \theta_i\}$ and $\widehat{D} = \{k : \hat{\theta}_k = \hat{\theta}_j\}$. Let ζ be a smooth real function defined on Θ such that :

$$\zeta(\theta) = \begin{cases} 1, & \text{if} \quad \theta = \theta_i \\ 0, & \text{if} \quad \theta \in \left(\{\theta_1, \ldots, \theta_N\} \cup \{\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{N}}\}\right) \setminus \{\theta_i\}. \end{cases} \tag{3.7}$$

By (3.6) and (3.7),

$$\begin{aligned} \sum_{k=1}^{N} \int_{\Theta} c_k \zeta(\theta) \delta_{\theta_k}(\theta) &= \sum_{k=1}^{\widehat{N}} \int_{\Theta} \widehat{c}_k \zeta(\theta) \delta_{\widehat{\theta}_k}(\theta) \\ \sum_{k=1}^{N} c_k \zeta(\theta_k) &= \sum_{k=1}^{\widehat{N}} \widehat{c}_k \zeta(\widehat{\theta}_k) \\ \sum_{k \in D} c_k &= \sum_{k \in \widehat{D}} \widehat{c}_k. \end{aligned} \tag{3.8}$$

For (b), let $c_i > 0$, for some i, $1 \leq i \leq N$. Suppose that $\theta_i \neq \widehat{\theta}_j$, for every $j = 1, \ldots, \widehat{N}$, then by (3.8)

$$\sum_{k \in D} c_k = 0,$$

implying $c_k = 0$ for all $k \in D$. Since $i \in D$, then $c_i = 0$, contradicting with $c_i > 0$. Thus, there must be j, $1 \leq j \leq \widehat{N}$, such that $\theta_i = \widehat{\theta}_j$

For (c) and (d), suppose that $c_i > 0$ and θ_i are distinct, for $i = 1, \ldots, N$. By part (b), for every $i = 1, \ldots, N$, there is j, $1 \leq j \leq \widehat{N}$, such that

$$\theta_i = \widehat{\theta}_j. \tag{3.9}$$

Since θ_i are all distinct for $i = 1, \ldots, N$, then it must be $\widehat{N} \geq N$. If $N = \widehat{N}$, the mapping $i \mapsto j$ is bijective. Let σ be the mapping and by (3.9),

$$\theta_i = \theta_{\sigma(i)}, \qquad \text{for} \quad i = 1, \ldots, N. \tag{3.10}$$

By (3.8) and (3.10),

$$c_i = c_{\sigma(i)}, \qquad \text{for} \quad i = 1, \ldots, N.$$

So the lemma is proved. ■

Lemma 3.2.10 *Let $\widetilde{\mathcal{H}}$ be the class of finite mixtures on $\mathcal{F}$. $\widetilde{\mathcal{H}}$ is identifiable if and only if*

$$\sum_{i=1}^{N} c_i F(\cdot,\theta_i) = \sum_{i=1}^{\widehat{N}} \widehat{c}_i F(\cdot,\widehat{\theta}_i) \qquad \Longrightarrow \qquad N=\widehat{N}, \quad \sum_{i=1}^{N} c_i \delta_{\theta_i} = \sum_{i=1}^{N} \widehat{c}_i \delta_{\widehat{\theta}_i}.$$

Proof :

To prove the lemma is enough to show that the necessary and sufficient condition for

$$\sum_{i=1}^{N} c_i \delta_{\theta_i} = \sum_{i=1}^{N} \widehat{c}_i \delta_{\widehat{\theta}_i}$$

where:

$$c_i, \widehat{c}_i > 0, \qquad i=1,\ldots,N, \qquad \sum_{i=1}^{N} c_1 = 1, \qquad \sum_{i=1}^{N} \widehat{c}_i = 1$$

θ_i are distinct for $i=1,\ldots,N$

$\widehat{\theta}_i$ are distinct for $i=1,\ldots,\widehat{N}$,

is for each $i=1,\ldots,N$, there is j, $1 \le j \le N$ such that $c_i = \widehat{c}_j$ and $\theta_i = \widehat{\theta}_j$. The sufficient condition is obvious and the necessity follows from part (d) of Lemma 3.2.9. ■

The following theorem is the first result concerning the sufficient conditions of the identifiability of finite mixtures.

Theorem 3.2.11 (Teicher [50]) *Let $\mathcal{F} = \{F(\cdot,\theta) : \theta \in \mathcal{B}\}$ be a family of one dimensional distribution functions, indexed by a point θ in a Borel subset $\mathcal{B}$ of Euclidean m-space $\boldsymbol{R}^m$ such that $F(\cdot,\cdot)$ is measurable in $\boldsymbol{R} \times \mathcal{B}$. Suppose there exists a transform*

$$M : F \mapsto \phi,$$

where ϕ is a real valued function defined on some S_ϕ, such that M is linear and injective. If there is a total ordering ($\preceq$) of $\mathcal{F}$ such that $F_1 \prec F_2$ implies

(a). $S_{\phi_1} \subset S_{\phi_2}$,

(b). The existence of some $t_1 \in \bar{S}_{\phi_1}$ (t_1 being independent of ϕ_2) such that :

$$\lim_{t \to t_1} \frac{\phi_2(t)}{\phi_1(t)} = 0,$$

then the class $\widetilde{\mathcal{H}}$ of all finite mixtures on $\mathcal{F}$ is identifiable.

Proof :

Suppose there are two finite sets of elements of $\mathcal{F}$, say $\mathcal{F}_1 = \{F(\cdot, \theta_i) : i = 1, \ldots, N\}$ and $\mathcal{F}_2 = \{F(\cdot, \widehat{\theta}_j) : j = 1, \ldots, \widehat{N}\}$ such that

$$\sum_{i=1}^{N} c_i F(y, \theta_i) = \sum_{j=1}^{\widehat{N}} \widehat{c}_j F(y, \widehat{\theta}_j), \qquad \forall y \in \boldsymbol{R}, \tag{3.11}$$

where

$$0 < c_i \leq 1, \qquad i = 1, \ldots, N, \qquad \sum_{i=1}^{N} c_i = 1$$

$$0 < \widehat{c}_i \leq 1, \qquad i = 1, \ldots, \widehat{N}, \qquad \sum_{i=1}^{\widehat{N}} \widehat{c}_i = 1.$$

Without loss of generality, index $\mathcal{F}_1$ and $\mathcal{F}_2$ such that for $i < j$,

$$F(\cdot, \theta_i) \prec F(\cdot, \theta_j) \qquad \text{and} \qquad F(\cdot, \widehat{\theta}_i) \prec F(\cdot, \widehat{\theta}_j).$$

If $F(\cdot, \theta_1) \neq F(\cdot, \widehat{\theta}_1)$, suppose without loss of generality that

$$F(\cdot, \theta_1) \prec F(\cdot, \widehat{\theta}_1),$$

then

$$F(\cdot, \theta_1) \prec F(\cdot, \widehat{\theta}_j), \qquad \text{for} \qquad j = 1, \ldots, \widehat{N}.$$

Apply the transform to (3.11). Then for $t \in T_1 = S_{\phi_1} \cap \{t : \phi_1(t) \neq 0\}$,

$$\begin{aligned} \sum_{i=1}^{N} c_i \phi_i(t) &= \sum_{j=1}^{\widehat{N}} \widehat{c}_j \widehat{\phi}_j(t) \\ c_1 + \sum_{i=2}^{N} c_i \frac{\phi_i(t)}{\phi_1(t)} &= \sum_{j=1}^{\widehat{N}} \widehat{c}_j \frac{\widehat{\phi}_j(t)}{\phi_1(t)}. \end{aligned}$$

Letting $t \to t_1$, through values in T_1, we have $c_1 = 0$, contradicting with the fact that $c_1 > 0$. Thus

$$F(\cdot, \theta_1) = F(\cdot, \widehat{\theta}_1)$$

and for any $t \in T_1$,

$$(c_1 - \widehat{c}_1) + \sum_{i=2}^{N} c_i \frac{\phi_i(t)}{\phi_1(t)} = \sum_{j=2}^{\widehat{N}} \widehat{c}_j \frac{\widehat{\phi}_j(t)}{\phi_1(t)}.$$

Again, letting $t \to t_1$, through values in T_1, we have $c_1 = \widehat{c}_1$.

So now,

$$\sum_{i=2}^{N} c_i F(y, \theta_i) = \sum_{j=2}^{\widehat{N}} \widehat{c}_j F(y, \widehat{\theta}_j), \qquad \forall y \in \boldsymbol{R}.$$

Repeating the same argument $\min(N, \widehat{N})$ times, we have

$$F(\cdot, \theta_i) = F(\cdot, \widehat{\theta}_i) \qquad \text{and} \qquad c_i = \widehat{c}_i,$$

for $i = 1, 2, \ldots, \min(N, \widehat{N})$.

If $N \neq \widehat{N}$, without loss of generality assume $N > \widehat{N}$. Then

$$\sum_{i=\widehat{N}+1}^{N} c_i F(y, \theta_i) = 0, \qquad \forall y \in \boldsymbol{R}.$$

Letting $y \to \infty$ in the above equation, implies $c_i = 0$, for $i = \widehat{N} + 1, \ldots, N$, in contradiction to the fact that $c_i > 0$, for $i = \widehat{N} + 1, \ldots, N$. Therefore

$$N = \widehat{N}, \qquad c_i = \widehat{c}_i \qquad \text{and} \qquad F(\cdot, \theta_i) = F(\cdot, \widehat{\theta}_i),$$

for $i = 1, 2, \ldots, N$. But $F(\cdot, \theta_i) = F(\cdot, \widehat{\theta}_i)$ imply $\theta_i = \widehat{\theta}_i$, for all $i = 1, 2, \ldots, N$. Then by definition $\widetilde{\boldsymbol{\mathcal{H}}}$ is identifiable. ■

An important application of Theorem 3.2.11 is the identifiability of the class of finite mixtures of one-dimensional normal distributions, the class of finite mixtures of one-dimensional gamma distributions and the class of finite mixtures of one-dimensional Poisson distributions.

Lemma 3.2.12 (Teicher [50]) *The class of all finite mixtures of one dimensional normal distributions is identifiable.*

Proof :

Let $\mathcal{N} = \{N(\cdot, \theta, \sigma^2) : \theta \in \boldsymbol{R}, \sigma > 0\}$ be a family of normal distributions, where $N(\cdot, \theta, \sigma^2)$ denotes a normal distribution with mean θ and variance σ^2.

Let $N(\cdot, \theta, \sigma^2) \in \mathcal{N}$ and define its (Laplace) transform by

$$\begin{aligned} \phi(t, \theta, \sigma^2) &= \int_{-\infty}^{\infty} N(y, \theta, \sigma^2)\, e^{-ty}\, dy \\ &= \int_{-\infty}^{\infty} \tfrac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2} e^{-ty}\, dy \\ &= e^{\frac{1}{2}\sigma^2 t^2 - \theta t}, \end{aligned}$$

where $t \in S_\phi = (-\infty, \infty)$.

Order $\mathcal{N}$ by

$$N_1 = N(\cdot, \theta_1, \sigma_1^2) \prec N(\cdot, \theta_2, \sigma_2^2) = N_2$$

if $\sigma_1 > \sigma_2$ or if $\sigma_1 = \sigma_2$, but $\theta_1 < \theta_2$.

Let $N_1 = N(\cdot, \theta_1, \sigma_1^2)$ and $N_2 = N(\cdot, \theta_2, \sigma_2^2)$ in $\mathcal{N}$ such that $N_1 \prec N_2$ and let $\phi_1(\cdot, \theta_1, \sigma_1^2)$ and $\phi_2(\cdot, \theta_2, \sigma_2^2)$ be their transforms respectively. $S_{\phi_1} = S_{\phi_2} = (-\infty, \infty)$. Take $t_1 = \infty$. If $\sigma_1 > \sigma_2$, then

$$\lim_{t\to\infty} \frac{\phi_2(t)}{\phi_1(t)} = \lim_{t\to\infty} e^{\left\{\frac{(\sigma_2^2 - \sigma_1^2)t^2}{2} - (\theta_2 - \theta_1)t\right\}} = 0,$$

since

$$\lim_{t\to\infty} e^{\left\{\frac{(\sigma_2^2 - \sigma_1^2)t^2}{2}\right\}} = 0.$$

If $\sigma_1 = \sigma_2$ and $\theta_1 < \theta_2$, then

$$\lim_{t\to\infty} \frac{\phi_2(t)}{\phi_1(t)} = \lim_{t\to\infty} e^{\{-(\theta_2 - \theta_1)t\}} = 0.$$

Then the identifiability of the class of finite mixtures of $\mathcal{N}$ follows from Theorem 3.2.11. ■

Lemma 3.2.13 (Teicher [50]) *The class of all finite mixtures of gamma distributions is identifiable*

Proof :

Let $\mathcal{F} = \{F(\cdot, \theta, \alpha) : \theta > 0, \alpha > 0\}$ be a family of gamma distrsibutions, where

$$F(y, \theta, \alpha) = \frac{\theta^\alpha}{\Gamma(\alpha)} \int_0^y x^{\alpha-1} e^{-\theta x}\, dx\,, \qquad \alpha > 0, \qquad \theta > 0.$$

For $F(\cdot, \theta, \alpha) \in \mathcal{F}$, define its (Laplace) transform as follows.

$$\begin{aligned} \phi(t, \theta, \alpha) &= \int_0^\infty \frac{\theta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x} e^{-tx}\, dx \\ &= \frac{\theta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{(\theta+t)^\alpha} \\ &= \left(1 + \frac{t}{\theta}\right)^{-\alpha}, \qquad \text{for} \quad t > -\theta. \end{aligned}$$

Order $\mathcal{F}$ by

$$F_1 = F(\cdot, \theta_1, \alpha_1) \prec F(\cdot, \theta_2, \alpha_2) = F_2$$

if $\theta_1 < \theta_2$ or $\theta_1 = \theta_2$ but $\alpha_1 > \alpha_2$.

Let $F_1 = F(\cdot, \theta_1, \alpha_1)$ and $F_2 = F(\cdot, \theta_2, \alpha_2)$ be any elements of $\mathcal{N}$, such that $F_1 \prec F_2$ and let $\phi_1(\cdot, \theta_1, \alpha_1)$ and $\phi_2(\cdot, \theta_2, \alpha_2)$ be their transforms respectively. Then $S_{\phi_1} = (-\theta_1, \infty) \subset S_{\phi_2} = (-\theta_2, \infty)$. If $\theta_1 < \theta_2$, then

$$\lim_{t \to -\theta_1} \frac{\phi_2(t, \theta_2, \alpha_2)}{\phi_1(t, \theta_1, \alpha_1)} = \lim_{t \to -\theta_1} \frac{\left(1 + \frac{t}{\theta_2}\right)^{-\alpha_2}}{\left(1 + \frac{t}{\theta_1}\right)^{-\alpha_1}} = 0,$$

since $\lim_{t \to -\theta_1} \left(1 + \frac{t}{\theta_1}\right)^{-\alpha_1} = \infty$. If $\theta_1 = \theta_2$, but $\alpha_1 > \alpha_2$, then

$$\lim_{t \to -\theta_1} \frac{\phi_2(t, \theta_2, \alpha_2)}{\phi_1(t, \theta_1, \alpha_1)} = \lim_{t \to -\theta_1} \frac{\left(1 + \frac{t}{\theta_2}\right)^{-\alpha_2}}{\left(1 + \frac{t}{\theta_1}\right)^{-\alpha_1}} = \lim_{t \to -\theta_1} \left(1 + \frac{t}{\theta_1}\right)^{\alpha_1 - \alpha_2} = 0,$$

since $\alpha_1 - \alpha_2 > 0$. Then by Theorem 3.2.11, the class of all finite mixtures of gamma distributions is identifiable. ■

Corollary 3.2.14 *The class of all finite mixtures of negative exponential distribution is identifiable*

Proof :

Let $\mathcal{F} = \{F(\cdot, \theta) : \theta > 0\}$ be the family of exponential distributions, where

$$F(y, \theta) = \int_0^y \theta e^{-\theta x}\, dx\,, \qquad \theta > 0.$$

It can be seen that $\mathcal{F}$ is a special case of the family of gamma distributions in Lemma 3.2.13 with $\alpha = 1$, then the result follows. ■

Lemma 3.2.15 *The class of all finite mixtures of Poisson distributions is identifiable.*

Proof :

Let $\mathcal{F} = \{f(\cdot, \theta) : \theta > 0\}$ be the family of Poisson distribution with mean θ, where

$$f(y, \theta) = \frac{e^{-\theta}\theta^y}{y!}, \qquad y = 0, 1, 2, \ldots \qquad \text{and} \qquad \theta > 0.$$

For $f(\cdot, \theta) \in \mathcal{F}$, define its transform as follows

$$\begin{aligned}
\phi(t, \theta) &= \sum_{y=0}^{\infty} e^{ty} \cdot \frac{e^{-\theta}\theta^y}{y!} \\
&= e^{-\theta}\left(\sum_{y=0}^{\infty} \frac{\theta^y}{y!} e^{ty}\right) \\
&= e^{-\theta}\left(\sum_{y=0}^{\infty} \frac{(e^t\theta)^y}{y!}\right) \\
&= e^{e^t\theta} \cdot e^{-\theta} \\
&= e^{\theta(e^t - 1)}, \qquad t \in \boldsymbol{R}.
\end{aligned}$$

Order $\mathcal{F}$ by

$$f_1 = f(\cdot, \theta_1) \prec f(\cdot, \theta_2) = f_2 \qquad \text{if} \qquad \theta_1 > \theta_2.$$

Let $f_1 = f(\cdot, \theta_1)$ and $f_2 = f(\cdot, \theta_2)$ be in $\mathcal{F}$, such that $f_1 \prec f_2$ and let $\phi_1(\cdot, \theta_1)$ and $\phi_2(\cdot, \theta_2)$ be their transforms respectively. Then $S_{\phi_1} = S_{\phi_2} = (-\infty, \infty)$ and

$$\begin{aligned}\lim_{t\to\infty} \frac{\phi_2(t)}{\phi_1(t)} &= \lim_{t\to\infty} \frac{e^{\theta_2(e^t-1)}}{e^{\theta_1(e^t-1)}} \\ &= \lim_{t\to\infty} e^{\theta_2(e^t-1)-\theta_1(e^t-1)} \\ &= \lim_{t\to\infty} e^{(\theta_2-\theta_1)(e^t-1)} \\ &= 0\,,\end{aligned}$$

since $\theta_2 - \theta_1 < 0$. Then the result follows from Theorem 3.2.11. ■

Yakowitz and Spragins [54] extended Teicher's results of identifiability to include multidimensional distribution functions. Let

$$\mathcal{F}_n = \{F(\cdot, \theta) : \theta \in \mathcal{B}\}$$

be a family of n-dimensional distribution functions taking values in $\boldsymbol{R}^n$ indexed by a point θ in a borel subset $\mathcal{B}$ of Euclidean m-space $\boldsymbol{R}^m$, such that $F(\cdot, \cdot)$ is measurable in $\boldsymbol{R}^n \times \mathcal{B}$.

Let $\widetilde{\mathcal{H}}_n$ be the class of all finite mixtures on $\mathcal{F}_n$ defined as in Definition 3.2.6, that is,

$$\widetilde{\mathcal{H}}_n = \left\{ H(\cdot) : H(\cdot) = \sum_{i=1}^{N} c_i F(\cdot, \theta_i), c_i > 0, \sum_{i=1}^{N} c_i = 1, F(\cdot, \theta_i) \in \mathcal{F}_n, N \in \boldsymbol{N} \right\}.$$

As in one dimensional case, for every finite mixture

$$H(\cdot) = \sum_{i=1}^{N} c_i F(\cdot, \theta_i),$$

$\theta_1, \ldots, \theta_N$ are assumed to be *distinct.*

Theorem 3.2.16 (Yakowitz and Spragins [54]) *A necessary and sufficient condition that the class $\widetilde{\mathcal{H}}_n$ of all finite mixtures on $\mathcal{F}_n$ be identifiable is that $\mathcal{F}_n$ be a linearly independent set over the field of real numbers.*

Proof :

Necessity :

Suppose $\mathcal{F}_n$ is not a linearly independent set over the field of real numbers. Let

$$\sum_{i=1}^{N} a_i F(y,\theta_i) = 0, \qquad \forall y \in \boldsymbol{R}^n,$$

where $a_i \in \boldsymbol{R}$, $i = 1, 2, \ldots, N$, be a linear relation in $\mathcal{F}_n$.

Assume the a_i's are subscripted so that

$$a_i < 0 \quad \Longleftrightarrow \quad i \leq M\,.$$

Then

$$\sum_{i=1}^{M} |a_i| F(y,\theta_i) = \sum_{i=M+1}^{N} |a_i| F(y,\theta_i), \qquad \forall y \in \boldsymbol{R}^n\,. \tag{3.12}$$

By letting $y \to \infty$ in (3.12), where $\infty = (\infty, \infty, \ldots, \infty)$,

$$\sum_{i=1}^{M} |a_i| = \sum_{i=M+1}^{N} |a_i|. \tag{3.13}$$

Let

$$b = \sum_{i=1}^{M} |a_i| \qquad \text{and} \qquad c_i = \frac{|a_i|}{b}, \quad i = 1, \ldots, N. \tag{3.14}$$

By (3.13) and (3.14),

$$b > 0$$

$$c_i > 0, \qquad i = 1, \ldots, M, \qquad \sum_{i=1}^{M} c_i = 1$$

$$c_i \geq 0, \qquad i = M+1, \ldots, N, \qquad \sum_{i=M+1}^{N} c_i = 1.$$

Then

$$\sum_{i=1}^{M} c_i F(\cdot,\theta_i) = \sum_{i=M+1}^{N} c_i F(\cdot,\theta_i)$$

are two distinct representations of the same finite mixture and therefore $\widetilde{\mathcal{H}}_n$ can not be identifiable.

Sufficiency :

Let $\langle \mathcal{F}_n \rangle$ be the span of $\mathcal{F}_n$. If $\mathcal{F}_n$ is linearly independent, then it is a bases for

$\langle \mathcal{F}_n \rangle$. Two distinct representations of the same mixture implied by the non-identifiability of $\widetilde{\mathcal{H}}_n \subset \langle \mathcal{F}_n \rangle$ would contradict the uniqueness of representation of bases. ■

From the properties of isomorphisms, $\mathcal{F}_n$ is linearly independent if and only if the image of the isomorphism is linearly independent in the image space, the corollary below follows.

Corollary 3.2.17 *The class $\widetilde{\mathcal{H}}_n$ of all finite mixtures of the family $\mathcal{F}_n$ is identifiable if and only if the image $\mathcal{F}_n$ under any vector isomorphism of $\langle \mathcal{F}_n \rangle$ be linearly independent in the image space.*

The most important result of the application of Theorem 3.2.16 is the identifiability of the family of finite mixtures of multidimensional normal distributions.

Lemma 3.2.18 (Yakowitz and Spragins [54]) *The family of n dimensional normal distribution functions generates identifiable finite mixtures.*

Proof :

Let

$$\mathcal{N} = \{N(\cdot, \theta, \Lambda) : \theta \in \boldsymbol{R}^n \text{ and } \Lambda \text{ is an } n \times n \text{ positive definite matrix } \}$$

be a family of n-dimensional normal distribution with mean vector θ and covariance matrix Λ.

For $N(\cdot, \theta, \Lambda) \in \mathcal{N}$, let $M(\cdot, \theta, \Lambda)$ be its moment generating function defined by

$$\begin{aligned} M(t, \theta, \Lambda) &= \int_{\boldsymbol{R}^n} \exp\{-t^T y\} N(y, \theta, \Lambda)\, dy \\ &= \exp\left\{\theta^T t + \tfrac{1}{2} t^T \Lambda t\right\}, \qquad t \in \boldsymbol{R}^n. \end{aligned}$$

Note that θ, t and y are n-dimensional column vectors. It is clear that the mapping $N \mapsto M$ is an isomorphism.

Suppose that $\mathcal{N}$ does not generate identifiable finite mixtures. Then by Corollary 3.2.17, the set

$$\mathcal{M} = \{M(\cdot, \theta, \Lambda) : \theta \in \boldsymbol{R}^n \text{ and } \Lambda \text{ is an } n \times n \text{ positive definite matrix}\}$$

is a linearly dependent set over $\boldsymbol{R}$. There are $M \geq 1$, $d_i \in \boldsymbol{R}$, $d_i \neq 0$, $i = 1, \ldots, M$ and distinct pairs (θ_i, Λ_i), $i = 1, \ldots, M$ such that

$$\sum_{i=1}^{M} d_i \exp\left\{\theta_i^T t + \tfrac{1}{2} t^T \Lambda_i t\right\} = 0, \qquad t \in \boldsymbol{R}^n. \tag{3.15}$$

Consider a special case of (3.15), when $t = \alpha s$, for a fixed vector s and $\alpha \in \boldsymbol{R}$. Then (3.15) becomes

$$\sum_{i=1}^{M} d_i \exp\left\{\alpha(\theta_i^T s) + \tfrac{1}{2}\alpha^2 (s^T \Lambda_i s)\right\} = 0, \qquad \alpha \in \boldsymbol{R}. \tag{3.16}$$

If all θ_i, $i = 1, \ldots, M$ are identical, then all Λ_i, $i = 1, \ldots, M$ are distinct. For $i \neq j$, $1 \leq i, j \leq M$,

$$s^T \Lambda_i s = s^T \Lambda_j s \quad \Longleftrightarrow \quad s \in \left\{z : z^T(\Lambda_i - \Lambda_j) z = 0\right\}.$$

So if

$$s \notin \bigcup_{\substack{i \neq j \\ 1 \leq i, j \leq M}} \left\{z : z^T(\Lambda_i - \Lambda_j) z = 0\right\},$$

then $s^T \Lambda_i s$, for $i = 1, \ldots, M$, are all distinct positive real numbers, implying the pairs of real numbers $(\theta_i^T s, s^t \Lambda_i s)$, for $i = 1, \ldots, M$, are distinct.

Otherwise, Suppose without loss of generality that $\theta_1, \ldots, \theta_k$, for some k, $k < M$, are the only distinct vectors among $\theta_1, \ldots, \theta_M$. Then for $i \neq j$, $1 \leq i, j \leq k$,

$$\theta_i^T s = \theta_j^T s \quad \Longleftrightarrow \quad s \in \left\{z : (\theta_i^T - \theta_j^T) z = 0\right\}.$$

So if

$$s \notin \bigcup_{\substack{i \neq j \\ 1 \leq i,j \leq k}} \left\{ z : (\theta_i^T - \theta_j^T) z = 0 \right\},$$

then the real numbers $\theta_i^T s$, for $i = 1, \ldots, k$, are distinct. Since the (θ_i, Λ_i), $i = 1, \ldots, M$ are all distinct, then the Λ_i, $i = k+1, \ldots, M$ with the same θ_i, are different. So if

$$s \notin \bigcup_{\substack{i \neq j \\ k+1 \leq i,j \leq M}} \left\{ z : z^T (\Lambda_i - \Lambda_j) z = 0 \right\},$$

then the real numbers $s^T \Lambda_i s$, for $i = k+1, \ldots, M$, are distinct. Consequently, for

$$s \notin \bigcup_{\substack{i \neq j \\ 1 \leq i,j \leq k}} \left\{ z : (\theta_i^T - \theta_j^T) z = 0 \right\} \bigcup \bigcup_{\substack{i \neq j \\ k+1 \leq i,j \leq M}} \left\{ z : z^T (\Lambda_i - \Lambda_j) z = 0 \right\},$$

the pairs of real numbers $(\theta_i^T s, s^T \Lambda_i s)$, for $i = 1, \ldots, M$, are distinct.

Therefore, for such a choice of s, the equation (3.16) asserts that there is $M \geq 1$, $d_i \in \boldsymbol{R}$, $d_i \neq 0$, $i = 1, \ldots, M$ and distinct pairs (μ_i, σ_i^2), where

$$\mu_i = \theta_i^T s \qquad \text{and} \qquad \sigma_i^2 = s^T \Lambda_i s, \qquad \text{for} \qquad i = 1, \ldots, M,$$

such that

$$\sum_{i=1}^{M} d_i \exp\left\{ \mu_i \alpha + \tfrac{1}{2} \sigma_i^2 \alpha^2 \right\} = 0, \qquad \alpha \in \boldsymbol{R}.$$

Corollary 3.2.17 implies that the class of finite mixtures of one dimensional normal distributions is not identifiable, contrary to Lemma 3.2.12. ■

Teicher's result which is concerned with mixtures of product measures will be presented next. Teicher [51] stated that the identifiability of mixture distributions can be carried over to mixtures of product distributions.

Recall that for any $k \in \boldsymbol{N}$, we have defined

$$\mathcal{F}_k = \{ F(\cdot, \alpha) : \alpha \in \mathcal{B} \},$$

as a family of k-dimensional distribution functions indexed by a point α in a Borel subset $\mathcal{B}$ of Euclidean m-space $\boldsymbol{R}^m$, such that $F(\cdot,\cdot)$ is measurable in $\boldsymbol{R}^k \times \mathcal{B}$.

Define for every $k, n \in \boldsymbol{N}$,

$$\mathcal{F}^*_{k,n} = \left\{ F^*(\cdot,\alpha) : F^*(\cdot,\alpha) = \prod_{i=1}^{n} F(\cdot,\alpha_i),\ F(\cdot,\alpha_i) \in \mathcal{F}_k,\ i = 1,\ldots,n \right\}. \tag{3.17}$$

Notice that in (3.17), $F(\cdot,\cdot)$ is defined on $\boldsymbol{R}^k \times \mathcal{B}$ and $F^*(\cdot,\cdot)$ is defined on $\boldsymbol{R}^{kn} \times \mathcal{B}^n$.

Theorem 3.2.19 (Teicher [51]) *If the class of all mixtures on $\mathcal{F}_1$ is identifiable, then for every $n > 1$, the class of mixtures on $\mathcal{F}^*_{1,n}$ is identifiable. Conversely, if for some $n > 1$, the class of all mixtures on $\mathcal{F}^*_{1,n}$ is identifiable, then the class of mixtures on $\mathcal{F}_1$ is identifiable.*

Proof :

To prove the second part, suppose that the class of all mixtures on $\mathcal{F}^*_{1,n}$ is identifiable for some $n > 1$. Let $F(\cdot,\alpha) \in \mathcal{F}_1$. If

$$\int_{\mathcal{B}} F(y,\alpha)\, dG(\alpha) = \int_{\mathcal{B}} F(y,\alpha)\, d\widehat{G}(\alpha),$$

then multiplying both sides by $\prod_{i=1}^{n-1} F(y_i,\alpha_o)$, $\alpha_o \in \mathcal{B}$, necessarily,

$$I_{\alpha_o} \times \cdots \times I_{\alpha_o} \times G = I_{\alpha_o} \times \cdots \times I_{\alpha_o} \times \widehat{G},$$

where I_{α_o} is a characteristic function $\chi_{[\alpha_o,\infty)}$. Hence by the hypothesis $G = \widehat{G}$.

To prove the first part of the theorem, the mathematical induction will be used. Suppose the class of mixtures on $\mathcal{F}_1$ is identifiable and also suppose the class of mixtures on $\mathcal{F}^*_{1,n}$ is identifiable, for fixed but arbitrary n. It will be shown that the class of mixtures on $\mathcal{F}^*_{1,(n+1)}$ is also identifiable.

Suppose that for $F^* \in \mathcal{F}_{1,n}$ and $F \in \mathcal{F}_1$,

$$\int F^*(x,\alpha)F(y,\beta)\,dG(\alpha,\beta) = \int F^*(x,\alpha)F(y,\beta)\,d\widehat{G}(\alpha,\beta)\,. \tag{3.18}$$

Let $G_2(\beta)$ and $\widehat{G}_2(\beta)$ denote the marginal distribution of β corresponding to G and $\widehat{G}$. Let $G(\alpha|\beta)$, $\widehat{G}(\alpha|\beta)$ denote versions of the conditional probabilities, such that, for each β, $G(\alpha|\beta)$ and $\widehat{G}(\alpha|\beta)$ are distribution functions in the variable α, and for each α, $G(\alpha|\beta)$ and $\widehat{G}(\alpha|\beta)$ are equal almost every where to measurable functions of β. Then (3.18) may be rewritten as,

$$\int F(y,\beta)H(x,\beta)\,dG_2(\beta) = \int F(y,\beta)\widehat{H}(x,\beta)\,d\widehat{G}_2(\beta)\,, \tag{3.19}$$

where

$$H(x,\beta) = \int F^*(x,\alpha)\,d_\alpha G(\alpha|\beta) \tag{3.20}$$

$$\widehat{H}(x,\beta) = \int F^*(x,\alpha)\,d_\alpha \widehat{G}(\alpha|\beta). \tag{3.21}$$

In turn, (3.19) may be expressed as

$$\int F(y,\beta)\,dJ_x(\beta) = \int F(y,\beta)\,d\widehat{J}_x(\beta), \tag{3.22}$$

where

$$J_x(\beta) = \int_{-\infty}^{\beta} H(x,\gamma)\,dG_2(\gamma) \le G_2(\beta) \tag{3.23}$$

$$\widehat{J}_x(\beta) = \int_{-\infty}^{\beta} \widehat{H}(x,\gamma)\,d\widehat{G}_2(\gamma) \le \widehat{G}_2(\beta), \tag{3.24}$$

as $H(x,\gamma) \le 1$ and $\widehat{H}(x,\gamma) \le 1$. Dominated convergence applied to (3.22), to ensure that

$$J_x(\infty) = \widehat{J}_x(\infty),$$

since this common value is finite by (3.23) and (3.24). Thus from (3.22) and since the class of mixture on $\mathcal{F}_1$ is identifiable by the hypothesis, then

$$J_x = \widehat{J}_x.$$

Or equivalently from (3.23) and (3.24),

$$\int_{-\infty}^{\beta} H(x,\gamma)\, dG_2(\gamma) = \int_{-\infty}^{\beta} \widehat{H}(x,\gamma)\, d\widehat{G}_2(\gamma). \tag{3.25}$$

On the other hand, letting $x \to \infty$ in (3.19) and since

$$\lim_{x\to\infty} H(x,\beta) = \lim_{x\to\infty} \int F^*(x,\alpha)\, d_\alpha G(\alpha|\beta) = 1 \tag{3.26}$$

$$\lim_{x\to\infty} \widehat{H}(x,\beta) = \lim_{x\to\infty} \int F^*(x,\alpha)\, d_\alpha \widehat{G}(\alpha|\beta) = 1 \tag{3.27}$$

by monotone convergence theorem, then (3.19) gives

$$\int F(y,\beta)\, dG_2(\beta) = \int F(y,\beta)\, d\widehat{G}_2(\beta). \tag{3.28}$$

By the hypothesis,

$$G_2(\beta) = \widehat{G}_2(\beta). \tag{3.29}$$

However, (3.29) in conjunction with (3.25) necessitates

$$H(x,\beta) = \widehat{H}(x,\beta)\,, \tag{3.30}$$

for almost all β. Equation (3.30) together with (3.20) and (3.21), gives

$$\int F^*(x,\alpha)\, d_\alpha G(\alpha|\beta) = \int F^*(x,\alpha)\, d_\alpha \widehat{G}(\alpha|\beta). \tag{3.31}$$

By the induction hypothesis, that is, the class of mixtures on $\mathcal{F}^*_{1,n}$ is identifiable, and (3.31) imply

$$G(\alpha|\beta) = \widehat{G}(\alpha|\beta) \tag{3.32}$$

Finally, combining (3.29) and (3.32) we have

$$G(\alpha,\beta) = G(\alpha|\beta)\, G(\beta) = \widehat{G}(\alpha|\beta)\, \widehat{G}(\beta) = \widehat{G}(\alpha,\beta).$$

So that the class of mixtures on $\mathcal{F}^*_{1,(n+1)}$ is identifiable. ■

Since Theorem 3.2.20 applies for general mixtures, then we have the following theorem for finite mixtures.

Theorem 3.2.20 *If the class of all finite mixtures on $\mathcal{F}_1$ is identifiable, then for every $n > 1$, the class of finite mixtures on $\mathcal{F}^*_{1,n}$ is identifiable. Conversely, if for some $n > 1$, the class of all finite mixtures on $\mathcal{F}^*_{1,n}$ is identifiable, then the class of finite mixtures on $\mathcal{F}_1$ is identifiable.*

Analogous result hold, with $\mathcal{F}_1$ and $\mathcal{F}^*_{1,n}$ is replaced by $\mathcal{F}_k$ and $\mathcal{F}^*_{k,n}$, where $k > 1$.

3.3 Identifiability of Hidden Markov Models

Let $\{(X_t, Y_t)\}$ be a hidden Markov model with representation $\phi = (K, A, \pi, \theta)$ $\in \Phi_K$. From section 2.5, the parameters A, π and θ satisfy :

$$A = (\alpha_{ij}), \qquad \alpha_{ij} \geq 0, \qquad \sum_{j=1}^{K} \alpha_{ij} = 1, \qquad i, j = 1, \ldots, K$$

$$\pi = (\pi_i), \qquad \pi_i \geq 0, \qquad i = 1, \ldots, K, \qquad \sum_{i=1}^{K} \pi_i = 1$$

$$\theta = (\theta_i)^T, \qquad \theta_i \in \Theta, \qquad i = 1, \ldots, K.$$

Notice that $\theta_1, \theta_2, \ldots, \theta_K$ *need not all to be distinct.*

Under ϕ, for any $n \in \boldsymbol{N}$, the joint density function of $Y_1, \ldots, Y_n$ is

$$p_\phi(y_1, \ldots, y_n) = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} f(y_1, \theta_{x_1}) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t} f(y_t, \theta_{x_t}). \tag{3.33}$$

Let

$$Q_\phi = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t} \delta_{(\theta_{x_1}, \ldots, \theta_{x_n})}, \tag{3.34}$$

then (3.33) can be written as

$$p_\phi(y_1, y_2, \ldots, y_n) = \int_{\Theta^n} f(y_1, \zeta_1) \cdots f(y_n, \zeta_n) Q_\phi(d\zeta_1, \ldots, d\zeta_n). \tag{3.35}$$

Equations (3.33), (3.34) and (3.35) assert that, for $n = 1$, p_ϕ is a *finite mixture* with *non-negative coefficients* $\pi_1, \ldots, \pi_K$ and *may not be distinct support points* $\theta_1, \ldots, \theta_K$. For $n \geq 2$, p_ϕ is a *finite mixture of product measures* with *non-negative coefficients* $\left(\pi_{x_1} \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}\right)$ and *may not be distinct support points* $(\theta_{x_1}, \ldots, \theta_{x_n})$, for $x_1, \ldots, x_n \in \{1, \ldots, K\}$.

In order to apply the identifiability of finite mixtures to hidden Markov models, Definition 3.2.8 has to be relaxed to allow the above possibilities.

Definition 3.3.1 *Let $\mathcal{F} = \{F(\cdot, \theta) : \theta \in \Theta\}$ be a family of one dimensional distribution functions, defined on $\mathcal{Y}$, indexed by $\theta \in \Theta$. Let*

$$\widehat{\mathcal{H}} = \left\{ H(\cdot) \; : \; H(\cdot) = \sum_{i=1}^{K} c_i F(\cdot, \theta_i), \; c_i \geq 0, \; \sum_{i=1}^{K} c_i = 0, \right. \\ \left. \theta_i \in \Theta, \; i = 1, 2, \ldots, K, \; K \in \boldsymbol{N} \right\}. \quad (3.36)$$

*$\widehat{\mathcal{H}}$ is said to be **identifiable** if and only if*

$$\sum_{i=1}^{K} c_i F(\cdot, \theta_i) = \sum_{i=1}^{\widehat{K}} \widehat{c}_i F(\cdot, \widehat{\theta}_i) \quad \Longrightarrow \quad \sum_{i=1}^{K} c_i \delta_{\theta_i} = \sum_{i=1}^{\widehat{K}} \widehat{c}_i \delta_{\widehat{\theta}_i}. \quad (3.37)$$

where δ_θ denotes the Dirac distribution of a point mass at θ.

Remarks 3.3.2 In every expression of

$$H(\cdot) = \sum_{i=1}^{K} c_i F(\cdot, \theta_i) \in \widehat{\mathcal{H}},$$

the parameters $\theta_1, \ldots, \theta_K$ *need not all to be distinct.*

Next lemma shows the relation between Definition 3.2.8 and Definition 3.3.1.

Lemma 3.3.3 *$\widehat{\mathcal{H}}$ is identifiable according to Definition 3.3.1 if and only if $\widetilde{\mathcal{H}}$ is identifiable according to Definition 3.2.8.*

Proof :

Necessity :

Assume that $\widehat{\mathcal{H}}$ is identifiable according to Definition 3.3.1. We will prove that $\widetilde{\mathcal{H}}$ is identifiable according to Definition 3.2.8. Suppose

$$\sum_{i=1}^{K} c_i F(\cdot, \theta_i) = \sum_{i=1}^{\widehat{K}} \widehat{c}_i F(\cdot, \widehat{\theta}_i), \tag{3.38}$$

where :

$$c_i > 0, \qquad i = 1, \ldots, K, \qquad \sum_{i=1}^{K} c_i = 1$$

$$\widehat{c}_i > 0, \qquad i = 1, \ldots, \widehat{K} \qquad \sum_{i=1}^{\widehat{K}} \widehat{c}_i = 1$$

$$\theta_i \quad \text{are distinct for} \quad i = 1, \ldots, K$$

$$\widehat{\theta}_i \quad \text{are distinct for} \quad i = 1, \ldots, \widehat{K}.$$

By Definition 3.3.1, equation (3.38) implies

$$\sum_{i=1}^{K} c_i \delta_{\theta_i} = \sum_{i=1}^{\widehat{K}} \widehat{c}_i \delta_{\widehat{\theta}_i}. \tag{3.39}$$

Since $c_i > 0$ and θ_i are distinct for $i = 1, \ldots, K$, then by part (c) of Lemma 3.2.9, $\widehat{K} \geq K$. On the otherhand, since $\widehat{c}_i > 0$ and $\widehat{\theta}_i$ are distinct for $i = 1, \ldots, \widehat{K}$, then by part (c) of Lemma 3.2.9, we also have $K \geq \widehat{K}$. Hence, we have $K = \widehat{K}$ and by (3.39),

$$\sum_{i=1}^{K} c_i \delta_{\theta_i} = \sum_{i=1}^{K} \widehat{c}_i \delta_{\widehat{\theta}_i}.$$

By Lemma 3.2.10, $\widetilde{\mathcal{H}}$ is identifiable according to Definition 3.2.8.

Sufficiency :

Assume that $\widetilde{\mathcal{H}}$ is identifiable according to Definition 3.2.8. We will prove that $\widehat{\mathcal{H}}$ is identifiable according to Definition 3.3.1. Suppose

$$\sum_{i=1}^{K} c_i F(\cdot, \theta_i) = \sum_{i=1}^{\widehat{K}} \widehat{c}_i F(\cdot, \widehat{\theta}_i), \tag{3.40}$$

where :

$$c_i \geq 0, \qquad i = 1, \ldots, K, \qquad \sum_{i=1}^{K} c_i = 1$$

$$\widehat{c}_i \geq 0, \qquad i = 1, \ldots, \widehat{K}, \qquad \sum_{i=1}^{\widehat{K}} \widehat{c}_i = 1$$

θ_i need not all to be distinct, for $i = 1, 2, \ldots, K$

$\widehat{\theta}_i$ need not all to be distinct, for $i = 1, \ldots, \widehat{K}$.

Let

$$F_+ = \{i \ : \ c_i > 0, \ i = 1, \ldots, K\}$$
$$\widehat{F}_+ = \{i \ : \ \widehat{c}_i > 0, \ i = 1, \ldots, \widehat{K}\}.$$

Let r be the number of distinct θ_i, $i \in F_+$, and $\widehat{r}$ be the number of distinct $\widehat{\theta}_i$, $i \in \widehat{F}_+$. Without loss of generality, suppose that $\theta_1, \ldots, \theta_r$ are distinct and also $\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{r}}$. Let

$$R_i = \{j \ : \ j \in F_+, \ \theta_j = \theta_i\}, \qquad i = 1, \ldots, r$$
$$\widehat{R}_i = \{j \ : \ j \in \widehat{F}_+, \ \widehat{\theta}_j = \widehat{\theta}_i\}, \qquad i = 1, \ldots, \widehat{r}.$$

Equation (3.40) then can be written as

$$\sum_{i=1}^{r} a_i F(\cdot, \theta_i) = \sum_{i=1}^{\widehat{r}} \widehat{a}_i F(\cdot, \widehat{\theta}_i), \tag{3.41}$$

where

$$a_i = \sum_{j \in R_i} c_j \qquad \text{and} \qquad \widehat{a}_i = \sum_{j \in \widehat{R}_i} \widehat{c}_j.$$

Since $a_i > 0$ and θ_i are distinct for $i = 1, \ldots, r$, and $\widehat{a}_i > 0$ and $\widehat{\theta}_i$ are distinct for $i = 1, \ldots, \widehat{r}$, then by Definition 3.2.8, equation (3.41) implies $r = \widehat{r}$ and

$$\sum_{i=1}^{r} a_i \delta_{\theta_i} = \sum_{i=1}^{r} \widehat{a}_i \delta_{\widehat{\theta}_i}. \tag{3.42}$$

But this is equivalent with

$$\begin{aligned} \sum_{i=1}^{r} \sum_{j \in R_i} c_j \delta_{\theta_j} &= \sum_{i=1}^{\widehat{r}} \sum_{j \in \widehat{R}_i} \widehat{c}_j \delta_{\widehat{\theta}_j} \\ \sum_{i \in F_+} c_i \delta_{\theta_i} &= \sum_{i \in \widehat{F}_+} \widehat{c}_i \delta_{\widehat{\theta}_i}. \end{aligned} \tag{3.43}$$

Since $c_i = 0$, for $i \notin F_+$ and also $\widehat{c}_i = 0$, for $i \notin \widehat{F}_+$, then by (3.43)

$$\sum_{i=1}^{K} c_i \delta_{\theta_i} = \sum_{i=1}^{\widehat{K}} \widehat{c}_i \delta_{\widehat{\theta}_i}.$$

Hence, $\widehat{\mathcal{H}}$ is identifiable according to Definition 3.3.1. ■

Remarks 3.3.4 As a consequence of Lemma 3.3.3, all the results of identifiability in section 3.2 are now applicable for hidden Markov models. So from now on, when we say $\widehat{\mathcal{H}}$ is identifiable, we mean it in the sense of Definition 3.3.1.

The following lemma gives a necessary and sufficient condition for two representations of a hidden Markov model to be equivalent.

Lemma 3.3.5 *Suppose that $\widehat{\mathcal{H}}$ is identifiable. Let $\phi, \widehat{\phi} \in \Phi$, where $\phi = (K, \pi, A, \theta)$ and $\widehat{\phi} = (\widehat{K}, \widehat{A}, \widehat{\pi}, \widehat{\theta})$. Then $\phi \sim \widehat{\phi}$ if and only if for every $n \in \boldsymbol{N}$,*

$$\sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} \prod_{t=2}^{n} \alpha_{x_{t-1},x_t} \delta_{(\theta_{x_1},\ldots,\theta_{x_n})}$$
$$= \sum_{x_1=1}^{\widehat{K}} \cdots \sum_{x_n=1}^{\widehat{K}} \widehat{\pi}_{x_1} \prod_{t=2}^{n} \widehat{\alpha}_{x_{t-1},x_t} \delta_{(\widehat{\theta}_{x_1},\ldots,\widehat{\theta}_{x_n})}. \tag{3.44}$$

Proof :

The sufficient is obvious. We will prove the necessity. Suppose that $\phi \sim \widehat{\phi}$, then for any $n \in \boldsymbol{N}$, the n- dimensional joint density functions of $Y_1, \ldots, Y_n$ under ϕ and $\widehat{\phi}$ are the same, that is

$$p_\phi(y_1, \ldots, y_n) = p_{\widehat{\phi}}(y_1, \ldots, y_n),$$

which can be written as

$$\sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} f(y_1, \theta_{x_1}) \prod_{t=2}^{n} \alpha_{x_{t-1},x_t} f(y_t, \theta_{x_t})$$
$$= \sum_{x_1=1}^{\widehat{K}} \cdots \sum_{x_n=1}^{\widehat{K}} \widehat{\pi}_{x_1} f(y_1, \widehat{\theta}_{x_1}) \prod_{t=2}^{n} \widehat{\alpha}_{x_{t-1},x_t} f(y_t, \widehat{\theta}_{x_t}). \tag{3.45}$$

Since $\widehat{\mathcal{H}}$ is identifiable, by Theorem 3.2.20, equation (3.45) implies (3.44). ■

In particular, Lemma 3.3.5 gives necessary and sufficient condition for representations to be equivalent with the true parameter.

Corollary 3.3.6 *Let $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$. Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, then $\phi \sim \phi^o$, if and only if for every $n \in \boldsymbol{N}$,*

$$\sum_{x_1=1}^{K^o} \cdots \sum_{x_n=1}^{K^o} \pi^o_{x_1} \prod_{t=2}^{n} \alpha^o_{x_{t-1},x_t} \delta_{(\theta^o_{x_1},\ldots,\theta^o_{x_n})} = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1} \prod_{t=2}^{n} \alpha_{x_{t-1},x_t} \delta_{(\theta_{x_1},\ldots,\theta_{x_n})}.$$

Next lemma gives an example of such parameter that can be a true parameter of a hidden Markov model.

Lemma 3.3.7 *Assume that $\widehat{\mathcal{H}}$ is identifiable. Let $\phi = (K, A, \pi, \theta) \in \Phi_K$ satisfying :*

(a). $\pi_i > 0$, for $i = 1, \ldots, K$

(b). θ_i are distinct for $i = 1, \ldots, K$,

then the size K is minimum, that is, no $\widehat{\phi} \in \Phi_{\widehat{K}}$, with $\widehat{K} < K$, such that $\widehat{\phi} \sim \phi$.

Proof :

Suppose the size K is not minimum, then there is $\widehat{\phi} = (\widehat{K}, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{\widehat{K}}$, with $\widehat{K} < K$, such that $\widehat{\phi} \sim \phi$. Since $\widehat{\phi} \sim \phi$, by Lemma 3.3.5,

$$\sum_{i=1}^{K} \pi_i \delta_{\theta_i} = \sum_{i=1}^{\widehat{K}} \widehat{\pi}_i \delta_{\widehat{\theta}_i}. \tag{3.46}$$

Since $\pi_i > 0$ and θ_i are distinct for $i = 1, \ldots, K$, then by part (c) of Lemma 3.2.9, $\widehat{K} \geq K$, contradicting with the fact that $\widehat{K} < K$. Hence, the size K must be minimum. ■

The following three lemmas give characteristics of representations which equivalent with the true parameter.

Lemma 3.3.8 *Assume that $\widehat{\mathcal{H}}$ is identifiable. Let $\phi^o = (K^o, A^o, \pi^o, \theta^o) \in \Phi_{K^o}$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$ satisfying :*
(a). $\pi_i^o > 0$, for $i = 1, \ldots, K^o$
(b). θ_i^o are distinct for $i = 1, \ldots, K^o$.
Let $\widehat{\phi} = (K^o, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K^o}$, then $\widehat{\phi} \sim \phi^o$ if and only if $\widehat{\phi} = \sigma(\phi^o)$, for some permutation σ of $\{1, \ldots, K^o\}$.

Proof :
Let $\widehat{\phi} = (K^o, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K^o}$. If $\widehat{\phi} = \sigma(\phi^o)$, by Lemma 2.5.3, it is clear that $\widehat{\phi} \sim \phi^o$.

Now suppose that $\widehat{\phi} \sim \phi^o$. By Corollary 3.3.6,

$$\sum_{i=1}^{K^o} \pi_i^o \delta_{\theta_i^o} = \sum_{i=1}^{K^o} \widehat{\pi}_i \delta_{\widehat{\theta}_i}.$$

Since $\pi_i^o > 0$ and θ_i^o are distinct for $i = 1, \ldots, K^o$, then by part (d) of Lemma 3.2.9, there is a permutation σ of $\{1, \ldots, K^o\}$ such that

$$\widehat{\theta}_i = \theta_{\sigma(i)}^o, \qquad i = 1, \ldots, K^o \tag{3.47}$$

and

$$\widehat{\pi}_i = \pi_{\sigma(i)}^o, \qquad i = 1, \ldots, K^o. \tag{3.48}$$

Also by Corollary 3.3.6,

$$\sum_{i=1}^{K^o} \sum_{j=1}^{K^o} \pi_i^o \alpha_{ij}^o \delta_{(\theta_i^o, \theta_j^o)} = \sum_{i=1}^{K^o} \sum_{j=1}^{K^o} \widehat{\pi}_i \widehat{\alpha}_{ij} \delta_{(\widehat{\theta}_i, \widehat{\theta}_j)}. \tag{3.49}$$

Since θ_i^o are distinct for $i = 1, \ldots, K^o$, then by part (a) of Lemma 3.2.9 and (3.47), we have from (3.49),

$$\widehat{\pi}_i \widehat{\alpha}_{ij} = \pi_{\sigma(i)}^o \alpha_{\sigma(i),\sigma(j)}^o, \qquad i, j = 1, \ldots, K^o. \tag{3.50}$$

As $\pi_i^o > 0$, for $i = 1, \ldots, K^o$, then by (3.48) and (3.50)

$$\hat{\alpha}_{ij} = \alpha^o_{\sigma(i),\sigma(j)}, \qquad i, j = 1, \ldots, K^o. \tag{3.51}$$

Hence from (3.47), (3.48) and (3.51), $\hat{\phi} = \sigma(\phi)$. ■

Lemma 3.3.9 *Assume that $\widehat{\mathcal{H}}$ is identifiable. Let $\phi^o = (K^o, A^o, \pi^o, \theta^o) \in \Phi_{K^o}$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$ satisfying :*

(a). $\pi_i^o > 0$, for $i = 1, \ldots, K^o$

(b). θ_i^o are distinct for $i = 1, \ldots, K^o$.

Let $\hat{\phi} = (K^o + 1, \hat{A}, \hat{\pi}, \hat{\theta}) \in \Phi_{K^o+1}$, where $\hat{\phi} \sim \phi^o$. Then after suitable permutation of indices, parameters of $\hat{\phi}$ have one of the following forms :

(a).

$$\hat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & \alpha^o_{1,K^o} & 0 \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & \alpha^o_{2,K^o} & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & \alpha^o_{K^o-1,K^o} & 0 \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \alpha^o_{K^o,K^o} & 0 \\ \hat{\alpha}_{K^o+1,1} & \cdots & \hat{\alpha}_{K^o+1,K^o-1} & \hat{\alpha}_{K^o+1,K^o} & \hat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

$$\hat{\pi} = (\pi^o_1, \cdots, \pi^o_{K^o-1}, \pi^o_{K^o}, 0)$$

$$\hat{\theta} = (\theta^o_1, \cdots, \theta^o_{K^o-1}, \theta^o_{K^o}, \gamma)^T$$

where :

$$\hat{\alpha}_{K^o+1,i} \geq 0, \qquad i = 1, \ldots, K^o + 1, \qquad \sum_{i=1}^{K^o+1} \hat{\alpha}_{K^o+1,i} = 1$$

$$\gamma \in \Theta, \qquad \gamma \neq \theta^o_i, \qquad i = 1, \ldots, K^o$$

(b).

$$\hat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & \hat{\alpha}_{1,K^o} & \hat{\alpha}_{1,k^o+1} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & \hat{\alpha}_{2,K^o} & \hat{\alpha}_{2,K^o+1} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & \hat{\alpha}_{K^o-1,K^o} & \hat{\alpha}_{K^o-1,K^o+1} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \hat{\alpha}_{K^o,K^o} & \hat{\alpha}_{K^o,K^o+1} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \hat{\alpha}_{K^o+1,K^o} & \hat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

$$\hat{\pi} = (\pi^o_1, \cdots, \pi^o_{K^o-1}, a\pi^o_{K^o}, b\pi^o_{K^o})$$
$$\hat{\theta} = (\theta^o_1, \cdots, \theta^o_{K^o-1}, \theta^o_{K^o}, \theta^o_{K^o})^T$$

where :

$$\hat{\alpha}_{i,K^o} + \hat{\alpha}_{i,K^o+1} = \alpha^o_{i,K^o}, \qquad i = 1, \ldots, K^o$$
$$\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1} = \alpha^o_{K^o,K^o}$$
$$a, b \geq 0, \qquad a + b = 1$$
$$a\hat{\alpha}_{i,K^o+1} \neq b\hat{\alpha}_{i,K^o}, \qquad \textit{for some } i,\ 1 \leq i \leq K^o - 1.$$

(c).

$$\hat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & a\alpha^o_{1,K^o} & b\alpha^o_{1,k^o} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & a\alpha^o_{2,K^o} & b\alpha^o_{2,K^o} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & a\alpha^o_{K^o-1,K^o} & b\alpha^o_{K^o-1,K^o} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \hat{\alpha}_{K^o,K^o} & \hat{\alpha}_{K^o,K^o+1} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \hat{\alpha}_{K^o+1,K^o} & \hat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

$$\hat{\pi} = (\pi^o_1, \cdots, \pi^o_{K^o-1}, a\pi^o_{K^o}, b\pi^o_{K^o})$$
$$\hat{\theta} = (\theta^o_1, \cdots, \theta^o_{K^o-1}, \theta^o_{K^o}, \theta^o_{K^o})^T$$

where :

$$a, b \geq 0, \qquad a + b = 1$$
$$\hat{\alpha}_{i,K^o} + \hat{\alpha}_{i,K^o+1} = \alpha^o_{K^o,K^o}, \qquad i = K^o, K^o + 1$$
$$a(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}) \neq b(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o})$$

(d).

$$\hat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & a\alpha^o_{1,K^o} & b\alpha^o_{1,k^o} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & a\alpha^o_{2,K^o} & b\alpha^o_{2,K^o} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & a\alpha^o_{K^o-1,K^o} & b\alpha^o_{K^o-1,K^o} \\ \hat{\alpha}_{K^o,1} & \cdots & \hat{\alpha}_{K^o,K^o-1} & \hat{\alpha}_{K^o,K^o} & \hat{\alpha}_{K^o,K^o+1} \\ \hat{\alpha}_{K^o+1,1} & \cdots & \hat{\alpha}_{K^o+1,K^o-1} & \hat{\alpha}_{K^o+1,K^o} & \hat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

$$\begin{aligned}\widehat{\pi} &= (\pi^o_1, \cdots, \pi^o_{K^o-1}, a\pi^o_{K^o}, b\pi^o_{K^o}) \\ \widehat{\theta} &= (\theta^o_1, \cdots, \theta^o_{K^o-1}, \theta^o_{K^o}, \theta^o_{K^o})^T\end{aligned}$$

where :

$$\begin{aligned}&a, b \geq 0, \qquad a+b=1 \\ &a\widehat{\alpha}_{K^o,j} + b\widehat{\alpha}_{K^o+1,j} = \alpha^o_{K^o,j}, \qquad j = 1, \ldots, K^o - 1 \\ &a\widehat{\alpha}_{K^o,K^o} + b\widehat{\alpha}_{K^o+1,K^o} = a\alpha^o_{K^o,K^o} \\ &a\widehat{\alpha}_{K^o,K^o+1} + b\widehat{\alpha}_{K^o+1,K^o+1} = b\alpha^o_{K^o,K^o}.\end{aligned}$$

Remarks 3.3.10 Notice that the representation $\widehat{\phi} = (K^o+1, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K^o+1}$, where

$$\widehat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & a\alpha^o_{1,K^o} & b\alpha^o_{1,k^o} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & a\alpha^o_{2,K^o} & b\alpha^o_{2,K^o} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & a\alpha^o_{K^o-1,K^o} & b\alpha^o_{K^o-1,K^o} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & a\alpha^o_{K^o,K^o} & b\alpha^o_{K^o,K^o} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & a\alpha^o_{K^o,K^o} & b\alpha^o_{K^o,K^o} \end{pmatrix}$$

$$\begin{aligned}\widehat{\pi} &= (\pi^o_1, \cdots, \pi^o_{K^o-1}, a\pi^o_{K^o}, b\pi^o_{K^o}) \\ \widehat{\theta} &= (\theta^o_1, \cdots, \theta^o_{K^o-1}, \theta^o_{K^o}, \theta^o_{K^o})^T\end{aligned}$$

as in the proof of Lemma 2.5.9 can be clasified having (d)-form.

Proof :

Let $\widehat{\phi} = (K^o+1, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K^o+1}$ and $\widehat{\phi} \sim \phi^o$. As $\widehat{\phi} \sim \phi^o$, by Corollary 3.3.6,

$$\sum_{i=1}^{K^o} \pi^o_i \delta_{\theta^o_i} = \sum_{i=1}^{K^o+1} \widehat{\pi}_i \delta_{\widehat{\theta}_i}. \tag{3.52}$$

Since $\pi^o_i > 0$ and θ^o_i are distinct for $i = 1, \ldots, K^o$, then by part (c) of Lemma 3.2.9, for each $i = 1, \ldots, K^o$, there is j, $1 \leq j \leq K^o + 1$, such that $\widehat{\theta}_j = \theta^o_i$.

Without loss of generality, suppose that

$$\widehat{\theta}_i = \theta_i^o, \qquad \text{for} \quad i = 1, \ldots, K^o.$$

There are two possibilities for $\widehat{\theta}_{K^o+1}$, it may be equal to γ, for some $\gamma \in \Theta$, where $\gamma \neq \theta_i^o$, $i = 1, \ldots, K^o$, or it is equal to one of the θ_i^o. Without loss of generality, suppose we have two possibilities

$$\widehat{\theta}_{K^o+1} = \gamma \qquad \text{or} \qquad \widehat{\theta}_{K^o+1} = \theta_{K^o}^o.$$

If

$$\widehat{\theta}_i = \theta_i^o, \qquad \text{for} \qquad i = 1, \ldots, K^o \qquad \text{and} \qquad \widehat{\theta}_{K^o+1} = \gamma, \tag{3.53}$$

then by (3.52) and part (a) and (b) of Lemma 3.2.9,

$$\widehat{\pi}_i = \pi_i^o, \qquad \text{for} \qquad i = 1, \ldots, K^o \qquad \text{and} \qquad \widehat{\pi}_{K^o+1} = 0. \tag{3.54}$$

If

$$\widehat{\theta}_i = \theta_i^o, \qquad \text{for} \qquad i = 1, \ldots, K^o \qquad \text{and} \qquad \widehat{\theta}_{K^o+1} = \theta_{K^o}^o, \tag{3.55}$$

then by (3.52) and part (a) of Lemma 3.2.9,

$$\widehat{\pi}_i = \pi_i^o, \quad \text{for} \quad i = 1, \ldots, K^o - 1 \quad \text{and} \quad \widehat{\pi}_{K^o} = a\pi_{K^o}^o, \quad \widehat{\pi}_{K^o+1} = b\pi_{K^o}^o, \tag{3.56}$$

for some $a, b \in \boldsymbol{R}$, $\quad a, b \geq 0, \quad a + b = 1.$

Consider the first case. By (3.53) and (3.54), parameters of $\widehat{\phi}$ are of the form :

$$\begin{aligned} &\widehat{A} = (\widehat{\alpha}_{ij}) \\ &\widehat{\pi} = (\pi_1^o, \ldots, \pi_{K^o-1}^o, \pi_{K^o}^o, 0) \\ &\widehat{\theta} = (\theta_1^o, \ldots, \theta_{K^o-1}^o, \theta_{K^o}^o, \gamma)^T. \end{aligned} \tag{3.57}$$

To identify $\widehat{A}$, consider the following equation implied by Corollary 3.3.6,

$$\sum_{i=1}^{K^o} \sum_{j=1}^{K^o} \pi_i^o \alpha_{ij}^o \delta_{(\theta_i^o, \theta_j^o)} = \sum_{i=1}^{K^o+1} \sum_{j=1}^{K^o+1} \widehat{\pi}_i \widehat{\alpha}_{ij} \delta_{(\widehat{\theta}_i, \widehat{\theta}_j)}. \tag{3.58}$$

But by (3.57), the RHS of (3.58) can be written as,

$$\sum_{i=1}^{K^o+1}\sum_{j=1}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{ij}\delta_{(\widehat{\theta}_i,\widehat{\theta}_j)} = \sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\pi_i^o\widehat{\alpha}_{ij}\delta_{(\theta_i^o,\theta_j^o)} + \sum_{i=1}^{K^o}\pi_i\widehat{\alpha}_{i,K^o+1}\delta_{(\theta_i^o,\gamma)}. \tag{3.59}$$

Since $\pi_i^o > 0$ and θ_i^o are distinct for $i = 1,\ldots,K^o$ and $\gamma \neq \theta_i^o$, for $i = 1,\ldots,K^o$, then by Lemma 3.2.9, we have from (3.58) and (3.59)

$$\begin{aligned}
&\widehat{\alpha}_{ij} = \alpha_{ij}^o, \qquad i,j = 1,\ldots,K^o\\
&\widehat{\alpha}_{i,K^o+1} = 0, \qquad i = 1,\ldots,K^o.
\end{aligned}$$

So in this case, $\widehat{A}$ is of the form

$$\widehat{A} = \begin{pmatrix}
\alpha_{11}^o & \cdots & \alpha_{1,K^o-1}^o & \alpha_{1,K^o}^o & 0\\
\alpha_{21}^o & \cdots & \alpha_{2,K^o-1}^o & \alpha_{2,K^o}^o & 0\\
\vdots & \ddots & \vdots & \vdots & \vdots\\
\alpha_{K^o-1,1}^o & \cdots & \alpha_{K^o-1,K^o-1}^o & \alpha_{K^o-1,K^o}^o & 0\\
\alpha_{K^o,1}^o & \cdots & \alpha_{K^o,K^o-1}^o & \alpha_{K^o,K^o}^o & 0\\
\widehat{\alpha}_{K^o+1,1} & \cdots & \widehat{\alpha}_{K^o+1,K^o-1} & \widehat{\alpha}_{K^o+1,K^o} & \widehat{\alpha}_{K^o+1,K^o+1}
\end{pmatrix}$$

where

$$\widehat{\alpha}_{K^o+1,i} \geq 0, \qquad i = 1,\ldots,K^o+1, \qquad \sum_{i=1}^{K^o+1}\widehat{\alpha}_{K^o+1,i} = 1.$$

So (a) follows.

Now consider the second case. From (3.55), parameters of $\widehat{\phi}$ have forms :

$$\begin{aligned}
&\widehat{A} = (\widehat{\alpha}_{ij})\\
&\widehat{\pi} = (\pi_1^o,\ldots,\pi_{K^o-1}^o, a\pi_{K^o}^o, b\pi_{K^o}^o)\\
&\widehat{\theta} = (\theta_1^o,\ldots,\theta_{K^o-1}^o,\theta_{K^o}^o,\theta_{K^o}^o)^T,
\end{aligned} \tag{3.60}$$

where $a,b \in \boldsymbol{R}$, $a,b \geq 0$, $a+b = 1$.

To identify $\widehat{A}$, consider the following equation implied by Corollary 3.3.6,

$$\sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\pi_i^o\alpha_{ij}^o\delta_{(\theta_i^o,\theta_j^o)} = \sum_{i=1}^{K^o+1}\sum_{j=1}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{ij}\delta_{(\widehat{\theta}_i,\widehat{\theta}_j)}. \tag{3.61}$$

By (3.60), (3.61) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\pi_i^o\alpha_{ij}^o\delta_{(\theta_i^o,\theta_j^o)}+\sum_{i=1}^{K^o-1}\pi_i^o\alpha_{i,K^o}^o\delta_{(\theta_i^o,\theta_{K^o}^o)}\\
+&\sum_{j=1}^{K^o-1}\pi_{K^o}^o\alpha_{K^o,j}^o\delta_{(\theta_{K^o}^o,\theta_j^o)}+\pi_{K^o}^o\alpha_{K^o,K^o}^o\delta_{(\theta_{K^o}^o,\theta_{K^o}^o)}\\
=&\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\pi_i^o\widehat{\alpha}_{ij}\delta_{(\theta_i^o,\theta_j^o)}+\sum_{i=1}^{K^o-1}\sum_{j=K^o}^{K^o+1}\pi_i^o\widehat{\alpha}_{i,j}\delta_{(\theta_i^o,\theta_{K^o}^o)}\\
+&\sum_{i=K^o}^{K^o+1}\sum_{j=1}^{K^o-1}\widehat{\pi}_i\widehat{\alpha}_{ij}\delta_{(\theta_{K^o}^o,\theta_j^o)}+\sum_{i=K^o}^{K^o+1}\sum_{j=K^o}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{i,j}\delta_{(\theta_{K^o}^o,\theta_{K^o}^o)}. \qquad (3.62)
\end{aligned}$$

Since θ_i^o are distinct and $\pi_i^o>0$, for $i=1,\ldots,K^o$ and $\widehat{\pi}_{K^o}=a\pi_{K^o}^o$, $\widehat{\pi}_{K^o+1}^o=b\pi_{K^o}^o$, then by Lemma 3.2.9, from (3.62), it can be derived that

$$\widehat{\alpha}_{ij}=\alpha_{ij}^o,\qquad i,j=1,\ldots,K^o-1 \qquad (3.63)$$

$$\widehat{\alpha}_{i,K^o}+\widehat{\alpha}_{i,K^o+1}=\alpha_{i,K^o}^o,\qquad i=1,\ldots,K^o-1 \qquad (3.64)$$

$$a\widehat{\alpha}_{K^o,j}+b\widehat{\alpha}_{K^o+1,j}=\alpha_{K^o,j}^o,\qquad j=1,\ldots,K^o-1 \qquad (3.65)$$

$$a(\widehat{\alpha}_{K^o,K^o}+\widehat{\alpha}_{K^o,K^o+1})+b(\widehat{\alpha}_{K^o+1,K^o}+\widehat{\alpha}_{K^o+1,K^o+1})=\alpha_{K^o,K^o}^o. \qquad (3.66)$$

So from (3.63), (3.64), (3.65) and (3.66), $\widehat{A}$ can be identified having form

$$\widehat{A}=\begin{pmatrix}
\alpha_{11}^o & \cdots & \alpha_{1,K^o-1}^o & \widehat{\alpha}_{1,K^o} & \widehat{\alpha}_{1,k^o+1}\\
\alpha_{21}^o & \cdots & \alpha_{2,K^o-1}^o & \widehat{\alpha}_{2,K^o} & \widehat{\alpha}_{2,K^o+1}\\
\vdots & \ddots & \vdots & \vdots & \vdots\\
\alpha_{K^o-1,1}^o & \cdots & \alpha_{K^o-1,K^o-1}^o & \widehat{\alpha}_{K^o-1,K^o} & \widehat{\alpha}_{K^o-1,K^o+1}\\
\widehat{\alpha}_{K^o,1} & \cdots & \widehat{\alpha}_{K^o,K^o-1} & \widehat{\alpha}_{K^o,K^o} & \widehat{\alpha}_{K^o,K^o+1}\\
\widehat{\alpha}_{K^o+1,1} & \cdots & \widehat{\alpha}_{K^o+1,K^o-1} & \widehat{\alpha}_{K^o+1,K^o} & \widehat{\alpha}_{K^o+1,K^o+1}
\end{pmatrix}.$$

To identify $\widehat{A}$ further, let us consider another equation implied by Corollary 3.3.6, that is

$$\sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\sum_{k=1}^{K^o}\pi_i^o\alpha_{ij}^o\alpha_{jk}^o\delta_{(\theta_i^o,\theta_j^o,\theta_k^o)}=\sum_{i=1}^{K^o+1}\sum_{j=1}^{K^o+1}\sum_{k=1}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\delta_{(\widehat{\theta}_i,\widehat{\theta}_j,\widehat{\theta}_k)}. \qquad (3.67)$$

By (3.60), (3.63), (3.64) and (3.65), (3.67) can be rewritten as

$$
\begin{aligned}
&\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\sum_{k=1}^{K^o-1}\pi_i^o\alpha_{ij}^o\alpha_{jk}^o\delta_{(\theta_i^o,\theta_j^o,\theta_k^o)}+\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\pi_i^o\alpha_{ij}^o\alpha_{j,K^o}^o\delta_{(\theta_i^o,\theta_j^o,\theta_{K^o}^o)}\\
&+\sum_{i=1}^{K^o-1}\sum_{k=1}^{K^o-1}\pi_i^o\alpha_{i,K^o}^o\alpha_{K^o,k}^o\delta_{(\theta_i^o,\theta_{K^o}^o,\theta_k^o)}+\sum_{i=1}^{K^o-1}\pi_i^o\alpha_{i,K^o}^o\alpha_{K^o,K^o}^o\delta_{(\theta_i^o,\theta_{K^o}^o,\theta_{K^o}^o)}\\
&+\sum_{j=1}^{K^o-1}\sum_{k=1}^{K^o-1}\pi_{K^o}^o\alpha_{K^o,j}^o\alpha_{jk}^o\delta_{(\theta_{K^o}^o,\theta_j^o,\theta_k^o)}+\sum_{j=1}^{K^o-1}\pi_{K^o}^o\alpha_{K^o,j}^o\alpha_{j,K^o}^o\delta_{(\theta_{K^o}^o,\theta_j^o,\theta_{K^o}^o)}\\
&+\sum_{k=1}^{K^o-1}\pi_{K^o}^o\alpha_{K^o,K^o}^o\alpha_{K^o,k}^o\delta_{(\theta_{K^o}^o,\theta_{K^o}^o,\theta_k^o)}+\pi_{K^o}^o\alpha_{K^o,K^o}^o\alpha_{K^o,K^o}^o\delta_{(\theta_{K^o}^o,\theta_{K^o}^o,\theta_{K^o}^o)}\\
=&\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\sum_{k=1}^{K^o-1}\pi_i^o\alpha_{ij}^o\alpha_{jk}^o\delta_{(\theta_i^o,\theta_j^o,\theta_k^o)}+\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\sum_{k=K^o}^{K^o+1}\pi_i^o\alpha_{ij}^o\widehat{\alpha}_{j,K^o}\delta_{(\theta_i^o,\theta_j^o,\theta_{K^o}^o)}\\
&+\sum_{i=1}^{K^o-1}\sum_{j=K^o}^{K^o+1}\sum_{k=1}^{K^o-1}\pi_i^o\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\delta_{(\theta_i^o,\theta_{K^o}^o,\theta_k^o)}+\sum_{i=1}^{K^o-1}\sum_{j=K^o}^{K^o+1}\sum_{k=K^o}^{K^o+1}\pi_i^o\widehat{\alpha}_{ij}\widehat{\alpha}_{j,k}\delta_{(\theta_i^o,\theta_{K^o}^o,\theta_{K^o}^o)}\\
&+\sum_{i=K^o}^{K^o+1}\sum_{j=1}^{K^o-1}\sum_{k=1}^{K^o-1}\widehat{\pi}_i\widehat{\alpha}_{ij}\alpha_{jk}^o\delta_{(\theta_{K^o}^o,\theta_j^o,\theta_k^o)}+\sum_{i=K^o}^{K^o+1}\sum_{j=1}^{K^o-1}\sum_{k=K^o}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{j,k}\delta_{(\theta_{K^o}^o,\theta_j^o,\theta_{K^o}^o)}\\
&+\sum_{i=K^o}^{K^o+1}\sum_{j=K^o}^{K^o+1}\sum_{k=1}^{K^o-1}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\delta_{(\theta_{K^o}^o,\theta_{K^o}^o,\theta_k^o)}+\sum_{i=K^o}^{K^o+1}\sum_{j=K^o}^{K^o+1}\sum_{k=K^o}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{j,k}\delta_{(\theta_{K^o}^o,\theta_{K^o}^o,\theta_{K^o}^o)}\\
=&\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\sum_{k=1}^{K^o-1}\pi_i^o\alpha_{ij}^o\alpha_{jk}^o\delta_{(\theta_i^o,\theta_j^o,\theta_k^o)}+\sum_{i=1}^{K^o-1}\sum_{j=1}^{K^o-1}\pi_i^o\alpha_{ij}^o\alpha_{j,K^o}^o\delta_{(\theta_i^o,\theta_j^o,\theta_{K^o}^o)}\\
&+\sum_{i=1}^{K^o-1}\sum_{j=K^o}^{K^o+1}\sum_{k=1}^{K^o-1}\pi_i^o\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\delta_{(\theta_i^o,\theta_{K^o}^o,\theta_k^o)}+\sum_{i=1}^{K^o-1}\sum_{j=K^o}^{K^o+1}\sum_{k=K^o}^{K^o+1}\pi_i^o\widehat{\alpha}_{ij}\widehat{\alpha}_{j,k}\delta_{(\theta_i^o,\theta_{K^o}^o,\theta_{K^o}^o)}\\
&+\sum_{j=1}^{K^o-1}\sum_{k=1}^{K^o-1}\pi_{K^o}^o\alpha_{K^o,j}^o\alpha_{jk}^o\delta_{(\theta_{K^o}^o,\theta_j^o,\theta_k^o)}+\sum_{j=1}^{K^o-1}\pi_{K^o}^o\alpha_{K^o,j}^o\alpha_{j,K^o}^o\delta_{(\theta_{K^o}^o,\theta_j^o,\theta_{K^o}^o)}\\
&+\sum_{i=K^o}^{K^o+1}\sum_{j=K^o}^{K^o+1}\sum_{k=1}^{K^o-1}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\delta_{(\theta_{K^o}^o,\theta_{K^o}^o,\theta_k^o)}+\sum_{i=K^o}^{K^o+1}\sum_{j=K^o}^{K^o+1}\sum_{k=K^o}^{K^o+1}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{j,k}\delta_{(\theta_{K^o}^o,\theta_{K^o}^o,\theta_{K^o}^o)}
\end{aligned}
\tag{3.68}
$$

By Lemma 3.2.9, (3.68) implies

$$
\widehat{\alpha}_{i,K^o}\widehat{\alpha}_{K^o,k}+\widehat{\alpha}_{i,K^o+1}\widehat{\alpha}_{K^o+1,k}=\alpha_{i,K^o}^o\alpha_{K^o,k}^o \tag{3.69}
$$

for $i,k=1,\ldots,K^o-1$,

$$
\widehat{\alpha}_{i,K^o}\left(\widehat{\alpha}_{K^o,K^o}+\widehat{\alpha}_{K^o,K^o+1}\right)+\widehat{\alpha}_{i,K^o+1}\left(\widehat{\alpha}_{K^o+1,K^o}+\widehat{\alpha}_{K^o+1,K^o+1}\right)=\alpha_{i,K^o}^o\alpha_{K^o,K^o}^o \tag{3.70}
$$

for $i = 1, \ldots, K^o - 1$,

$$\left(a\widehat{\alpha}_{K^o,K^o} + b\widehat{\alpha}_{K^o+1,K^o}\right)\widehat{\alpha}_{K^o,k} + \left(a\widehat{\alpha}_{K^o,K^o+1} + \widehat{\alpha}_{K^o+1,K^o+1}\right)\widehat{\alpha}_{K^o+1,k} = \alpha^o_{K^o,K^o}\alpha^o_{K^o,k} \tag{3.71}$$

for $k = 1, \ldots, K^o - 1$ and

$$\begin{aligned} &\left(a\widehat{\alpha}_{K^o,K^o} + b\widehat{\alpha}_{K^o+1,K^o}\right)\left(\widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1}\right) \\ &+\left(a\widehat{\alpha}_{K^o,K^o+1} + b\widehat{\alpha}_{K^o+1,K^o+1}\right)\left(\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1}\right) \;=\; \alpha^o_{K^o,K^o}\alpha^o_{K^o,K^o}. \end{aligned} \tag{3.72}$$

By (3.69), (3.64) and (3.65), for $i, k = 1, \ldots, K^o - 1$,

$$\begin{aligned} \widehat{\alpha}_{i,K^o}\widehat{\alpha}_{K^o,k} + \widehat{\alpha}_{i,K^o+1}\widehat{\alpha}_{K^o+1,k} &= \left(\widehat{\alpha}_{i,K^o} + \widehat{\alpha}_{i,K^o+1}\right)\left(a\widehat{\alpha}_{K^o,k} + b\widehat{\alpha}_{K^o+1,k}\right) \\ &= a\widehat{\alpha}_{i,K^o}\widehat{\alpha}_{K^o,k} + b\widehat{\alpha}_{i,K^o}\widehat{\alpha}_{K^o+1,k} \\ &\quad +a\widehat{\alpha}_{i,K^o+1}\widehat{\alpha}_{K^o,k} + b\widehat{\alpha}_{i,K^o+1}\widehat{\alpha}_{K^o+1,k} \end{aligned}$$

which gives

$$b\widehat{\alpha}_{i,K^o}\widehat{\alpha}_{K^o,k} + a\widehat{\alpha}_{i,K^o+1}\widehat{\alpha}_{K^o+1,k} = b\widehat{\alpha}_{i,K^o}\widehat{\alpha}_{K^o+1,k} + a\widehat{\alpha}_{i,K^o+1}\widehat{\alpha}_{K^o,k}$$

or

$$\left(a\widehat{\alpha}_{i,K^o+1} - b\widehat{\alpha}_{i,K^o}\right)\left(\widehat{\alpha}_{K^o+1,k} - \widehat{\alpha}_{K^o,k}\right) = 0$$

implying

$$a\widehat{\alpha}_{i,K^o+1} = b\widehat{\alpha}_{i,K^o} \qquad \text{or} \qquad \widehat{\alpha}_{K^o+1,k} = \widehat{\alpha}_{K^o,k} \tag{3.73}$$

From (3.70), (3.64) and (3.66), for $i = 1, \ldots, K^o - 1$,

$$\begin{aligned} &\widehat{\alpha}_{i,K^o}\left(\widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1}\right) + \widehat{\alpha}_{i,K^o+1}\left(\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1}\right) \\ &= \left(\widehat{\alpha}_{i,K^o} + \widehat{\alpha}_{i,K^o+1}\right)\left\{a\left(\widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1}\right) + b\left(\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1}\right)\right\} \\ &= a\widehat{\alpha}_{i,K^o}\left(\widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1}\right) + b\widehat{\alpha}_{i,K^o}\left(\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1}\right) \\ &\quad +a\widehat{\alpha}_{i,K^o+1}\left(\widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1}\right) + b\widehat{\alpha}_{i,K^o+1}\left(\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1}\right) \end{aligned}$$

which gives

$$\begin{aligned} & b\hat{\alpha}_{i,K^o}\left(\hat{\alpha}_{K^o,K^o}+\hat{\alpha}_{K^o,K^o+1}\right)+a\hat{\alpha}_{i,K^o+1}\left(\hat{\alpha}_{K^o+1,K^o}+\hat{\alpha}_{K^o+1,K^o+1}\right) \\ & \quad = b\hat{\alpha}_{i,K^o}\left(\hat{\alpha}_{K^o+1,K^o}+\hat{\alpha}_{K^o+1,K^o+1}\right)+a\hat{\alpha}_{i,K^o+1}\left(\hat{\alpha}_{K^o,K^o}+\hat{\alpha}_{K^o,K^o+1}\right) \end{aligned}$$

or

$$\left(a\hat{\alpha}_{i,K^o+1}-b\hat{\alpha}_{i,K^o}\right)\left\{\left(\hat{\alpha}_{K^o+1,K^o}+\hat{\alpha}_{K^o+1,K^o+1}\right)-\left(\hat{\alpha}_{K^o,K^o}+\hat{\alpha}_{K^o,K^o+1}\right)\right\}=0$$

implying

$$a\hat{\alpha}_{i,K^o+1}=b\hat{\alpha}_{i,K^o} \quad \text{or} \quad \hat{\alpha}_{K^o+1,K^o}+\hat{\alpha}_{K^o+1,K^o+1}=\hat{\alpha}_{K^o,K^o}+\hat{\alpha}_{K^o,K^o+1}. \quad (3.74)$$

From (3.71), (3.66) and (3.65), for $k=1,\ldots,K^o-1$,

$$\begin{aligned} & \left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)\hat{\alpha}_{K^o,k}+\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right)\hat{\alpha}_{K^o+1,k} \\ & \quad = \left\{\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)+\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right)\right\} \\ & \qquad \times\left(a\hat{\alpha}_{K^o,k}+b\hat{\alpha}_{K^o+1,k}\right) \\ & \quad = a\hat{\alpha}_{K^o,k}\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)+b\hat{\alpha}_{K^o+1,k}\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)+ \\ & \qquad a\hat{\alpha}_{K^o,k}\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right)+b\hat{\alpha}_{K^o+1,k}\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right) \end{aligned}$$

which gives

$$\begin{aligned} & b\hat{\alpha}_{K^o,k}\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)+a\hat{\alpha}_{K^o+1,k}\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right) \\ & \quad = b\hat{\alpha}_{K^o+1,k}\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)+a\hat{\alpha}_{K^o,k}\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right) \end{aligned}$$

or

$$\left(\hat{\alpha}_{K^o+1,k}-\hat{\alpha}_{K^o,k}\right)\left\{a\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right)-b\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right)\right\}=0$$

implying

$$\begin{aligned} \hat{\alpha}_{K^o+1,k} &= \hat{\alpha}_{K^o,k} \\ & \text{or} \\ a\left(a\hat{\alpha}_{K^o,K^o+1}+b\hat{\alpha}_{K^o+1,K^o+1}\right) &= b\left(a\hat{\alpha}_{K^o,K^o}+b\hat{\alpha}_{K^o+1,K^o}\right). \end{aligned} \quad (3.75)$$

From (3.72) and (3.66),

$$\begin{aligned}
& \left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right)\left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right) \\
+ \; & \left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right)\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right) \\
= \; & \left\{\left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right) + \left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right)\right\} \\
\times \; & \left\{a\left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right) + b\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right)\right\} \\
= \; & a\left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right)\left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right) \\
+ \; & b\left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right)\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right) \\
+ \; & a\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right)\left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right) \\
+ \; & b\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right)\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right)
\end{aligned}$$

which gives

$$\begin{aligned}
& b\left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right)\left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right) \\
+ \; & a\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right)\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right) \\
= \; & b\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o}\right)\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right) \\
+ \; & a\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right)\left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right)
\end{aligned}$$

or

$$\begin{aligned}
& \left\{a\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right) - b\left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right)\right\} \\
& \times\left\{\left(\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1}\right) - \left(\hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1}\right)\right\} \; = \; 0,
\end{aligned}$$

implying

$$\begin{aligned}
\hat{\alpha}_{K^o+1,K^o} + \hat{\alpha}_{K^o+1,K^o+1} &= \hat{\alpha}_{K^o,K^o} + \hat{\alpha}_{K^o,K^o+1} \\
&\text{or} \qquad (3.76) \\
a\left(a\hat{\alpha}_{K^o,K^o+1} + b\hat{\alpha}_{K^o+1,K^o+1}\right) &= b\left(a\hat{\alpha}_{K^o,K^o} + b\hat{\alpha}_{K^o+1,K^o}\right).
\end{aligned}$$

From (3.73) and (3.74), we can consider two subcases.

Subcase (1): Suppose there is i, $1 \le i \le K^o - 1$, such that

$$a\hat{\alpha}_{i,K^o+1} \neq b\hat{\alpha}_{i,K^o}.$$

Then by (3.73) and (3.74),

$$\widehat{\alpha}_{K^o+1,k} = \widehat{\alpha}_{K^o,k} \qquad \text{for } k = 1, \ldots, K^o - 1 \tag{3.77}$$

$$\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1} = \widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1}\,. \tag{3.78}$$

By (3.77), (3.78), (3.65) and (3.66),

$$\widehat{\alpha}_{K^o,k} = \widehat{\alpha}_{K^o+1,k} = \alpha^o_{K^o,k} \qquad k = 1, \ldots, K^o - 1 \tag{3.79}$$

$$\widehat{\alpha}_{K^o,K^o} + \widehat{\alpha}_{K^o,K^o+1} = \widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1} = \alpha^o_{K^o,K^o}\,. \tag{3.80}$$

Notice that (3.77) and (3.78) also imply (3.75) and (3.76). Hence by (3.79) and (3.80), in this subcase, the matrix $\widehat{A}$ is of the form

$$\widehat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & \widehat{\alpha}_{1,K^o} & \widehat{\alpha}_{1,K^o+1} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & \widehat{\alpha}_{2,K^o} & \widehat{\alpha}_{2,K^o+1} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & \widehat{\alpha}_{K^o-1,K^o} & \widehat{\alpha}_{K^o-1,K^o+1} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \widehat{\alpha}_{K^o,K^o} & \widehat{\alpha}_{K^o,K^o+1} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \widehat{\alpha}_{K^o+1,K^o} & \widehat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

where :

$$\widehat{\alpha}_{i,K^o} + \widehat{\alpha}_{i,K^o+1} = \alpha^o_{i,K^o}, \qquad i = 1, \ldots, K^o$$

$$\widehat{\alpha}_{K^o+1,K^o} + \widehat{\alpha}_{K^o+1,K^o+1} = \alpha^o_{K^o,K^o}$$

$$a, b \geq 0, \qquad a + b = 1$$

$$a\widehat{\alpha}_{i,K^o+1} \neq b\widehat{\alpha}_{i,K^o}, \qquad \text{for some } i, \quad 1 \leq i \leq K^o - 1.$$

So (b) follows.

Subcase (2): Suppose that

$$a\widehat{\alpha}_{i,K^o+1} = b\widehat{\alpha}_{i,K^o}\,, \qquad \text{for } i = 1, \ldots, K^o - 1, \tag{3.81}$$

then by (3.64), (3.81) implies

$$\widehat{\alpha}_{i,K^o} = a\alpha^o_{i,K^o} \qquad \widehat{\alpha}_{i,K^o+1} = b\alpha^o_{i,K^o}, \qquad \text{for } i = 1, \ldots, K^o - 1. \tag{3.82}$$

Consider two sub-subcases.

Sub-subcase (2.i): If

$$a\Big(a\widehat{\alpha}_{K^o,K^o+1}+b\widehat{\alpha}_{K^o+1,K^o+1}\Big)\neq b\Big(a\widehat{\alpha}_{K^o,K^o}+b\widehat{\alpha}_{K^o+1,K^o}\Big)$$

then by (3.75) and (3.76)

$$\widehat{\alpha}_{K^o+1,k}=\widehat{\alpha}_{K^o,k},\qquad \text{for } k=1,\ldots,K^o-1$$
$$\widehat{\alpha}_{K^o+1,K^o}+\widehat{\alpha}_{K^o+1,K^o+1}=\widehat{\alpha}_{K^o,K^o}+\widehat{\alpha}_{K^o,K^o+1}\,.$$

Thus, as in the subcase (1), (3.79) and (3.80) holds. Hence by (3.79), (3.80) and (3.82), in this sub-subcase, the matrix $\widehat{A}$ takes form

$$\widehat{A}=\begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & a\alpha^o_{1,K^o} & b\alpha^o_{1,k^o} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & a\alpha^o_{2,K^o} & b\alpha^o_{2,K^o} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & a\alpha^o_{K^o-1,K^o} & b\alpha^o_{K^o-1,K^o} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \widehat{\alpha}_{K^o,K^o} & \widehat{\alpha}_{K^o,K^o+1} \\ \alpha^o_{K^o,1} & \cdots & \alpha^o_{K^o,K^o-1} & \widehat{\alpha}_{K^o+1,K^o} & \widehat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

where :

$$\widehat{\alpha}_{i,K^o}+\widehat{\alpha}_{i,K^o+1}=\alpha^o_{K^o,K^o},\qquad i=K^o,K^o+1$$
$$a(a\widehat{\alpha}_{K^o,K^o+1}+b\widehat{\alpha}_{K^o+1,K^o+1})\neq b(a\widehat{\alpha}_{K^o,K^o}+b\widehat{\alpha}_{K^o+1,K^o}).$$

So (c) follows.

Sub-subcase(2ii): If

$$a\Big(a\widehat{\alpha}_{K^o,K^o+1}+b\widehat{\alpha}_{K^o+1,K^o+1}\Big)=b\Big(a\widehat{\alpha}_{K^o,K^o}+b\widehat{\alpha}_{K^o+1,K^o}\Big),\tag{3.83}$$

then (3.75) and (3.76) hold. Also by (3.66), (3.83) implies

$$a\widehat{\alpha}_{K^o,K^o}+b\widehat{\alpha}_{K^o+1,K^o}=a\widehat{\alpha}_{K^o,K^o}\tag{3.84}$$
$$a\widehat{\alpha}_{K^o,K^o+1}+b\widehat{\alpha}_{K^o+1,K^o+1}=b\widehat{\alpha}_{K^o,K^o}\,.\tag{3.85}$$

Thus by (3.82), (3.84) and (3.85), $\widehat{A}$ is of the form

$$\widehat{A} = \begin{pmatrix} \alpha^o_{11} & \cdots & \alpha^o_{1,K^o-1} & a\alpha^o_{1,K^o} & b\alpha^o_{1,k^o} \\ \alpha^o_{21} & \cdots & \alpha^o_{2,K^o-1} & a\alpha^o_{2,K^o} & b\alpha^o_{2,K^o} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha^o_{K^o-1,1} & \cdots & \alpha^o_{K^o-1,K^o-1} & a\alpha^o_{K^o-1,K^o} & b\alpha^o_{K^o-1,K^o} \\ \widehat{\alpha}_{K^o,1} & \cdots & \widehat{\alpha}_{K^o,K^o-1} & \widehat{\alpha}_{K^o,K^o} & \widehat{\alpha}_{K^o,K^o+1} \\ \widehat{\alpha}_{K^o+1,1} & \cdots & \widehat{\alpha}_{K^o+1,K^o-1} & \widehat{\alpha}_{K^o+1,K^o} & \widehat{\alpha}_{K^o+1,K^o+1} \end{pmatrix}$$

where :

$$a\widehat{\alpha}_{K^o,j} + b\widehat{\alpha}_{K^o+1,j} = \alpha^o_{K^o,j}, \qquad j = 1, \ldots, K^o - 1$$

$$a\widehat{\alpha}_{K^o,K^o} + b\widehat{\alpha}_{K^o+1,K^o} = a\alpha^o_{K^o,K^o}$$

$$a\widehat{\alpha}_{K^o,K^o+1} + b\widehat{\alpha}_{K^o+1,K^o+1} = b\alpha^o_{K^o,K^o}\,.$$

So (d) follows.

Lemma 3.3.11 *Assume $\widehat{\mathcal{H}}$ is identifiable. Let $\phi^o = (K^o, A^o, \pi^o, \theta^o) \in \Phi_{K^o}$ be a true parameter of a hidden Markov model $\{(X_t, Y_t)\}$, such that $\pi^o_i > 0$, for $i = 1, \ldots, K^o$. Let $\widehat{\phi} = (K^o, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K^o}$ and $\widehat{\phi} \sim \phi^o$. Then*

$$\{\theta^o_i : i = 1, \ldots, K^o\} = \{\widehat{\theta}_i : i = 1, \ldots, K^o\}.$$

Proof :

Let $\widehat{\phi} = (K^o, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_{K^o}$ and $\widehat{\phi} \sim \phi^o$. Consider the following equation implied by Corollary 3.3.6,

$$\sum_{i=1}^{K^o} \pi^o_i \delta_{\theta^o_i} = \sum_{i=1}^{K^o} \widehat{\pi}_i \delta_{\widehat{\theta}_i}\,. \tag{3.86}$$

Since $\pi^o_i > 0$, for $i = 1, \ldots, K^o$, then by Lemma 3.2.9, part (b), for every $i = 1, \ldots, K^o$, there is j, $1 \le j \le K^o$, such that $\theta^o_i = \widehat{\theta}_j$. So

$$\{\theta^o_i : i = 1, \ldots, K^o\} \subset \{\widehat{\theta}_i : i = 1, \ldots, K^o\}. \tag{3.87}$$

Suppose there is j, $1 \le j \le K^o$, such that

$$\widehat{\theta}_j \notin \{\theta^o_i : i = 1, \ldots, K^o\}.$$

Without loss of generality, suppose that

$$\widehat{\theta}_i \in \{\theta_i^o : i = 1, \ldots, K^o\}, \qquad \text{for } i = 1, \ldots, K_1 \tag{3.88}$$

$$\widehat{\theta}_i \notin \{\theta_i^o : i = 1, \ldots, K^o\}, \qquad \text{for } i = K_1 + 1, \ldots, K^o, \tag{3.89}$$

for some $K_1 \in \boldsymbol{N}$, with $1 \leq K_1 < K^o$. Lemma 3.2.9, (3.86), (3.88) and (3.89) imply that

$$\widehat{\pi}_i = 0, \qquad \text{for } i = K_1 + 1, \ldots, K^o. \tag{3.90}$$

Also by Corollary 3.3.6 and (3.90), we have

$$\begin{aligned} \sum_{i=1}^{K^o}\sum_{j=1}^{K^o} \pi_i^o \alpha_{ij}^o \delta_{(\theta_i^o, \theta_j^o)} &= \sum_{i=1}^{K^o}\sum_{j=1}^{K^o} \widehat{\pi}_i \widehat{\alpha}_{ij} \delta_{(\widehat{\theta}_i, \widehat{\theta}_j)} \\ &= \sum_{i=1}^{K_1}\sum_{j=1}^{K^o} \widehat{\pi}_i \widehat{\alpha}_{ij} \delta_{(\widehat{\theta}_i, \widehat{\theta}_j)}. \end{aligned} \tag{3.91}$$

Since $\widehat{\theta}_i \notin \{\theta_i^o : i = 1, \ldots, K^o\}$, for $i = K_1 + 1, \ldots, K^o$, then (3.91) implies

$$\widehat{\pi}_i \widehat{\alpha}_{ij} = 0, \qquad \text{for} \qquad i = 1, \ldots, K_1, \qquad j = K_1 + 1, \ldots, K^o. \tag{3.92}$$

Consider two cases.

Case (1): If $\widehat{\pi}_i > 0$, for $i = 1, \ldots, K_1$, then by (3.92)

$$\widehat{\alpha}_{ij} = 0, \qquad \text{for} \qquad i = 1, \ldots, K_1, \quad j = K_1 + 1, \ldots, K^o. \tag{3.93}$$

By (3.90) and (3.93), in this case, parameters of $\widehat{\phi}$ are of the forms :

$$\begin{aligned} \widehat{A} &= \begin{pmatrix} B_{K_1,K_1} & 0_{K_1,L_1} \\ D_{L_1.K_1} & E_{L_1,L_1} \end{pmatrix} \\ \widehat{\pi} &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_1}, 0, \ldots, 0) \\ \widehat{\theta} &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_1}, \widehat{\theta}_{K_1+1}, \ldots, \widehat{\theta}_{K^o})^T, \end{aligned}$$

where $L_1 = K^o - K_1$. Let

$$\begin{aligned} \widetilde{A}^1 &= B_{K_1,K_1} \\ \widetilde{\pi}^1 &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_1}) \\ \widetilde{\theta}^1 &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_1})^T \end{aligned}$$

and

$$\tilde{\phi}_1 = (K_1, \tilde{A}^1, \tilde{\pi}^1, \tilde{\theta}^1),$$

then $\tilde{\phi}_1 \in \Phi_{K_1}$ and $\tilde{\phi}_1 \sim \hat{\phi}$. Since $\hat{\phi} \sim \phi^o$, thus $\tilde{\phi}_1 \sim \phi^o$, contradicting with the fact that K^o is minimum.

Case (2): Suppose there exist i, $1 \leq i \leq K_1$, such that $\hat{\pi}_i = 0$. Without loss of generality, suppose that

$$\hat{\pi}_i > 0, \qquad \text{for } i = 1, \ldots, K_2 \tag{3.94}$$

$$\hat{\pi}_i = 0, \qquad \text{for } i = K_2 + 1, \ldots, K_1, \tag{3.95}$$

for some K_2, $1 \leq K_2 < K_1$. By (3.92), (3.94) and (3.95), we have

$$\hat{\alpha}_{ij} = 0, \qquad \text{for} \qquad i = 1, \ldots, K_2, \qquad j = K_1 + 1, \ldots, K^o. \tag{3.96}$$

By (3.90), (3.94), (3.95) and (3.96), parameters of $\hat{\phi}$ take forms

$$\hat{A} = \begin{pmatrix} B^1_{K_2,K_2} & B^2_{K_2,L_2} & 0_{K_2,L_1} \\ B^3_{L_2,K_2} & B^4_{L_2,L_2} & C_{L_2,L_1} \\ D_{L_1,K_1} & & E_{L_1,L_1} \end{pmatrix} \tag{3.97}$$

$$\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_{K_2}, 0, \ldots, 0) \tag{3.98}$$

$$\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_{K_2}, \hat{\theta}_{K_2+1}, \ldots, \hat{\theta}_{K^o})^T, \tag{3.99}$$

where $K_2 + L_2 = K_1$ and $K_2 + L_2 + L_1 = K^o$.

To identify $\hat{A}$ further, let us consider another equation implied by Corollary 3.3.6, that is

$$\begin{aligned} \sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\sum_{k=1}^{K^o} \pi^o_i \alpha^o_{ij}\alpha^o_{jk}\delta_{(\theta^o_i,\theta^o_j,\theta^o_k)} &= \sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\sum_{k=1}^{K^o} \hat{\pi}_i\hat{\alpha}_{ij}\hat{\alpha}_{jk}\delta_{(\hat{\theta}_i,\hat{\theta}_j,\hat{\theta}_k)} \\ &= \sum_{i=1}^{K_2}\sum_{j=1}^{K_1}\sum_{k=1}^{K^o} \hat{\pi}_i\hat{\alpha}_{ij}\hat{\alpha}_{jk}\delta_{(\hat{\theta}_i,\hat{\theta}_j,\hat{\theta}_k)}. \end{aligned} \tag{3.100}$$

Equation (3.100) follows from equations (3.97), (3.98) and (3.99). Since $\hat{\theta}_i \notin \{\theta^o_i : i = 1, \ldots, K^o\}$, for $i = K_1 + 1, \ldots, K^o$, then (3.99) and (3.96) imply

$$\hat{\pi}_i\hat{\alpha}_{ij}\hat{\alpha}_{jk} = 0, \qquad \text{for } i = 1, \ldots, K_2,\ j = K_2 + 1, \ldots, K_1,\ k = K_1 + 1, \ldots, K^o.$$

But $\widehat{\pi}_i > 0$, for $i = 1, \ldots, K_2$, then

$$\widehat{\alpha}_{ij}\widehat{\alpha}_{jk} = 0, \qquad \text{for } i = 1, \ldots, K_2,\ j = K_2 + 1, \ldots, K_1,\ k = K_1 + 1, \ldots, K^o. \tag{3.101}$$

So from (3.97) and (3.101), our focus will be matrices $B^2_{K_2,L_2}$ and C_{L_2,L_1}.

Based on the value of $L_2 = K_1 - K_2$, we divide case (2) into two subcases.
Subcase (2i): Suppose that $L_2 = K_1 - K_2 = 1$, then equation (3.101) can be written as

$$\widehat{\alpha}_{i,K_1}\widehat{\alpha}_{K_1,k} = 0, \qquad \text{for } i = 1, \ldots, K_1 - 1,\ k = K_1 + 1, \ldots, K^o. \tag{3.102}$$

If there exist i, $1 \leq i \leq K_1 - 1$, such that $\widehat{\alpha}_{i,K_1} > 0$, then by (3.102)

$$\widehat{\alpha}_{K_1,k} = 0, \qquad \text{for} \qquad k = K_1 + 1, \ldots, K^o,$$

giving

$$C_{1,L_1} = 0_{1,L_1}. \tag{3.103}$$

Thus by (3.97), (3.98), (3.99) and (3.103), parameters of $\widehat{\phi}$ take forms

$$\begin{aligned}
\widehat{A} &= \begin{pmatrix} B^1_{K_1-1,K_1-1} & B^2_{K_1-1,1} & 0_{K_1-1,L_1} \\ B^3_{1,K_1-1} & B^4_{1,1} & 0_{1,L_1} \\ D_{L_1,K_1} & & E_{L_1,L_1} \end{pmatrix} \\
\widehat{\pi} &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_1-1}, 0, \ldots, 0) \\
\widehat{\theta} &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_1-1}, \widehat{\theta}_{K_1}, \ldots, \widehat{\theta}_{K^o})^T.
\end{aligned}$$

Let

$$\begin{aligned}
\widetilde{A}^2 &= \begin{pmatrix} B^1_{K_1-1,K_1-1} & B^2_{K_1-1,1} \\ B^3_{1,K_1-1} & B^4_{1,1} \end{pmatrix} \\
\widetilde{\pi}^2 &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_1-1}, 0) \\
\widetilde{\theta}^2 &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_1-1}, \widehat{\theta}_{K_1})^T
\end{aligned}$$

and

$$\widetilde{\phi}_2 = (K_1, \widetilde{A}^2, \widetilde{\pi}^2, \widetilde{\theta}^2)$$

then $\tilde{\phi}_2 \in \Phi_{K_1}$ and $\tilde{\phi}_2 \sim \hat{\phi} \sim \phi^o$, contradicting with the fact that K^o is minimum.

If

$$\hat{\alpha}_{i,K_1} = 0, \qquad \text{for } i = 1, \ldots, K_1 - 1,$$

then

$$B^2_{K_1-1,1} = 0_{K_1-1,1}. \tag{3.104}$$

Hence by (3.97), (3.98), (3.99) and (3.104), parameters of $\hat{\phi}$, in this case have forms :

$$\begin{aligned}
\hat{A} &= \begin{pmatrix} B^1_{K_1-1,K_1-1} & 0_{K_1-1,L_1} & 0_{K_1-1,L_1} \\ B^3_{1,K_1-1} & B^4_{1,1} & C_{1,L_1} \\ & D_{L_1,K_1} & E_{L_1,L_1} \end{pmatrix} \\
\hat{\pi} &= (\hat{\pi}_1, \ldots, \hat{\pi}_{K_1-1}, 0, \ldots, 0) \\
\hat{\theta} &= (\hat{\theta}_1, \ldots, \hat{\theta}_{K_1-1}, \hat{\theta}_{K_1}, \ldots, \hat{\theta}_{K^o})^T.
\end{aligned}$$

Let

$$\begin{aligned}
\tilde{A}^3 &= B^1_{K_1-1,K_1-1} \\
\tilde{\pi}^3 &= (\hat{\pi}_1, \ldots, \hat{\pi}_{K_1-1}) \\
\tilde{\theta}^3 &= (\hat{\theta}_1, \ldots, \hat{\theta}_{K_1-1})^T
\end{aligned}$$

and

$$\tilde{\phi}_3 = (K_1 - 1, \tilde{A}^3, \tilde{\pi}^3, \tilde{\theta}^3)$$

then $\tilde{\phi}_3 \in \Phi_{K_1-1}$ and $\tilde{\phi}_3 \sim \hat{\phi} \sim \phi^o$, contradicting with the fact that K^o is minimum.

Subcase (2.ii): Suppose that $L_2 = K_2 - K_1 > 1$.
If for every $j = K_2 + 1, \ldots, K_1$ and $i = 1, \ldots, K_2$, $\hat{\alpha}_{ij} = 0$, then

$$B^2_{K_2,L_2} = 0_{K_2,L_2}. \tag{3.105}$$

Hence by (3.97), (3.98), (3.99) and (3.105), parameters of $\widehat{\phi}$ take forms :

$$\begin{aligned}
\widehat{A} &= \begin{pmatrix} B^1_{K_2,K_2} & 0^2_{K_2,L_2} & 0_{K_2,L_1} \\ B^3_{L_2,K_2} & B^4_{L_2,L_2} & C_{L_2,L_1} \\ D_{L_1,K_1} & & E_{L_1,L_1} \end{pmatrix} \\
\widehat{\pi} &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_2}, 0, \ldots, 0) \\
\widehat{\theta} &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_2}, \ldots, \widehat{\theta}_{K^o})^T.
\end{aligned}$$

Let

$$\begin{aligned}
\widetilde{A}^4 &= B^1_{K_2,K_2} \\
\widetilde{\pi}^4 &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_2}) \\
\widetilde{\theta}^4 &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_2})^T
\end{aligned}$$

and

$$\widetilde{\phi}_4 = (K_2, \widetilde{A}^4, \widetilde{\pi}^4, \widetilde{\theta}^4)$$

then $\widetilde{\phi}_4 \in \Phi_{K_2}$ and $\widetilde{\phi}_4 \sim \widehat{\phi} \sim \phi^o$, contradicting with the fact that K^o is minimum.

If for every $j = K_2 + 1, \ldots, K_1$, there is i, $1 \leq i \leq K_2$, such that $\widehat{\alpha}_{ij} > 0$, then by (3.101),

$$\widehat{\alpha}_{j,k} = 0, \qquad \text{for} \qquad j = K_2 + 1, \ldots, K_1, \qquad k = K_1 + 1, \ldots, K^o, \quad (3.106)$$

that is

$$C_{L_2,L_1} = 0_{L_2,L_1}. \quad (3.107)$$

Thus by (3.97), (3.98), (3.99) and (3.107), parameters of $\widehat{\phi}$ have forms

$$\begin{aligned}
\widehat{A} &= \begin{pmatrix} B^1_{K_2,K_2} & B^2_{K_2,L_2} & 0_{K_2,L_1} \\ B^3_{L_2,K_2} & B^4_{L_2,L_2} & 0_{L_2,L_1} \\ D_{L_1,K_1} & & E_{L_1,L_1} \end{pmatrix} \\
\widehat{\pi} &= (\widehat{\pi}_1, \ldots, \widehat{\pi}_{K_2}, 0, \ldots, 0) \\
\widehat{\theta} &= (\widehat{\theta}_1, \ldots, \widehat{\theta}_{K_2}, \ldots, \widehat{\theta}_{K^o})^T.
\end{aligned}$$

Let

$$\begin{aligned}\tilde{A}^5 &= \begin{pmatrix} B^1_{K_2,K_2} & B^2_{K_2,L_2} \\ B^3_{L_2,K_2} & B^4_{L_2,L_2} \end{pmatrix} \\ \tilde{\pi}^5 &= (\hat{\pi}_1, \ldots, \hat{\pi}_{K_2}, \ldots, 0) \\ \tilde{\theta}^5 &= (\hat{\theta}_1, \ldots, \hat{\theta}_{K_2}, \ldots, \hat{\theta}_{K_1})^T\end{aligned}$$

and

$$\tilde{\phi}_5 = (K_1, \tilde{A}^5, \tilde{\pi}^5, \tilde{\theta}^5)$$

then $\tilde{\phi}_5 \in \Phi_{K_1}$ and $\tilde{\phi}_5 \sim \hat{\phi} \sim \phi^o$, contradicting with the fact that K^o is minimum.

Suppose without loss of generality, there is $K_3 \in \boldsymbol{N}$ with $K_2 < K_3 < K_1$ satisfying :

(a). for every $j = K_2 + 1, \ldots, K_3$, there is i, $1 \leq i \leq K_2$, such that $\hat{\alpha}_{ij} > 0$,

(b). for every $j = K_3 + 1, \ldots, K_1$ and $i = 1, \ldots, K_2$, $\hat{\alpha}_{ij} = 0$.

By (3.101) and (a),

$$\hat{\alpha}_{jk} = 0, \qquad \text{for} \qquad j = K_2 + 1, \ldots, K_3, \qquad k = K_1 + 1, \ldots, K^o. \tag{3.108}$$

So in this case, by (3.108), (3.97), (3.98) and (3.99), parameters of $\hat{\phi}$ are of the forms :

$$\hat{A} = \begin{pmatrix} B^1_{K_2,K_2} & B^{21}_{K_2,L_3} & 0_{K_2,L_4} & 0_{K_2,L_1} \\ B^{31}_{L_3,K_2} & B^{41}_{L_3,L_3} & B^{42}_{L_3,L_4} & 0_{L_3,L_1} \\ B^{32}_{L_4,K_2} & B^{43}_{L_4,L_3} & B^{44}_{L_4,L_4} & C^{11}_{L_4,L_1} \\ & D_{L_1,K_1} & & E_{L_1,L_1} \end{pmatrix} \tag{3.109}$$

$$\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_{K_2}, 0, \ldots, 0) \tag{3.110}$$

$$\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_{K_2}, \ldots, \hat{\theta}_{K^o})^T, \tag{3.111}$$

where $K_2+L_3=K_3$, $K_2+L_3+L_4=K^1$ and $K_2+L_3+L_4+L_1=K^o$. Notice that by (a)

$$B^{21}_{K_2,L_3} \neq 0_{K_2,L_3}.$$

To identify $\widehat{A}$ further, consider the following equation implied by Corollary 3.3.6,

$$\begin{aligned}\sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\sum_{k=1}^{K^o}\sum_{l=1}^{K^o}\pi_i^o\alpha_{ij}^o\alpha_{jk}^o\alpha_{kl}^o\delta_{(\theta_i^o,\theta_j^o,\theta_k^o,\theta_l^o)} &= \sum_{i=1}^{K^o}\sum_{j=1}^{K^o}\sum_{k=1}^{K^o}\sum_{l=1}^{K^o}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\widehat{\alpha}_{kl}\delta_{(\widehat{\theta}_i,\widehat{\theta}_j,\widehat{\theta}_k,\widehat{\theta}_l)} \\ &= \sum_{i=1}^{K_2}\sum_{j=1}^{K_3}\sum_{k=1}^{K_1}\sum_{l=1}^{K^o}\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\widehat{\alpha}_{kl}\delta_{(\widehat{\theta}_i,\widehat{\theta}_j,\widehat{\theta}_k,\widehat{\theta}_l)}.\end{aligned} \tag{3.112}$$

Equation (3.112) follows from equations (3.109), (3.110) and (3.111). Since $\widehat{\theta}_i \notin \{\theta_i^o : i=1,\ldots,K^o\}$, for $i=K_1+1,\ldots,K^o$, then (3.109), (3.110), (3.111), (3.112) implies

$$\widehat{\pi}_i\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\widehat{\alpha}_{kl}=0,$$

for $i=1,\ldots,K_2$, $j=K_2+1,\ldots,K_3$, $k=K_3+1,\ldots,K_1$ and $l=K_1+1,\ldots,K^o$. As $\widehat{\pi}_i>0$, for $i=1,\ldots,K_2$, then we have

$$\widehat{\alpha}_{ij}\widehat{\alpha}_{jk}\widehat{\alpha}_{kl}=0, \tag{3.113}$$

for $i=1,\ldots,K_2$, $j=K_2+1,\ldots,K_3$, $k=K_3+1,\ldots,K_1$ and $l=K_1+1,\ldots,K^o$. By (3.113) and (3.109), our focus will be matrices $B^{21}_{K_2,L_3}$, $B^{42}_{L_3,L_4}$ and $C^{11}_{L_4,L_1}$. However from (a), for every $j=K_2+1,\ldots,K_3$, there is i, $1\leq i\leq K_2$, such that $\widehat{\alpha}_{ij}>0$. So $B^{21}_{K_2,L_3}\neq 0_{K_2,L_3}$. As a result, we only have to focus ourself to matrices $B^{42}_{L_3,L_4}$ and $C^{11}_{L_4,L_1}$.

If $K_1-K_3=1$, then we have a case similar to subcase (2.i). If $K_1-K_3>1$, then we have a case similar to subcase (2.ii). We must consider the equation in Corollary 3.3.6, for $n=5$. Since K^o is finite, then maximum number of procedure that we have to follow is $(K_1-K_3)/2$.

Since the assumption : "there is j, $1 \leq j \leq K^o$, such that $\widehat{\theta}_j \notin \{\theta_i^o : i = 1, \ldots, K^o\}$" lead to contradiction, then it must be, for every $j = 1, \ldots, K^o$, $\widehat{\theta}_j \in \{\theta_i^o : i = 1, \ldots, K^o\}$. Thus

$$\{\widehat{\theta}_i : i = 1, \ldots, K^o\} \subset \{\theta_i^o : i = 1, \ldots, K^o\}. \tag{3.114}$$

By (3.84) and (3.114), the conclusion of the lemma follows. ■

Chapter 4

Maximum Likelihood Estimation for Hidden Markov Models

Numerous concrete phenomena provide numerical sequences of observations, for instance, in econometrics: stock prices, interest rates, exchange rates, etc; or in meteorology: daily temperatures, weekly rain levels, etc. In general, these sequences have two characteristics :

(a). The graph of the observations is *irregular* and *complex* which is impossible to model by a *simple* curve depending on a *small* number of parameters.

(b). It is impossible to have the sequence again in *identical* conditions.

Suppose a series of observations $\{y_1, \ldots, y_n\}$ is given to be modelled for some specific reason. In view of (a) and (b), it is convenient to model the sequence $\{y_1, \ldots, y_n\}$ as observations of some *unknown stochastic process* $\{Y_t : t \in \boldsymbol{N}\}$. The observed data sample $\{y_1, \ldots, y_n\}$ is now interpreted as the initial segment of the realization $\{Y_t\}$.

Based on prior information, insight and mathematical tractability, we suppose $\{Y_t\}$ is *equivalent* with the observation process of a hidden Markov model, which is generated by an *unknown true parameter* $\phi^o = (K^o, A^o, \pi^o, \theta^o)$, in a sense that, $\{Y_t\}$ and the observation process have the same finite dimensional joint density functions. $\{Y_t\}$ is defined as the coordinate projection process on $(\mathcal{Y}^\infty, \mathcal{B}, P_{\phi^o})$, where $\mathcal{Y}^\infty$ is the set of all realizations $y = \{y_t\}$ and $\mathcal{B}$ is the Borel σ-field of $\mathcal{Y}^\infty$, that is, for $y = \{y_t\} \in \mathcal{Y}^\infty$,

$$Y_t(y) = y_t, \qquad t \in \boldsymbol{N}.$$

Under ϕ^o, for any $n \in \boldsymbol{N}$, the n-dimensional joint density function of $Y_1, \ldots, Y_n$ is

$$p_{\phi^o}(y_1, \ldots, y_n) = \sum_{x_1=1}^{K^o} \cdots \sum_{x_n=1}^{K^o} \pi^o_{x_1} f(y_1, \theta^o_{x_1}) \prod_{t=2}^{n} \alpha^o_{x_{t-1}, x_t} f(y_t, \theta^o_{x_t}).$$

The modelling problem is now reduced to an *estimation problem.* To estimate the true parameter ϕ^o, we have to select a class of hidden Markov models and define a likelihood function on it. The true parameter ϕ^o is then estimated by parameters in the class which maximize the likelihood function.

Consider two approaches for estimating ϕ^o.

The first approach :

This approach is based on the assumption that the selected class of hidden Markov models is the one that consists of all models having size K^o, same as the order of the true parameter ϕ^o. So in this approach,

$$\Phi_{K^o} = \Big\{ \phi \ : \ \phi = (K^o, A, \pi, \theta), \text{ where } A, \pi \text{ and } \theta \text{ satisfy :}$$

$$A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K^o} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K^o$$

$$\pi = (\pi_i), \quad \pi_i \geq 0, \quad i = 1, \ldots, K^o, \quad \sum_{i=1}^{K^o} \pi_i = 1$$

$$\theta = (\theta_i)^T, \quad \theta_i \in \Theta, \quad i = 1, \ldots, K^o \Big\}$$

will be this class

For $i, j = 1, \ldots, K^o$, define *coordinate projections* $\alpha_{ij}(\cdot)$, $\pi_i(\cdot)$ and $\theta_i(\cdot)$ on Φ_{K^o} by

$$\alpha_{ij}(\phi) = \alpha_{ij}, \qquad \pi_i(\phi) = \pi_i, \qquad \theta_i(\phi) = \theta_i,$$

for $\phi = (K^o, A, \pi, \theta) \in \Phi_{K^o}$.

Under $\phi \in \Phi_{K^o}$, $\{Y_t\}$ is defined as the coordinate projection process on $(\mathcal{Y}^\infty, \mathcal{B}, P_\phi)$ and for any $n \in \boldsymbol{N}$, $Y_1, \ldots, Y_n$ has the n-dimensional joint density function

$$p_\phi(y_1, \ldots, y_n) = \sum_{x_1=1}^{K^o} \cdots \sum_{x_n=1}^{K^o} \pi_{x_1}(\phi) f(y_1, \theta_{x_1}(\phi)) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}(\phi) f(y_t, \theta_{x_t}(\phi)).$$

Define the *log-likelihood function* on Φ_{K^o} by

$$L_n(\phi, y) = \frac{1}{n} \log p_\phi(y_1, \ldots, y_n), \qquad n \in \boldsymbol{N}, \tag{4.1}$$

for $y = \{y_t\} \in \mathcal{Y}^\infty$, and the *maximum likelihood estimator* $\{\hat{\phi}_n : n \in \boldsymbol{N}\}$ such that

$$\hat{\phi}_n(y) = \left\{ \tilde{\phi} : L_n(\tilde{\phi}, y) = \sup_{\phi \in \Phi_{K^o}} L_n(\phi, y) \right\}, \qquad n \in \boldsymbol{N}. \tag{4.2}$$

Since a true parameter for a hidden Markov model is not unique, let

$$\mathcal{T}^o = \left\{ \phi \in \Phi_{K^o} : \phi \sim \phi^o \right\}.$$

The maximum likelihood estimator $\{\hat{\phi}_n\}$ is then judged to be *good*, if it is *strongly consistent*, that is

$$\hat{\phi}_n(y) \longrightarrow \mathcal{T}^o, \qquad \text{for almost all y under } \phi^o,$$

when $n \to \infty$, or in another words:

$$\hat{\phi}_n \longrightarrow \mathcal{T}^o, \qquad \text{with probability one under } \phi^o, \tag{4.3}$$

when $n \to \infty$.

The second approach : (which we prefer)

Since the order K^o is *unknown*, let

$$\Phi_K = \Big\{ \phi \ : \ \phi = (K, A, \pi, \theta), \text{ where } A, \ \pi \text{ and } \theta \text{ satisfy :}$$

$$A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K$$

$$\pi = (\pi_i), \quad \pi_i \geq 0, \quad i = 1, \ldots, K, \quad \sum_{i=1}^{K} \pi_i = 1$$

$$\theta = (\theta_i)^T, \quad \theta_i \in \Theta, \quad i = 1, \ldots, K \Big\}$$

be the selected class of hidden Markov models, for a *fixed* $K \in \boldsymbol{N}$.

For $i, j = 1, \ldots, K$, define *coordinate projections* $\alpha_{ij}(\cdot)$, $\pi_i(\cdot)$ and $\theta_i(\cdot)$ on Φ_K by

$$\alpha_{ij}(\phi) = \alpha_{ij}, \qquad \pi_i(\phi) = \pi_i, \qquad \theta_i(\phi) = \theta_i,$$

for $\phi = (K, A, \pi, \theta) \in \Phi_K$.

Under $\phi \in \Phi_K$, $\{(Y_t)\}$ is defined as the coordinate projection process on $(\mathcal{Y}^\infty, \mathcal{B}, P_\phi)$ and for any $n \in \boldsymbol{N}$, $Y_1, \ldots, Y_n$ has the n-dimensional joint density function

$$p_\phi(y_1, \ldots, y_n) = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1}(\phi) f(y_1, \theta_{x_1}(\phi)) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}(\phi) f(y_t, \theta_{x_t}(\phi)).$$

Notice that a true parameter ϕ^o may not be in Φ_K. A true parameter estimation is then transformed into one of *best approximation.* A *distance* between $\phi \in \Phi_K$ and ϕ^o is introduced and a *quasi true parameter* is defined as parameter which minimizes distance to ϕ^o.

Define a distance between ϕ^o and ϕ in Φ_K as *Kullback-Leibler divergence*

$$K(\phi^o, \phi) = \lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\left[\log \frac{p_{\phi^o}(Y_1, \ldots, Y_n)}{p_\phi(Y_1, \ldots, Y_n)}\right], \tag{4.4}$$

where E_{ϕ^o} is the expectation with respect to ϕ^o. As a distance, we will prove later in section 4.3, that $K(\phi^o, \phi) \geq 0$, for every $\phi \in \Phi_K$.

Define the *quasi true parameter* set as

$$\mathcal{N} = \left\{\tilde{\phi} : K(\phi^o, \tilde{\phi}) = \inf_{\phi\in\Phi_K} K(\phi^o, \phi)\right\}. \tag{4.5}$$

Later in section 4.8, it will be shown that if $K \geq K^o$, then

$$\mathcal{N} = \{\phi \in \Phi_K : \phi \sim \phi^o\}. \tag{4.6}$$

Now define the *log-likelihood function* on Φ_K by

$$L_n(\phi, y) = \frac{1}{n} \log p_\phi(y_1, \ldots, y_n), \qquad n \in \boldsymbol{N}, \tag{4.7}$$

for $y = \{y_t\} \in \mathcal{Y}^\infty$. The *maximum likelihood estimator* $\{\hat{\phi}_n : n \in \boldsymbol{N}\}$ is defined by

$$\hat{\phi}_n(y) = \left\{\tilde{\phi} : L_n(\tilde{\phi}, y) = \sup_{\phi\in\Phi_K} L_n(\phi, y)\right\}, \qquad n \in \boldsymbol{N}.$$

The estimator $\{\hat{\phi}_n\}$ is then judged to be *good*, if it is *strongly consistent*, that is

$$\hat{\phi}_n \longrightarrow \mathcal{N}, \qquad \text{with probability one under } \phi^o, \tag{4.8}$$

when $n \to \infty$.

Since for $K = K^o$, $\mathcal{N} = \mathcal{T}^o$ by (4.6), then by comparing (4.3) and (4.8), it can be concluded that the first approach is a special case of the second.

The consistency of the maximum likelihood estimator in the sense of the first approach has been established by Leroux in [34]. In this chapter we will prove

the consistency of the maximum likelihood estimator using the second approach. So our result will be a generalization of [34].

This chapter begins by giving some topology to the hidden Markov model space Φ_K in section 4.1. This section also gives some regularity conditions for Φ_K and the true parameter ϕ^o. In section 4.2, some properties of the log-likelihood function are presented. In section 4.3, the Kullback-Leibler divergence is discussed, starting from general to hidden Markov case.

The focus of section 4.4 is to find the relation between the Kullback-Leibler divergence and the log-likelihood process. The result of this section shows that the Kullback-Leibler divergence does not depend on the initial probality distribution, which gives the idea for simplify the hidden Markov model space Φ_K in section 4.5.

Section 4.6 studies the relation between the Kullback-Leibler divergence and parameters which are equivalent to the true parameter. In section 4.7, uniform convergence of the likelihood process is derived. Based on the results of section 4.6 and section 4.7, the quasi true parameter set can then be analysed in section 4.8. Consistency of the maximum likelihood estimator, which is the main result of this chapter, is presented in section 4.9.

4.1 Parameter Restriction

Let

$$\Phi_K = \Big\{ \phi \ : \ \phi = (K, A, \pi, \theta), \text{ where } A, \pi \text{ and } \theta \text{ satisfy :}$$

$$A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K$$

$$\pi = (\pi_i), \quad \pi_i \geq 0, \quad i = 1, \ldots, K, \quad \sum_{i=1}^{K} \pi_i = 1$$

$$\theta = (\theta_i)^T, \quad \theta_i \in \Theta, \quad i = 1, \ldots, K \Big\} \tag{4.9}$$

be the hidden Markov model parameter space for a *fixed* $K \in \boldsymbol{N}$. Since all hidden Markov models in Φ_K have the same size K, then for each $\phi = (K, A, \pi, \theta) \in \Phi_K$, we may consider it as

$$\phi = (A, \pi, \theta).$$

For estimation purposes, we need to compactify Φ_K. Let

$$\Psi_K = \Big\{ (A, \pi) \ : \ \text{where } A \text{ and } \pi \text{ satisfy :}$$

$$A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K$$

$$\pi = (\pi_i), \quad \pi_i \geq 0, \quad i = 1, \ldots, K, \quad \sum_{i=1}^{K} \pi_i = 1 \Big\}. \tag{4.10}$$

Ψ_K is a *compact* subset of $\boldsymbol{R}^{K^2+K}$ with respect to the Euclidean norm $\|\cdot\|$. By (4.9) and (4.10),

$$\Phi_K = \Psi_K \times \Theta^K. \tag{4.11}$$

Suppose that $\Theta \subset \boldsymbol{R}^n$, for some $n \in \boldsymbol{N}$. From examples, Θ is usually *locally compact* and *not compact*. Hence by (4.11), to compactify Φ_K, we need to compactify Θ.

Let Θ^c be the *one-point compactification* of Θ (see for example [17], page 321). Θ^c is obtained by attaching a point *infinity*, denoted by ∞, to Θ. Extend $f(y, \cdot)$ to Θ^c by defining $f(y, \infty) = 0$.

Suppose $\Theta = \boldsymbol{R}^n$, then $\Theta^c = \boldsymbol{R}^n \cup \{\infty\}$. From [17], page 196, $\boldsymbol{R}^n \cup \{\infty\}$ is *homeomorphic* to $\mathcal{S}_n = \{x \in \boldsymbol{R}^{n+1} : \|x\| = 1\}$ by homeomorphism f :

$\boldsymbol{R}^n \cup \{\infty\} \cong \mathcal{S}_n$ which is defined by

$$\begin{aligned} f(x) &= \frac{2}{\|x\|^2+1}(x_1,\ldots,x_n,0) + \frac{\|x\|^2-1}{\|x\|^2+1}(0,\ldots,0,1) \\ f(\infty) &= (0,\ldots,0,1), \end{aligned}$$

for $x = (x_1,\ldots,x_n) \in \boldsymbol{R}^n$. Define norm $\|\cdot\|_\infty$ on $\boldsymbol{R}^n \cup \{\infty\}$ by

$$\|x\|_\infty = \|f(x)\|. \tag{4.12}$$

Let Φ_K^c be the *compactification space* of Φ_K. Thus

$$\Phi_K^c = \Psi_K \times \left(\Theta^c\right)^K.$$

Notice that Φ_K is *dense* on Φ_K^c. Define norm $\|\cdot\|_K$ in Φ_K^c by

$$\|\phi\|_K = \|(A,\pi)\| + \sum_{i=1}^{K} \|\theta_i\|_\infty, \tag{4.13}$$

for $\phi = (K, A, \pi, \theta) \in \Phi_K^c$.

For each $i, j = 1,\ldots,K$, define the *coordinate projections* $\alpha_{ij}(\cdot)$, $\pi_i(\cdot)$ and $\theta_i(\cdot)$ on Φ_K^c by

$$\alpha_{ij}(\phi) = \alpha_{ij}, \qquad \pi_i(\phi) = \pi_i, \qquad \theta_i(\phi) = \theta_i,$$

for $\phi = (K, A, \pi, \theta) \in \Phi_K^c$

Based on the results of Chapter 2 and Chapter 3, and to simplify theoretical considerations, our model parameters will be restricted to a certain class which satisfy the following conditions.

A1. The transition probability matrix A^o is *irreducible.*

A2. π^o is a *stationary* probability distribution of A^o.

A3. The family of finite mixtures on $\{f(\cdot,\theta) : \theta \in \Theta\}$ is *identifiable*

A4. $f(\cdot,\cdot) > 0$ and continuous on $\mathcal{Y}\times\Theta$. For each y, $f(y,\cdot)$ *vanishes* at infinity.

A5. For each $i, j = 1, 2, \ldots, K$, $\pi_i(\cdot)$, $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ are continuous functions on Φ_K^c

A6. $E_{\phi^o}\left[|\log f(Y_1, \theta_i^o)|\right] < \infty, \quad$ for $i = 1, \ldots, K^o$.

A7. For every $\theta \in \Theta$, $E_{\phi^o}\left[(\log f(Y_1, \theta))^+\right] < \infty$, where $x^+ = \max\{x, 0\}$.

Remarks 4.1.1

(a). By conditions A1 and A2, $\pi_i^o > 0$, for $i = 1, \ldots, K^o$.

(b). Conditions A1 and A2 guarantee that the observed process $\{Y_t\}$ is *stationary* and *ergodic* under ϕ^o and under all ϕ which are equivalent to ϕ^o.

(c). From Chapter 3, condition A3 is needed for parameter identification purposes.

(d). Conditions A4 and A5 together imply that the likelihood function $L_n(\cdot, y)$, for each $n \in \boldsymbol{N}$ and $y \in \mathcal{Y}^\infty$, is *continuous* on compact space Φ_K^c. This guarantees the *existence* of maximum likelihood estimator.

(e). Conditions A6 and A7 are uniform integrability conditions, which are essential for the Kullback-Leibler divergence.

Next, some examples of distribution families which satisfy conditions A3, A4, A6 and A7 are given. For these, the following lemma is needed.

Lemma 4.1.2 (Farrel and Ross [21]) *Let* $a, b, c \in \boldsymbol{R}$. *If* $b, c > 0$, *and* $a > -1$, *then*

$$\int_0^\infty t^a e^{-bt^c}\, dt = \frac{\Gamma(\frac{a+1}{c})}{c b^{\frac{a+1}{c}}}.$$

Proof :

Let $x = bt^c$, then $t = \frac{x^{\frac{1}{c}}}{b^{\frac{1}{c}}}$ and $dx = cbt^{c-1}\,dt$. Therefore

$$\begin{aligned}\int_0^\infty t^a e^{-bt^c}\,dt &= \int_0^\infty \frac{x^{\frac{a}{c}}}{b^{\frac{a}{c}}} e^{-x} \frac{x^{\frac{1}{c}-1}}{cb^{\frac{1}{c}}}\,dx \\ &= \frac{1}{cb^{\frac{a+1}{c}}} \int_0^\infty e^{-x} x^{\frac{a+1}{c}-1}\,dx \\ &= \frac{\Gamma(\frac{a+1}{c})}{cb^{\frac{a+1}{c}}}.\end{aligned}$$

■

Using Lemma 4.1.2, conditions A3, A4, A6 and A7 hold for three important families of distributions.

1. The family of Poisson distributions.

- $f(y,\theta) = \dfrac{e^{-\theta}\theta^y}{y!}, \qquad y \in \mathcal{Y} = \{0,1,2,\ldots\}, \qquad \theta \in \Theta = (0,\infty).$
- $\{f(\cdot,\theta) : \theta \in \Theta\}$ is identifiable (see section 3.2).
- $f(\cdot,\cdot)$ is continuous on $\mathcal{Y} \times \Theta$ and for each $y \in \mathcal{Y}$, $f(y,\cdot)$ vanishes at infinity.
- Let $\theta \in \Theta$.

$$\begin{aligned}E_{\phi^o}\left[|\log f(Y_1,\theta)|\right] &= \sum_{y_1=0}^{\infty} p_{\phi^o}(y_1)|\log f(y_1,\theta)| \\ &= \sum_{y_1=0}^{\infty}\sum_{j=1}^{K^o} \pi_j^o f(y_1,\theta_j^o)|\log f(y_1,\theta)| \\ &\le \sum_{y_1=0}^{\infty}\sum_{j=1}^{K^o} \pi_j^o \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!}\Big\{\theta + y_1|\log\theta| + \log(y_1!)\Big\} \\ &= \sum_{j=1}^{K^o} \pi_j^o \sum_{y_1=0}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!}\Big\{\theta + y_1|\log\theta| + \log(y_1!)\Big\}\end{aligned} \tag{4.14}$$

$$\sum_{j=1}^{K^o} \pi_j^o \sum_{y_1=0}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!} = 1$$

$$\begin{aligned}
\sum_{y_1=0}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!} y_1 |\log\theta| &= \theta_j^o |\log\theta| e^{-\theta_j^o} \sum_{y_1=1}^{\infty} \frac{(\theta_j^o)^{y_1-1}}{(y_1-1)!} \\
&= \theta_j^o |\log\theta| e^{-\theta_j^o} e^{\theta_j^o} \\
&= \theta_j^o |\log\theta| \qquad (4.15)
\end{aligned}$$

$$\begin{aligned}
\sum_{y_1=0}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!} \log(y_1!) &= \sum_{y_1=1}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!} \log(y_1!) \\
&\leq \sum_{y_1=2}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!} y_1 \log y_1 \\
&\leq \sum_{y_1=2}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{y_1!} y_1(y_1-1) \\
&= \sum_{y_1=2}^{\infty} \frac{e^{-\theta_j^o}(\theta_j^o)^{y_1}}{(y_1-2)!} \\
&= (\theta_j^o)^2 e^{-\theta_j^o} \sum_{y_1=2}^{\infty} \frac{(\theta_j^o)^{y_1-2}}{(y_1-2)!} \\
&= (\theta_j^o)^2 e^{-\theta_j^o} e^{\theta_j^o} \\
&= (\theta_j^o)^2 \qquad (4.16)
\end{aligned}$$

where the first and the second inequalities come respectively from $y_1! \leq (y_1)^{y_1}$ and $\log y_1 \leq y_1 - 1$ (see [1], page 68).

From (4.14), (4.15) and (4.16), then

$$\begin{aligned}
E_{\phi^o}\Big[|\log f(Y_1,\theta)|\Big] &\leq \theta + \sum_{j=1}^{K^o} \pi_j^o \Big\{\theta_j^o |\log\theta| + (\theta_j^o)^2\Big\} \\
&< \infty.
\end{aligned}$$

Hence condition A6 and A7 hold.

2. The family of negative exponential distributions.

- $f(y,\theta) = \theta\, e^{-\theta y}, \qquad \theta \in \Theta = (0,\infty), \qquad y \in \mathcal{Y} = (0,\infty).$
- $\{f(\cdot,\theta) : \theta \in \Theta\}$ is identifiable (see section 3.2).

- $f(\cdot,\cdot)$ is continuous on $\mathcal{Y} \times \Theta$ and for each $y \in \mathcal{Y}$, $f(y,\cdot)$ vanishes at infinity.

- Let $\theta \in \Theta$.

$$\begin{aligned} E_{\phi^o}\left[|\log f(Y_1,\theta)|\right] &= \int_0^\infty p_{\phi^o}(y_1)\,|\log f(y_1,\theta)|\,dy_1 \\ &= \int_0^\infty \sum_{j=1}^{K^o} \pi_j^o f(y_1,\theta_j^o)\,|\log f(y_1,\theta)|\,dy_1 \\ &= \sum_{j=1}^{K^o} \int_0^\infty \pi_j^o\,\theta_j^o\,e^{-(\theta_j^o)y_1}\,|\log\theta - \theta y_1|\,dy_1 \\ &\leq \sum_{j=1}^{K^o} \pi_j^o \left\{|\log\theta| + \int_0^\infty \theta_j^o\,e^{-(\theta_j^o)y_1}\,\theta\,y_1\,dy_1\right\} \end{aligned} \tag{4.17}$$

and

$$\begin{aligned} \int_0^\infty \theta_j^o\,e^{-(\theta_j^o)y_1}\,\theta\,y_1\,dy_1 &= \tfrac{\theta}{\theta_j^o}\int_0^\infty u e^{-u}\,du \\ &= \tfrac{\theta}{\theta_j^o}\,\Gamma(2) \\ &= \tfrac{\theta}{\theta_j^o} \end{aligned} \tag{4.18}$$

where the second equality comes from Lemma 4.1.2.

From (4.17) and (4.18), then for every $\theta \in \Theta$

$$\begin{aligned} E_{\phi^o}\left[|\log f(Y_1,\theta)|\right] &\leq \sum_{j=1}^{K^o} \pi_j^o \left\{|\log\theta| + \tfrac{\theta}{\theta_j^o}\right\} \\ &= |\log\theta| + \tfrac{\theta}{\theta_j^o} < \infty\,. \end{aligned}$$

Therefore, conditions A6 and A7 hold.

3. The family of Normal distributions with fixed (known) variance.

- $f(y,\theta) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2}$, $y \in \mathcal{Y} = \boldsymbol{R}$, $\theta \in \Theta = \boldsymbol{R}$, and $\sigma > 0$ constant (known).

- $\{f(\cdot,\theta) : \theta \in \Theta\}$ is identifiable (see section 3.2).

- $f(\cdot,\cdot)$ is continuous on $\mathcal{Y} \times \Theta$ and for each $y \in \mathcal{Y}$, $f(y,\cdot)$ vanishes at infinity.

- Let $\theta \in \Theta$.

$$\begin{aligned} E_{\phi^o}\left[|\log f(Y_1,\theta)|\right] &= \int_{-\infty}^{\infty} p_{\phi^o}(y_1)\,|\log f(y_1,\theta)|\,dy_1 \\ &= \int_{-\infty}^{\infty} \sum_{j=1}^{K^o} \pi_j^o f(y_1,\theta_j^o)\,|\log f(y_1,\theta)|\,dy_1 \\ &= \sum_{j=1}^{K^o} \pi_j^o \int_{-\infty}^{\infty} \left\{ \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1-\theta_j^o}{\sigma}\right)^2} \right. \\ &\qquad \left. \times \left| -\log \sigma\sqrt{2\pi} - \tfrac{1}{2}\left(\tfrac{y_1-\theta}{\sigma}\right)^2 \right| \right\} dy_1 \\ &\le \sum_{j=1}^{K^o} \pi_j^o \left\{ |\log \sigma\sqrt{2\pi}| \right. \\ &\qquad \left. + \int_{-\infty}^{\infty} \tfrac{1}{2}\left(\tfrac{y_1-\theta}{\sigma}\right)^2 \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1-\theta_j^o}{\sigma}\right)^2} dy_1 \right\}. \end{aligned} \tag{4.19}$$

$$\begin{aligned} \left(\tfrac{y_1-\theta}{\sigma}\right)^2 &= \left(\tfrac{y_1-\theta_j^o+\theta_j^o-\theta}{\sigma}\right)^2 \\ &= \left(\tfrac{y_1-\theta_j^o}{\sigma}\right)^2 + 2\left(\tfrac{\theta_j^o-\theta}{\sigma}\right)\left(\tfrac{y_1-\theta_j^o}{\sigma}\right) + \left(\tfrac{\theta_j^o-\theta}{\sigma}\right)^2. \end{aligned} \tag{4.20}$$

$$\begin{aligned} \int_{-\infty}^{\infty} \tfrac{1}{\sigma\sqrt{2\pi}} \tfrac{1}{2}\left(\tfrac{y_1-\theta_j^o}{\sigma}\right)^2 e^{-\frac{1}{2}\left(\frac{y_1-\theta_j^o}{\sigma}\right)^2} dy_1 &= \tfrac{1}{\sqrt{2\pi}} \int_0^{\infty} u^2 e^{-\frac{1}{2}u^2}\,du \\ &= \tfrac{1}{\sqrt{2\pi}} \cdot \tfrac{\Gamma\left(\frac{3}{2}\right)}{2\cdot\frac{1}{\sqrt{8}}} \\ &= \tfrac{1}{\sqrt{2\pi}} \cdot \tfrac{1}{4}\sqrt{8\pi} \\ &= \tfrac{1}{2} \end{aligned} \tag{4.21}$$

where the second equality comes from Lemma 4.1.2.

$$\begin{aligned} \int_{-\infty}^{\infty} \tfrac{1}{\sigma\sqrt{2\pi}} \left(\tfrac{\theta_j^o-\theta}{\sigma}\right)\left(\tfrac{y_1-\theta_j^o}{\sigma}\right) e^{-\frac{1}{2}\left(\frac{y_1-\theta_j^o}{\sigma}\right)^2} dy_1 \\ = \tfrac{1}{\sqrt{2\pi}} \left(\tfrac{\theta_j^o-\theta}{\sigma}\right) \int_{-\infty}^{\infty} u e^{-\frac{1}{2}u^2}\,du \\ = 0 \end{aligned} \tag{4.22}$$

since $g(u) = ue^{-\frac{1}{2}u^2}$ is an odd function.

From (4.19), (4.20), (4.21) and (4.22), for every $\theta \in \Theta$,

$$\begin{aligned} E_{\phi^o}\left[|\log f(Y_1,\theta)|\right] &\leq \sum_{j=1}^{K^o} \pi_j^o \left\{ |\log(\sigma\sqrt{2\pi})| + \tfrac{1}{2} + \tfrac{1}{2}\left(\tfrac{\theta_j^o - \theta}{\sigma}\right)^2 \right\} \\ &= |\log(\sigma\sqrt{2\pi})| + \tfrac{1}{2} + \tfrac{1}{2}\sum_{j=1}^{K^o} \pi_j^o \left(\tfrac{\theta_j^o - \theta}{\sigma}\right)^2 \\ &< \infty . \end{aligned}$$

Hence conditions A6 and A7 hold.

4.2 The Log-likelihood Function

For $y = \{y_t\} \in \mathcal{Y}^\infty$ and $n \in \boldsymbol{N}$, the log-likelihood function $L_n(\cdot, y)$ is defined on Φ_K by

$$\begin{aligned} L_n(\phi, y) &= \frac{1}{n}\log p_\phi(y_1,\ldots,y_n) \\ &= \frac{1}{n}\log \sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} \pi_{x_1}(\phi) f(y_1,\theta_{x_1}(\phi)) \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\phi) f(y_t,\theta_{x_t}(\phi)). \end{aligned} \tag{4.23}$$

Condition A4 ensures that for each y_t, $f(y_t,\cdot)$ vanishes at infinity. By defining

$$f(y_t,\infty) = \lim_{\theta\to\infty} f(y_t,\theta) = 0, \tag{4.24}$$

the log-likelihood function $L_n(\cdot, y)$ in (4.23) can then be extended to Φ_K^c.

Since for each $i, j = 1,\ldots,K$, the coordinate projections $\pi_i(\cdot)$, $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ are continuous on Φ_K^c by A5, and $f(\cdot,\cdot)$ continuous on $\mathcal{Y}\times\Theta^c$ by A4 and (4.24), the log-likelihood function $L_n(\cdot, y)$ is continuous on Φ_K^c.

The next lemma shows that $\{L_n(\cdot, y) : n \in \boldsymbol{N}\}$ is an equicontinuous sequence on Φ_K^c.

Lemma 4.2.1 *Assume conditions A4 and A5 hold. Then for each $y \in \mathcal{Y}$, $\{L_n(\cdot, y) : n \in \boldsymbol{N}\}$ is an equicontinuous sequence on Φ_K^c.*

Proof :

Let $y = \{y_t\} \in \mathcal{Y}$. Since Φ_K is dense in Φ_K^c, it is enough to show that $\{L_n(\cdot, y)\}$ is an equicontinuous sequence on Φ_K.

Given $\epsilon > 0$. We will prove that there exists $\delta(\epsilon) > 0$ such that for every n,

$$\left|L_n(\phi, y) - L_n(\widehat{\phi}, y)\right| < \epsilon \qquad \text{if } \phi, \widehat{\phi} \in \Phi_K, \quad \|\phi - \widehat{\phi}\|_K < \delta\,.$$

For each $t \in \boldsymbol{N}$, let $s_t = (x_t, y_t)$, where $x_t \in \{1, \ldots, K\}$ and

$$p_\phi(s_1, \ldots, s_n) = \pi_{x_1}(\phi) f(y_1, \theta_{x_1}(\phi)) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}(\phi) f(y_t, \theta_{x_t}(\phi)).$$

Then

$$p_\phi(s_1, \ldots, s_n) = \pi_{x_1}(\phi) \left(\prod_{i=1}^{K} \prod_{t \in T_i} f(y_t, \theta_i(\phi)) \right) \left(\prod_{i,j=1}^{K} \alpha_{ij}(\phi)^{N_{ij}} \right) \tag{4.25}$$

where

$$T_i = \{1 \le t \le n \ : \ \theta_t = \theta_i\}, \qquad i = 1, \ldots, K$$
$$N_{i,j} = \sum_{t=2}^{n} 1_{\{X_{t-1}=i, X_t=j\}}\,, \qquad i, j = 1, \ldots, K.$$

For each $n \in \boldsymbol{N}$, define

$$L_n(\phi, s) = \frac{1}{n} \log p_\phi(s_1, \ldots, s_n). \tag{4.26}$$

By (4.25) and (4.26), for $\phi, \widehat{\phi} \in \Phi_K$,

$$\begin{aligned} \left|L_n(\phi, s) - L_n(\widehat{\phi}, s)\right| \ \le \ & \frac{1}{n} \left|\log \pi_{x_1}(\phi) - \log \pi_{x_1}(\widehat{\phi})\right| \\ & + \frac{1}{n} \sum_{i=1}^{K} \sum_{t \in T_i} \left|\log f(y_t, \theta_i(\phi)) - \log f(y_t, \theta_i(\widehat{\phi}))\right| \\ & + \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{K} N_{i,j} \left|\log \alpha_{i,j}(\phi) - \log \alpha_{ij}(\widehat{\phi})\right|. \end{aligned} \tag{4.27}$$

Since K is finite, then by A5, there is $\delta_1 > 0$, such that for every $i = 1, \ldots, K$,

$$\left|\log \pi_i(\phi) - \log \pi_i(\widehat{\phi})\right| < \frac{\epsilon}{3}, \qquad \text{if } \phi, \widehat{\phi} \in \Phi_K, \quad \|\phi - \widehat{\phi}\|_K < \delta_1 . \tag{4.28}$$

Also by A4, there is $\delta_2 > 0$, such that, for every $i = 1, \ldots, K$ and for every $t \in T_i$,

$$\left|\log f(y_t, \theta_i(\phi)) - \log f(y_t, \theta_i(\widehat{\phi}))\right| < \frac{\epsilon}{3}, \tag{4.29}$$

if

$$\begin{aligned} \|(y_t, \theta_i(\phi)) - (y_t, \theta_i(\widehat{\phi}))\| &= \|y_t - y_t\| + \|\theta_i(\phi) - \theta_i(\widehat{\phi})\|_\infty \\ &= \|\theta_i(\phi) - \theta_i(\widehat{\phi})\|_\infty \\ &< \delta_2 . \end{aligned} \tag{4.30}$$

However, by A5, there is $\delta_3 > 0$, such that for every $i = 1, \ldots, K$,

$$\|\theta_i(\phi) - \theta_i(\widehat{\phi})\|_\infty < \delta_2 \qquad \text{if } \phi, \widehat{\phi} \in \Phi_K, \quad \|\phi - \widehat{\phi}\|_K < \delta_3 . \tag{4.31}$$

Moreover, by A5, there is $\delta_4 > 0$, such that for every $i, j = 1, \ldots, K$,

$$\left|\log \alpha_{ij}(\phi) - \log \alpha_{ij}(\widehat{\phi})\right| < \frac{e}{3}, \qquad \text{if } \phi, \widehat{\phi} \in \Phi_K, \quad \|\phi - \widehat{\phi}\|_K < \delta_4 . \tag{4.32}$$

Let $0 < \delta < \min\{\delta_1, \delta_3, \delta_4\}$. Then by (4.27), (4.28), (4.29), (4.30), (4.31) and (4.32), if $\phi, \widehat{\phi} \in \Phi_K$ and $\|\phi - \widehat{\phi}\|_K < \delta$, then for every $n \in \boldsymbol{N}$,

$$\begin{aligned} \|L_n(\phi, s) - L_n(\widehat{\phi}, s)\| &< \frac{1}{n} \cdot \frac{\epsilon}{3} + \frac{1}{n} \cdot n \cdot \frac{\epsilon}{3} + \frac{1}{n} \cdot (n-1) \cdot \frac{\epsilon}{3} \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\ &= \epsilon \end{aligned} \tag{4.33}$$

The equation (4.33) can be written as

$$\left|\frac{1}{n} \log \frac{p_\phi(s_1, \ldots, s_n)}{p_{\widehat{\phi}}(s_1, \ldots, s_n)}\right| < \epsilon.$$

Therefore,

$$p_\phi(s_1,\ldots,s_n) < \exp(n\epsilon)p_{\widehat{\phi}}(s_1,\ldots,s_n),$$

implying

$$\begin{aligned} p_\phi(y_1,\ldots,y_n) &= \sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} p_\phi(x_1,y_1,\ldots,x_n,y_n) \\ &= \sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} p_\phi(s_1,\ldots,s_n) \\ &< \sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} \exp(n\epsilon)p_{\widehat{\phi}}(s_1,\ldots,s_n) \\ &= \exp(n\epsilon)\sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} p_{\widehat{\phi}}(x_1,y_1,\ldots,x_n,y_n) \\ &= \exp(n\epsilon)p_{\widehat{\phi}}(y_1,\ldots,y_n). \end{aligned}$$

Similarly by exchanging the roles of ϕ and $\widehat{\phi}$, we have for every $n \in \boldsymbol{N}$,

$$\left|L_n(\phi,y) - L_n(\widehat{\phi},y)\right| = \left|\frac{1}{n}\log\frac{p_\phi(y_1,\ldots,y_n)}{p_{\widehat{\phi}}(y_1,\ldots,y_n)}\right| < \epsilon,$$

if $\phi,\widehat{\phi} \in \Phi_K$, and $\|\phi - \widehat{\phi}\|_K < \delta$. ■

Let $Y = \{Y_t\}$. For each $\phi \in \Phi_K^c$, define

$$L_n(\phi,Y) = \frac{1}{n}\log p_\phi(Y_1,\ldots,Y_n), \qquad n \in \boldsymbol{N}. \tag{4.34}$$

Notice that $\{L_n(\phi,Y) : n \in \boldsymbol{N}\}$ is a stochastic process defined on $(\mathcal{Y}^\infty, \mathcal{B}, P_{\phi^\circ})$. Such process will be called the *likelihood-process.*

As a consequence of Lemma 4.2.1, if conditions A4 and A5 hold, then we have that for every $\epsilon > 0$, there is $\delta(\epsilon,y) > 0$, such that for every n,

$$\left|L_n(\phi,Y(y)) - L_n(\widehat{\phi},Y(y))\right| < \epsilon \qquad \text{if } \phi,\widehat{\phi} \in \Phi_K, \quad \|\phi - \widehat{\phi}\|_K < \delta\,.$$

So the corollary below follows.

Corollary 4.2.2 *Assume conditions A4 and A5 hold, then* $\{L_n(\cdot,Y) : n \in \boldsymbol{N}\}$ *is an equicontinuous sequence on* Φ_K^c.

4.3 Kullback- Leibler Divergence

This section is divided into two subsections. The first subsection discusses Kullback-Leibler divergence in general. Here, definitions and characteristics of Kullback-Leibler divergence are given. Using definition in subsection 4.3.1, the Kullback-Leibler divergence for hidden Markov models is then derived in subsection 4.3.2.

4.3.1 General case

This subsection begins by defining absolute continuity between two measures.

Definition 4.3.1 *Let λ an ν be two measures defined on a measurable space $(\Omega, \mathcal{F})$. A measure ν is said to be* ***absolutely continuous*** *with respect to the measure λ, if for each $A \in \mathcal{F}$, $\lambda(A) = 0$ implies $\nu(A) = 0$. The relation is indicated by $\nu \ll \lambda$. If $\nu \ll \lambda$, it is said that ν is* ***dominated*** *by λ.*

The Radon-Nikodym theorem (e.g. [44], page 276) states that if $(\Omega, \mathcal{F}, \lambda)$ is a σ-finite measure space and ν is a measure defined on $\mathcal{F}$ such that $\nu \ll \lambda$, then there is a non-negative measurable function g, such that for each set $A \in \mathcal{F}$,

$$\nu(A) = \int_A g \, d\lambda.$$

The function g is *unique* in the sense that, if h is any measurable function with this property, then $h = g$ λ-almost sure. The function g is called the *Radon-Nikodym derivative* of ν with respect to λ and denoted by $\frac{d\nu}{d\lambda}$.

Now we are ready to define Kullback-Leibler divergence. According to [2], the Kullback-Leibler divergence is defined as follows.

Definition 4.3.2 *Let P and Q be two probability measures defined on the measurable space $(\boldsymbol{R}^n, \mathcal{R}^n)$ dominated by a measure ν. Let*

$$p = \frac{dP}{d\nu} \qquad \text{and} \qquad q = \frac{dQ}{d\nu}.$$

Suppose that q is ν-almost sure strictly positive. The ***Kullback-Leibler divergence*** *of Q with respect to P is defined by*

$$K(P,Q) = \begin{cases} \int p \log \frac{p}{q} \, d\nu & , \text{ if } P \ll Q \\ +\infty & , \text{ otherwise.} \end{cases} \tag{4.35}$$

From (4.35), the Kullback-Leibler divergence of Q with respect to P can be expressed as

$$K(P,Q) = \begin{cases} E_P \left[\log \frac{p}{q} \right] & , \text{ if } P \ll Q \\ +\infty & , \text{ otherwise} \end{cases} \tag{4.36}$$

where E_P is the expectation with respect to P. Notice that $\frac{p}{q} = \frac{dP}{dQ}$, so (4.36) can also be written as

$$K(P,Q) = \begin{cases} E_P \left[\log \frac{dP}{dQ} \right] & , \text{ if } P \ll Q \\ +\infty & , \text{ otherwise.} \end{cases} \tag{4.37}$$

Lemma 4.3.3 *Let P and Q be two probability measures defined on the measurable space $(\boldsymbol{R}^n, \mathcal{R}^n)$ dominated by a measure ν. Let*

$$p = \frac{dP}{d\nu} \qquad \text{and} \qquad q = \frac{dQ}{d\nu}.$$

Suppose that q is ν-almost sure strictly positive. Then

$$K(P,Q) \geq 0 \quad \text{and} \quad K(P,Q) = 0 \quad \text{if and only if} \quad P = Q.$$

Proof :

Assume without loss of generality that $P \ll Q$. Let

$$g(x) = x \log x + 1 - x, \qquad x > 0.$$

Since

$$g'(x) = \log x, \qquad x > 0,$$

then $g(x)$ is always positive and zero only for $x = 1$. Notice that

$$\begin{aligned}
\int \left(\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}\right) dQ &= \int \left(\frac{p}{q}\log\frac{p}{q} + 1 - \frac{p}{q}\right) q\, d\nu \\
&= \int \left(p\log\frac{p}{q} + q - p\right) d\nu \\
&= \int p\log\frac{p}{q}\, d\nu + \int q\, d\nu - \int p\, d\nu \\
&= K(P,Q) + \int dQ - \int dP \\
&= K(P,Q) + 1 - 1 \\
&= K(P,Q). \qquad (4.38)
\end{aligned}$$

By (4.38) and the characteristics of g, we have $K(P,Q) \geq 0$ and $K(P,Q) = 0$ if and only if $\frac{p}{q} = 1$, Q-almost sure, which implies $P = Q$. ■

Remarks 4.3.4 Notice that in general the Kullback-Leibler divergence $K(\cdot,\cdot)$ is not symmetric, so it is not a *metric*.

Definition 4.3.2 can be extended for probability measures defined on $(\boldsymbol{R}^\infty, \boldsymbol{\mathcal{R}}^\infty)$. Using projections, the Kullback-Leibler divergence of probability measures on $(\boldsymbol{R}^\infty, \boldsymbol{\mathcal{R}}^\infty)$ can be defined in the following way.

Let P be a probability measures defined on $(\boldsymbol{R}^\infty, \boldsymbol{\mathcal{R}}^\infty)$. For each $n \in \boldsymbol{N}$, let P_n be a probability measure on $(\boldsymbol{R}^n, \boldsymbol{\mathcal{R}}^n)$ which is defined by

$$P_n\Big\{(y_1,\ldots,y_n) \in \boldsymbol{R}^n : (y_1,\ldots,y_n) \in A\Big\} = P\Big\{\{y_t\} \in \boldsymbol{R}^\infty : (y_1,\ldots,y_n) \in A\Big\}.$$

The probability measures P_n, $n \in \boldsymbol{N}$, are called the *projections* of P on $(\boldsymbol{R}^\infty, \boldsymbol{\mathcal{R}}^\infty)$.

Definition 4.3.5 *Let P and Q be two probability measures defined on the measurable space $(\boldsymbol{R}^\infty, \boldsymbol{\mathcal{R}}^\infty)$, with projections P_n and Q_n on $(\boldsymbol{R}^n, \boldsymbol{\mathcal{R}}^n)$. The* ***Kullback-Leibler divergence*** *of Q with respect to P is defined by*

$$K(P,Q) = \lim_{n\to\infty} \frac{1}{n} K(P_n, Q_n), \tag{4.39}$$

if this limit exists.

Lemma 4.3.6 *Let P and Q be two probability measures defined on the measurable space $(\boldsymbol{R}^\infty, \boldsymbol{\mathcal{R}}^\infty)$, then $K(P,Q) \geq 0$. If $P = Q$, then $K(P,Q) = 0$*

Proof :

For each $n \in \boldsymbol{N}$, let P_n and Q_n be the projections of P and Q on $(\boldsymbol{R}^n, \boldsymbol{\mathcal{R}}^n)$. By Lemma 4.3.3,

$$K(P_n, Q_n) \geq 0, \qquad n \in \boldsymbol{N},$$

implying

$$K(P,Q) = \lim_{n\to\infty} \frac{1}{n} K(P_n, Q_n) \geq 0.$$

If $P = Q$, it is clear that $P_n = Q_n$, $n \in \boldsymbol{N}$, implying $K(P_n, Q_n) = 0$, $n \in \boldsymbol{N}$, thus $K(P,Q) = 0$. ■

4.3.2 Hidden Markov case

The idea of using Kullback-Leibler divergence to measure a distance between $\phi \in \Phi_K^c$ and the true parameter ϕ^o comes from [22], who used it for a hidden Markov model, in which the observation process takes values on a finite set.

Recall that under the true parameter $\phi^o = (K^o, A^o, \pi^o, \theta^o)$, $Y = \{Y_t\}$ is defined as a coordinate projection process on $(\mathcal{Y}^\infty, \mathcal{B}, P_{\phi^o})$ having n-dimensional joint

density function

$$p_{\phi^o}(y_1,\ldots,y_n)=\sum_{x_1=1}^{K^o}\cdots\sum_{x_n=1}^{K^o}\pi^o_{x_1}f(y_1,\theta^o_{x_1})\prod_{t=2}^{n}\alpha^o_{x_{t-1},x_t}f(y_t,\theta^o_{x_t}), \tag{4.40}$$

with respect to the measure μ. Also under each $\phi\in\Phi^c_K$, Y is defined as a coordinate projection process on $(\mathcal{Y}^\infty,\mathcal{B},P_\phi)$, having n-dimensional density function

$$p_\phi(y_1,\ldots,y_n)=\sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K}\pi_{x_1}(\phi)f(y_1,\theta_{x_1}(\phi))\prod_{t=2}^{n}\alpha_{x_{t-1},x_t}(\phi)f(y_t,\theta_{x_t}(\phi)), \tag{4.41}$$

with respect to the measure μ.

Define a distance between ϕ^o and $\phi\in\Phi^c_k$, as the Kullback-Leibler divergence of P_ϕ with respect to P_{ϕ^o}, that is by (4.35), (4.36), (4.40) and (4.41),

$$K(\phi^o,\phi) = \lim_{n\to\infty}\frac{1}{n}\int p_{\phi^o}(y_1,\ldots,y_n)\log\frac{p_{\phi^o}(y_1,\ldots,y_n)}{p_\phi(y_1,\ldots,y_n)}\,d\mu(y_1)\cdots d\mu(y_n) \tag{4.42}$$

$$= \lim_{n\to\infty}\frac{1}{n}E_{\phi^o}\left[\log\frac{p_{\phi^o}(Y_1,\ldots,Y_n)}{p_\phi(Y_1,\ldots,Y_n)}\right], \tag{4.43}$$

if this limits exist.

By Lemma 4.3.6, it is clear that $K(\phi^o,\phi)\geq 0$, for every $\phi\in\Phi^c_K$.

4.4 Relation between the Kullback-Leibler Divergence and the Log-likelihood Process

The main issue of this section is to find relation between the log-likelihood process

$$L_n(\phi,Y)=\frac{1}{n}\log p_\phi(Y_1,\ldots,Y_n),\qquad n\in\boldsymbol{N} \tag{4.44}$$

and the Kullback-Leibler divergence

$$K(\phi^o, \phi) = \lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\left[\log \frac{p_{\phi^o}(Y_1, \ldots, Y_n)}{p_\phi(Y_1, \ldots, Y_n)}\right], \tag{4.45}$$

for $\phi \in \Phi_K^c$.

In this section we will adapt the work of Leroux [34] to our case. From (4.45), the Kullback-Leibler divergence of $\phi \in \Phi_K^c$ with respect to ϕ^o can be expressed as

$$K(\phi^o, \phi) = \lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\left[\log p_{\phi^o}(Y_1, \ldots, Y_n)\right] - \lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\left[\log p_\phi(Y_1, \ldots, Y_n)\right], \tag{4.46}$$

provided the two limits in the right hand side exist. So the main interest now is to investigate the characteristics of these limits.

By condition A1 and A2, $\{Y_t\}$, under the true parameter ϕ^o is stationary and ergodic. As in section 2.7, we can give a past to $\{Y_t\}$ without destroying its stationarity. So, without lost of generality, we may consider $\{Y_t\}$ as a stationary and ergodic process indexed by $t \in \boldsymbol{Z}$.

Define the entropy of $\{Y_t\}$ as follows.

Definition 4.4.1 *The* ***entropy*** *of the stationary process* $\{Y_t\}$ *under the true parameter* ϕ^o *is defined by*

$$H(\phi^o) = -E_{\phi^o}\Big[\log p_{\phi^o}(Y_1 | Y_0, Y_{-1}, \ldots)\Big]. \tag{4.47}$$

In order for this definition to have meaning, we must show the existence of the conditional density $p_{\phi^o}(Y_1 \,|\, Y_0, Y_{-1}, \ldots)$.

From sections 2.3 and 2.4,

$$p_{\phi^o}(Y_1 | Y_0, \ldots, Y_{-n}) = \sum_{i=1}^{K^o} P_{\phi^o}(X_1 = i | Y_0, \ldots, Y_{-n}) f(Y_1, \theta_i^o). \tag{4.48}$$

Lévy martingale convergence theorem ([48], page 478) states that if Z is an integrable random variable and $\{\mathcal{F}_t\}$ is an increasing sequence of σ-fields, then

$$\lim_{n\to\infty} E[Z|\mathcal{F}_n] = E[Z|\mathcal{F}_\infty],$$

with probability one, where $\mathcal{F}_\infty$ is the σ-field generated by $\bigcup_t \mathcal{F}_t$. Applying the theorem for

$$Z = I_{\{X_1=i\}} \quad \text{and} \quad \mathcal{F}_n = \sigma(Y_0, \ldots, Y_{-n})$$

gives

$$\lim_{n\to\infty} P_{\phi^o}(X_1 = i|Y_0, \ldots, Y_{-n}) = P_{\phi^o}(X_1 = i|Y_0, Y_{-1}, \ldots), \tag{4.49}$$

with probability one under ϕ^o. Define

$$p_{\phi^o}(Y_1|Y_0, Y_{-1}, \ldots) = \sum_{i=1}^{K^o} P_{\phi^o}(X_1 = i|Y_0, Y_{-1}, \ldots) f(Y_1, \theta_i^o). \tag{4.50}$$

Then by (4.48), (4.49) and (4.50),

$$\lim_{n\to\infty} p_{\phi^o}(Y_1|Y_0, \ldots, Y_{-n}) = p_{\phi^o}(Y_1|Y_0, Y_{-1}, \ldots), \tag{4.51}$$

with probability one under ϕ^o.

The characteristics of the first limit of (4.46) are given by the following theorem.

Theorem 4.4.2 (Leroux [34]) *If conditions A1, A2 and A6 hold, then*

(a). $H(\phi^o) = -E_{\phi^o}\left[\log p_{\phi^o}(Y_1|Y_0, Y_{-1}, \ldots)\right]$ *is finite.*

(b). $\lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\left[\log p_{\phi^o}(Y_1, \ldots, Y_n)\right] = -H(\phi^o).$

(c). $\lim_{n\to\infty} \frac{1}{n} \log p_{\phi^o}(Y_1, \ldots, Y_n) = -H(\phi^o)$ *with probability one under* ϕ^o.

Proof :

To prove (a), from (4.48), for $n = 0, 1, \ldots,$

$$p_{\phi^o}(Y_1|Y_0, \ldots, Y_{-n}) = \sum_{i=1}^{K^o} P_{\phi^o}(X_1 = i|Y_0, \ldots, Y_{-n}) f(Y_1, \theta_i^o),$$

implying

$$\min_{1\le i\le K^o} f(Y_1,\theta_i^o) \le p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-n}) \le \max_{1\le i\le K^o} f(Y_1,\theta_i^o).$$

Since

$$E_{\phi^o}\Big[|\log f(Y_1,\theta_i^o)|\Big] < \infty, \qquad \text{for } i=1,2,\ldots,K^o,$$

by A6, then $\{\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-n}) : n=0,1,\ldots\}$ is uniformly integrable under ϕ^o. Also from (4.51),

$$\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-n}) \longrightarrow \log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots), \tag{4.52}$$

with probability one under ϕ^o, when $t\to\infty$, implying $\log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots)$ is integrable and (4.52) also holds in L^1, that is,

$$\begin{aligned}\lim_{n\to\infty} E_{\phi^o}\Big[\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-n})\Big] &= E_{\phi^o}\Big[\lim_{n\to\infty}\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-n})\Big]\\ &= E_{\phi^o}\Big[\log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots)\Big]\\ &= -H(\phi^o), \end{aligned}\tag{4.53}$$

which is finite.

To prove (b), using Cesaro convergence theorem ([16], page 83), stationarity of $\{Y_t\}$ and (4.53),

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}E_{\phi^o}\Big[\log p_{\phi^o}(Y_1,\ldots,Y_n)\Big] &= \lim_{n\to\infty}\frac{1}{n}E_{\phi^o}\left[\log\prod_{t=1}^{n} p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)\right]\\ &= \lim_{n\to\infty}\frac{1}{n}E_{\phi^o}\left[\sum_{t=1}^{n}\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)\right]\\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}E_{\phi^o}\Big[\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)\Big]\\ &= \lim_{t\to\infty}E_{\phi^o}\Big[\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)\Big]\\ &= \lim_{t\to\infty}E_{\phi^o}\Big[\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-t})\Big]\\ &= -H(\phi^o).\end{aligned}$$

To prove (c), consider

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\log p_{\phi^o}(Y_1,\ldots,Y_n)\\
&\quad=\lim_{n\to\infty}\frac{1}{n}\log\prod_{t=1}^{n}p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)\\
&\quad=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)\\
&\quad=\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{t=1}^{n}\log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots)+\right.\\
&\qquad\left.\frac{1}{n}\sum_{t=1}^{n}\Big(\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)-\log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots)\Big)\right\}. \qquad (4.54)
\end{aligned}$$

Since $\{Y_t\}$ is stationary and ergodic under ϕ^o, by A1 and A2, then $\{\log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots) : t\in \boldsymbol{N}\}$ is also stationary and ergodic under ϕ^o, which by ergodic theorem ([31], page 488) implies

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots) &= E_{\phi^o}\Big[\log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots)\Big]\\
&= -H(\phi^o), \qquad (4.55)
\end{aligned}$$

with probability one under ϕ^o.

Following [31], page 502, we will prove that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\Big(\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1)-\log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots)\Big)=0. \qquad (4.56)$$

Let N be any positive integer. Set

$$\lambda_N(Y_1,Y_0,\ldots)=\sup_{t\geq N}|\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-t})-\log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots)|. \qquad (4.57)$$

Let

$$Z_t^N=\lambda_N(Y_t,Y_{t-1},\ldots), \qquad t\in \boldsymbol{Z}.$$

Since $\{Y_t\}$ is stationary and ergodic under ϕ^o, then the process $\{Z_t^N : t\in \boldsymbol{Z}\}$ is also stationary and ergodic under ϕ^o. Moreover,

$$\begin{aligned}
E_{\phi^o}\left[Z_1^N\right] &= E_{\phi^o}\left[\sup_{t\geq N}|\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-t})-\log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots)|\right]\\
&< \infty,
\end{aligned}$$

as

$$\log p_{\phi^o}(Y_1|Y_0,\ldots,Y_{-t}) \longrightarrow \log p_{\phi^o}(Y_1|Y_0,Y_{-1},\ldots), \tag{4.58}$$

with probability one under ϕ^o and in L^1, when $t \to \infty$. The ergodic theorem then implies

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} Z_t^N = E_{\phi^o}\left[Z_1^N\right].$$

and hence

$$\begin{aligned}
&\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{t=1}^{n}\left(\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1) - \log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots)\right)\right| \\
&\quad\leq \limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left|\log p_{\phi^o}(Y_t|Y_{t-1},\ldots,Y_1) - \log p_{\phi^o}(Y_t|Y_{t-1},Y_{t-2},\ldots)\right| \\
&\quad\leq \limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} Z_t^N \\
&\quad= E_{\phi^o}\left[Z_1^N\right].
\end{aligned} \tag{4.59}$$

for any positive integer N.

However, by definition of Z_1^N and by (4.58),

$$Z_1^N \searrow 0,$$

with probability one under ϕ^o, when $N \to \infty$. By monotone convergence theorem,

$$\begin{aligned}
\lim_{N\to\infty} E_{\phi^o}\left[Z_1^N\right] &= E_{\phi^o}\left[\lim_{N\to\infty} Z_1^N\right] \\
&= E_{\phi^o}[0] \\
&= 0.
\end{aligned} \tag{4.60}$$

Hence by (4.59) and (4.60), (4.56) follows.

Combining (4.54), (4.55) and (4.56), we have

$$\lim_{n\to\infty}\frac{1}{n}\log p_{\phi^o}(Y_1,\ldots,Y_n) = -H(\phi^o),$$

with probability one under ϕ^o. ■

The next theorem shows the characteristics of the second limit of (4.46)

Theorem 4.4.3 *Assume conditions A1, A2, A4, A5 and A7 hold. Let* $\phi \in \Phi_K^c$, *with* $\pi_i(\phi) > 0$, *for* $i = 1, \ldots, K$. *Then for* ϕ, *there is a constant* $H(\phi^o, \phi)$, *such that :*

(a). $-\infty \leq H(\phi^o, \phi) < \infty$

(b). $\lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\big[\log p_\phi(Y_1, \ldots, Y_n)\big] = H(\phi^o, \phi)$

(c). $\lim_{n\to\infty} \frac{1}{n} \log p_\phi(Y_1, \ldots, Y_n) = H(\phi^o, \phi)$ *with probability one under* ϕ^o.

Proof :

Since Φ_K dense in Φ_K^c, then to prove the theorem, it is sufficient to prove only for $\phi \in \Phi_K$, with $\pi_i(\phi) > 0$, for $i = 1, \ldots, K$. Here we use the proof of Theorem 2 in [34].

Let $\phi \in \Phi_K$ with $\pi_i(\phi) > 0$, for $i = 1, \ldots, K$. Let

$$\pi_i(\phi) = \pi_i, \qquad \alpha_{ij}(\phi) = \alpha_{ij}, \qquad \theta_i(\phi) = \theta_i,$$

for $i, j = 1, \ldots, K$.

For $i = 1, \ldots, K$, $m, n \in \boldsymbol{Z}$, with $m < n$ and any realization $\{(x_t, y_t)\}$, define

$$\begin{aligned} q_\phi(y_{m+1}, \ldots, y_n|i) &= f(y_{m+1}, \theta_i) \sum_{x_{m+2}=1}^{K} \cdots \sum_{x_n=1}^{K} \alpha_{i,x_{m+2}} f(y_{m+2}, \theta_{m+2}) \\ &\quad \times \prod_{t=m+3}^{n} \alpha_{x_{t-1},x_t} f(y_t, \theta_{x_t}). \end{aligned} \tag{4.61}$$

Notice that for $i = 1, \ldots, K$ and $n \in \boldsymbol{N}$,

$$q_\phi(y_1, \ldots, y_n|i) = p_\phi(y_1, \ldots, y_n|i), \tag{4.62}$$

that is, the conditional density of $Y_1, \ldots, Y_n$ given $X_1 = i$, under ϕ. Hence from (4.62), the joint density function of $Y_1, \ldots, Y_n$ under ϕ can be expressed

as

$$\begin{aligned} p_\phi(y_1,\ldots,y_n) &= \sum_{i=1}^{K} \pi_i\, p_\phi(y_1,\ldots,y_n|i). \\ &= \sum_{i=1}^{K} \pi_i\, q_\phi(y_1,\ldots,y_n|i). \end{aligned} \tag{4.63}$$

Define for $m, n \in \boldsymbol{N}$ and $m < n$,

$$q_\phi(y_{m+1},\ldots,y_n) = \max_{1\le i\le K} q_\phi(y_{m+1},\ldots,y_n|i). \tag{4.64}$$

Then from (4.63) and (4.64),

$$\begin{aligned} p_\phi(y_1,\ldots,y_n) &\le \sum_{i=1}^{K} \pi_i \left(\max_{1\le i\le K} q_\phi(y_1,\ldots,y_n|i)\right) \\ &= q_\phi(y_1,\ldots,y_n)\sum_{i=1}^{K}\pi_i \\ &= q_\phi(y_1,\ldots,y_n). \end{aligned} \tag{4.65}$$

and

$$\begin{aligned} p_\phi(y_1,\ldots,y_n) &\ge \sum_{i=1}^{K} \left(\min_{1\le i\le K} \pi_i\right) q_\phi(y_1,\ldots,y_n|i) \\ &= \left(\min_{1\le i\le K} \pi_i\right) \sum_{i=1}^{K} q_\phi(y_1,\ldots,y_n|i) \\ &\ge \left(\min_{1\le i\le K} \pi_i\right) \left(\max_{1\le i\le K} q_\phi(y_1,\ldots,y_n|i)\right) \\ &= \left(\min_{1\le i\le K} \pi_i\right) q_\phi(y_1,\ldots,y_n). \end{aligned} \tag{4.66}$$

Therefore from (4.65) and (4.66), for each $n \in \boldsymbol{N}$ and any realization $\{y_t\}$,

$$\left(\min_{1\le i\le K} \pi_i\right) q_\phi(y_1,\ldots,y_n) \le p_\phi(y_1,\ldots,y_n) \le q_\phi(y_1,\ldots,y_n),$$

implying

$$\left(\min_{1\le i\le K} \pi_i\right) q_\phi(Y_1,\ldots,Y_n) \le p_\phi(Y_1,\ldots,Y_n) \le q_\phi(Y_1,\ldots,Y_n).$$

Thus for each $n \in \boldsymbol{N}$,

$$\begin{aligned} \frac{1}{n}\log\left(\min_{1\le i\le K} \pi_i\right) + \frac{1}{n}\log q_\phi(Y_1,\ldots,Y_n) &\le \frac{1}{n}\log p_\phi(Y_1,\ldots,Y_n) \\ &\le \frac{1}{n}\log q_\phi(Y_1,\ldots,Y_n) \end{aligned} \tag{4.67}$$

and

$$\begin{aligned}\frac{1}{n}\log\left(\min_{1\le i\le K}\pi_i\right)+\frac{1}{n}E_{\phi^\circ}\Big[\log q_\phi(Y_1,\ldots,Y_n)\Big] &\le \frac{1}{n}E_{\phi^\circ}\Big[\log p_\phi(Y_1,\ldots,Y_n)\Big]\\ &\le \frac{1}{n}E_{\phi^\circ}\Big[\log q_\phi(Y_1,\ldots,Y_n)\Big].\end{aligned} \tag{4.68}$$

As π_i are fixed, for $i=1,\ldots,K$ and

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\min_{1\le i\le K}\pi_i\right)=0,$$

then taking $n\to\infty$ on (4.67) and (4.68) gives

$$\lim_{n\to\infty}\frac{1}{n}\log p_\phi(Y_1,\ldots,Y_n)=\lim_{n\to\infty}\frac{1}{n}\log q_\phi(Y_1,\ldots,Y_n) \tag{4.69}$$

and

$$\lim_{n\to\infty}\frac{1}{n}E_{\phi^\circ}\Big[\log p_\phi(Y_1,\ldots,Y_n)\Big]=\lim_{n\to\infty}\frac{1}{n}E_{\phi^\circ}\Big[\log q_\phi(Y_1,\ldots,Y_n)\Big], \tag{4.70}$$

provided the separate limits exist. Therefore, the conclusions of the theorem will follow from the corresponding conclusions applied to q_ϕ.

Let $l,m,n\in\boldsymbol{Z}$, where $m<l<n$, then for $i=1,\ldots,K$,

$$\begin{aligned}q_\phi(y_{m+1},\ldots,y_n|i) &= f(y_{m+1},\theta_i)\sum_{x_{m+2}=1}^{K}\cdots\sum_{x_n=1}^{K}\alpha_{i,x_{m+2}}f(y_{m+2},\theta_{x_{m+2}})\\ &\quad\times\prod_{t=m+3}^{n}\alpha_{x_{t-1},x_t}f(y_t,\theta_{x_t})\\ &= f(y_{m+1},\theta_i)\sum_{x_{m+2}=1}^{K}\cdots\sum_{x_l=1}^{K}\alpha_{i,x_{m+2}}f(y_{m+2},\theta_{x_{m+2}})\\ &\quad\times\prod_{t=m+3}^{l}\alpha_{x_{t-1},x_t}f(y_t,\theta_{x_t})\\ &\quad\times\sum_{j=1}^{K}\sum_{x_{l+2}=1}^{K}\cdots\sum_{x_n=1}^{K}\alpha_{x_l,j}f(y_{l+1},\theta_j)\alpha_{j,x_{l+2}}f(y_{l+2},\theta_{x_{l+2}})\\ &\quad\times\prod_{s=l+3}^{n}\alpha_{x_{s-1},x_s}f(y_s,\theta_{x_s})\\ &= f(y_{m+1},\theta_i)\sum_{x_{m+2}=1}^{K}\cdots\sum_{x_l=1}^{K}\alpha_{i,x_{m+2}}f(y_{m+2},\theta_{x_{m+2}})\end{aligned}$$

$$
\begin{aligned}
&\times \prod_{t=m+3}^{l} \alpha_{x_{t-1},x_t} f(y_t, \theta_{x_t}) \sum_{j=1}^{K} \alpha_{x_l,j}\, q_\phi(y_{l+1}, \ldots, y_n | j) \\
&\leq f(y_{m+1}, \theta_i) \sum_{x_{m+2}=1}^{K} \cdots \sum_{x_l=1}^{K} \alpha_{i,x_{m+2}} f(y_{m+2}, \theta_{x_{m+2}}) \\
&\quad \times \prod_{t=m+3}^{l} \alpha_{x_{t-1},x_t} f(y_t, \theta_{x_t})\, q_\phi(y_{l+1}, \ldots, y_n) \left(\sum_{j=1}^{K} \alpha_{x_l,j} \right) \\
&= q_\phi(y_{m+1}, \ldots, y_l | i) \cdot q_\phi(y_{l+1}, \ldots, y_n) \\
&\leq q_\phi(y_{m+1}, \ldots, y_l) \cdot q_\phi(y_{l+1}, \ldots, y_n). \qquad (4.71)
\end{aligned}
$$

Since (4.71) holds for every $i = 1, \ldots, K$, then

$$
q_\phi(y_{m+1}, \ldots, y_n) \leq q_\phi(y_{m+1}, \ldots, y_l) \cdot q_\phi(y_{l+1}, \ldots, y_n). \qquad (4.72)
$$

Equation (4.72) holds for any $\{y_t\}$, implying

$$
q_\phi(Y_{m+1}, \ldots, Y_n) \leq q_\phi(Y_{m+1}, \ldots, Y_l) \cdot q_\phi(Y_{l+1}, \ldots, Y_n), \qquad (4.73)
$$

for $m < l < n$.

Now define a doubly indexed sequence of random variables $W = \{W_{s,t} : s, t \in \boldsymbol{Z}, s < t\}$ on $(\mathcal{Y}^\infty, \mathcal{B}, P_\phi^o)$ by

$$
W_{s,t} = \log q_\phi(Y_{s+1}, \ldots, Y_t), \qquad \text{for } s < t. \qquad (4.74)
$$

From (4.73), for $m < l < n$,

$$
\begin{aligned}
W_{m,n} &= \log q_\phi(Y_{m+1}, \ldots, Y_l, Y_{l+1}, \ldots, Y_n) \\
&\leq \log q_\phi(Y_{m+1}, \ldots, Y_l) + \log q_\phi(Y_{l+1}, \ldots, Y_n) \\
&= W_{m,l} + W_{l,n}. \qquad (4.75)
\end{aligned}
$$

Since $\{Y_t\}$ is stationary and ergodic under ϕ^o by A1 and A2; and $\{W_{s,t}\}$ are functions of $\{Y_t\}$, then $\{W_{s,t}\}$ is also stationary and ergodic under ϕ^o, relative to the shift transformation

$$
\{w_{s,t}\} \mapsto \{w_{s+1,t+1}\}. \qquad (4.76)
$$

Moreover,

$$\begin{aligned} E_{\phi^o}\left[W_{0,1}^{+}\right] &= E_{\phi^o}\left[(\ log q_{\phi}(Y_1))^{+}\right] \\ &= E_{\phi^o}\left[\left\{\log\left(\max_{1\leq i\leq K} f(Y_1,\theta_i)\right)\right\}^{+}\right] \\ &< \infty. \end{aligned}$$

by condition A7.

Kingman ([33], theorem 1.5 and theorem 1.8) proved that a process $\{W_{s,t} : s,t \in \boldsymbol{Z}, s<t\}$ defined on a probability space $(\Omega,\mathcal{F},P)$ and satisfying :

(a). $W_{m,n} \leq W_{m,l} + W_{l,n}$, for $m<l<n$

(b). $\{W_{s,t}\}$ is stationary relative to the shift transformation (4.76)

(c). $E\left[W_{0,1}^{+}\right] < \infty$,

also satisfies the conclusions of the ergodic theorem, namely,

(a). $\lim_{n\to\infty} \frac{1}{n} W_{0,n} = W$ exists, with probability one, where $-\infty \leq W < \infty$.

(b). $E[W] = \lim_{n\to\infty} \frac{1}{n} E\left[W_{0,n}\right]$.

(c). W is degenerate (constant), if the process $\{W_{s,t}\}$ is ergodic: that is, the σ-field of events invariant under the shift transformation in (4.76) is trivial.

An application of Kingman ergodic theorem to

$$W_{0,n} = \log q_{\phi}(Y_1,\ldots,Y_n), \qquad n \in \boldsymbol{N},$$

gives the existence of $H(\phi^o,\phi)$, such that $-\infty \leq H(\phi^o,\phi) < \infty$ and

$$\lim_{n\to\infty} \frac{1}{n} q_{\phi}(Y_1,\ldots,Y_n) = H(\phi^o,\phi), \tag{4.77}$$

with probability one under ϕ^o. Since $\{W_{s,t}\}$ is ergodic under ϕ^o, then $H(\phi^o, \phi)$ is constant and

$$\lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\Big[\log q_\phi(Y_1, \ldots, Y_n)\Big] = E\Big[H(\phi^o, \phi)\Big] = H(\phi^o, \phi). \tag{4.78}$$

By (4.69), (4.70), (4.77) and (4.78), the conclusions of the theorem follows. ■

Remarks 4.4.4 The proof of Theorem 4.4.3 shows that the value of $H(\phi^o, \phi)$, for $\phi \in \Phi_K^c$, with $\pi_i(\phi) > 0$, for $i = 1, \ldots, K$, depends on the value of

$$\begin{aligned} q_\phi(Y_1, \ldots, Y_n|i) &= f(Y_1, \theta_i(\phi)) \sum_{x_2=1}^{K} \cdots \sum_{x_n=1}^{K} \alpha_{i,x_2}(\phi) f(Y_2, \theta_{x_2}(\phi)) \\ &\quad \times \prod_{t=3}^{n} \alpha_{x_{t-1},x_t}(\phi) f(Y_t, \theta_{x_t}(\phi)), \end{aligned}$$

for $i = 1, \ldots, K$ and $n \in \boldsymbol{N}$, which does not depend on the value of the initial distribution $\pi(\phi)$.

As a direct consequence of Theorem 4.4.2 and Theorem 4.4.3, we have the following corollary which shows the relation between the Kullback-Leibler divergence and the log-likelihood process.

Corollary 4.4.5 *Assume conditions A1, A2, A4, A5, A6 and A7 hold. Then for $\phi \in \Phi_K^c$, with $\pi_i(\phi) > 0$, for $i = 1, \ldots, K$,*

$$\begin{aligned} K(\phi^o, \phi) &= \lim_{n\to\infty} L_n(\phi^o, Y) - \lim_{n\to\infty} L_n(\phi, Y) \\ &= -H(\phi^o) - H(\phi^o, \phi). \end{aligned}$$

Remarks 4.4.6 If $K \geq K^o$, then for $\phi \sim \phi^o$, $H(\phi^o) = -H(\phi^o, \phi)$.

4.5 Simplified Parameter Space for Hidden Markov Models

Let $\phi = (K, A, \pi, \theta) \in \Phi_K^c$, with $\pi_i > 0$, for $i = 1, \ldots, K$. From Theorem 4.4.3, for ϕ, there is a constant $H(\phi^o, \phi)$, such that :

(a). $-\infty \leq H(\phi^o, \phi) < \infty$

(b). $\lim_{n\to\infty} \frac{1}{n} E_{\phi^o}\Big[\log p_\phi(Y_1, \ldots, Y_n)\Big] = H(\phi^o, \phi)$

(c). $\lim_{n\to\infty} \frac{1}{n} \log p_\phi(Y_1, \ldots, Y_n) = H(\phi^o, \phi)$ with probability one under ϕ^o.

The proof of Theorem 4.4.3 shows that the value of $H(\phi^o, \phi)$ depends only on the values of A and θ, and does not depend on the value of π. This means that

$$H(\phi^o, \phi) = H(\phi^o, \widehat{\phi}), \tag{4.79}$$

for all $\widehat{\phi} \in \mathcal{S}(K, A, \theta)$, where

$$\mathcal{S}(K, A, \theta) = \Big\{\widehat{\phi} \in \Phi_K^c \ : \ \widehat{\phi} = (K, A, \widehat{\pi}, \theta,), \quad \widehat{\pi}_i > 0, \quad i = 1, \ldots, K\Big\}$$

By Corollary 4.4.5 and (4.79)

$$K(\phi^o, \phi) = K(\phi^o, \widehat{\phi}),$$

for all $\widehat{\phi} \in \mathcal{S}(K, A, \theta)$, which implies that all parameters in the set $\mathcal{S}(K, A, \theta)$ are *indistinguishable* in term of Kullback-Leibler divergence. This suggests that the set $\mathcal{S}(K, A, \theta)$ can be simply represented by a single parameter

$$\widetilde{\phi} = (K, A, \alpha^K, \theta),$$

where α^K is an arbitrary initial probability distribution with $\alpha_i^K > 0$, for $i = 1, \ldots, K$ and α^K is *independent* of A and θ.

Let $\phi = (K, A, \pi, \theta) \in \Phi_K$, with $\pi_i = 0$, for some i, $1 \leq i \leq K$. As K^o is minimum, if $K < K^o$, then $\phi \not\sim \phi^o$. If $K = K^o$, then by Corollary 2.6.5, $\phi \not\sim \phi^o$. If $K > K^o$ and $\phi \sim \phi^o$, then by Lemma 2.6.4, the number of non-zero π_i, that is N, satisfies $K^o \leq N \leq K$. By Lemma 2.5.5, there is $\widehat{\phi} = (N, \widehat{A}, \widehat{\pi}, \widehat{\theta}) \in \Phi_N^c$, with $\widehat{\pi}_i > 0$, for $i = 1, \ldots, N$ such that $\phi \sim \widehat{\phi} \sim \phi^o$.

Two facts above suggest that we may ignore every parameter in Φ_K^c which has zero elements in its initial probability distribution and simplify every set

$$\mathcal{S}(K, A, \theta)$$

by a single parameter

$$\widetilde{\phi} = (K, A, \alpha^K, \theta).$$

So Φ_K^c can be simplified by

$$\begin{aligned}\widetilde{\Phi}_K^c = \Big\{ \widetilde{\phi} \ : \ & \widetilde{\phi} = (K, A, \alpha^K, \theta), \text{ where } A \text{ and } \theta \text{ satisfy :} \\ & A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K \\ & \theta = (\theta_i)^T, \quad \theta_i \in \Theta^c, \quad i = 1, \ldots, K \Big\}. \end{aligned} \tag{4.80}$$

Since α^K is arbitrary but fixed for all (A, θ), then for convenience, we will write (4.80) as

$$\begin{aligned}\widetilde{\Phi}_K^c = \Big\{ \widetilde{\phi} \ : \ & \widetilde{\phi} = (K, A, \theta), \text{ where } A \text{ and } \theta \text{ satisfy :} \\ & A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K \\ & \theta = (\theta_i)^T, \quad \theta_i \in \Theta^c, \quad i = 1, \ldots, K \Big\}. \end{aligned} \tag{4.81}$$

Notation 4.5.1 For convenience, we will use tilde for every parameter in $\widetilde{\Phi}_K^c$ and without tilde for every parameter in Φ_K^c, for example, $\widetilde{\phi} \in \widetilde{\Phi}_K^c$ and $\phi \in \Phi_K^c$.

To extend the idea of the equivalence relation $\sim$ defined on

$$\Phi^c = \bigcup_{K \in \boldsymbol{N}} \Phi^c_K$$

to the new parameter space

$$\tilde{\Phi}^c = \bigcup_{K \in \boldsymbol{N}} \tilde{\Phi}^c_K,$$

define a new relation $\simeq$ on $\tilde{\Phi}^c$ as follows.

Definition 4.5.2 *Let $\tilde{\phi}_1 = (K_1, A_1, \theta_1)$ and $\tilde{\phi}_2 = (K_2, A_2, \theta_2)$ be two elements of $\tilde{\Phi}^c$. Define*

$$\tilde{\phi}_1 \simeq \tilde{\phi}_2$$

if and only if

(a). $K(\phi^o, \tilde{\phi}_1) = K(\phi^o, \tilde{\phi}_2)$

(b). there are initial probability distributions π_1 and π_2, such that

$$\phi_1 \sim \phi_2,$$

where $\phi_1 = (K_1, A_1, \pi_1, \theta_1)$ and $\phi_2 = (K_2, A_2, \pi_2, \theta_2)$.

It is clear that $\simeq$ is an *equivalence* relation on $\tilde{\Phi}^c$.

From Definition 4.5.2, if $\tilde{\phi}_1 \simeq \tilde{\phi}_2$, where $\tilde{\phi}_1 = (K_1, A_1, \theta_1)$ and $\tilde{\phi}_2 = (K_2, A_2, \theta_2)$, then there are initial probability distributions π_1 and π_2, such that $\phi_1 \sim \phi_2$, where $\phi_1 = (K_1, A_1, \pi_1, \theta_1)$ and $\phi_2 = (K_2, A_2, \pi_2, \theta_2)$, but in general the converse is not true. The following example shows this case.

Example 4.5.3 Let $\phi_1 = (2, A_1, \pi_1, \theta_1) \in \Phi^c_2$, with

$$A_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$

$$\begin{aligned}\pi_1 &= (\tfrac{1}{2}, \tfrac{1}{2}) \\ \theta_1 &= (\alpha_1, \alpha_2)^T,\end{aligned}$$

where $\alpha_1 \neq \alpha_2$ and $\phi_2 = (3, A_2, \pi_2, \theta_2) \in \Phi_3^c$, with

$$A_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$\begin{aligned}\pi_1 &= (\tfrac{1}{2}, \tfrac{1}{2}, 0) \\ \theta_1 &= (\alpha_1, \alpha_2, \alpha_3)^T,\end{aligned}$$

where $\alpha_3 \neq \alpha_1$ and $\alpha_3 \neq \alpha_2$. Let $\tilde{\phi}_1 = (2, A_1, \theta_1)$ and $\tilde{\phi}_2 = (3, A_2, \theta_2)$.

It is clear that $\phi_1 \sim \phi_2$, but since $\alpha_1, \alpha_2, \alpha_3$ are distinct, then $K(\phi^o, \tilde{\phi}_1) \neq K(\phi^o, \tilde{\phi}_2)$. So $\tilde{\phi}_1 \not\simeq \tilde{\phi}_2$.

Then next lemma gives a sufficient condition for the converse to hold.

Lemma 4.5.4 *Let* $\phi_1 = (K_1, A_1, \pi_1, \theta_1) \in \Phi_{K_1}^c$ *and* $\phi_2 = (K_2, A_2, \pi_2, \theta_2) \in \Phi_{K_2}^c$, *with* $\pi_{1,i} > 0$, *for* $i = 1, \ldots, K_1$ *and* $\pi_{2,i} > 0$, *for* $i = 1, \ldots, K_2$. *Let* $\tilde{\phi}_1 = (K_1, A_1, \theta_1)$ *and* $\tilde{\phi}_2 = (K_2, A_2, \theta_2)$. *If* $\phi_1 \sim \phi_2$, *then* $\tilde{\phi}_1 \simeq \tilde{\phi}_2$.

Proof :

Since $\pi_{1,i} > 0$, for $i = 1, \ldots, K_1$, $\phi_1 \sim \phi_2$ and $\pi_{2,i} > 0$, for $i = 1, \ldots, K_2$, then

$$\begin{aligned}H(\phi^o, \tilde{\phi}_1) &= \lim_{n\to\infty} \frac{1}{n} \log p_{\phi_1}(Y_1, \ldots, Y_n) \\ &= \lim_{n\to\infty} \frac{1}{n} \log p_{\phi_2}(Y_1, \ldots, Y_n) \\ &= H(\phi^o, \tilde{\phi}_2),\end{aligned}$$

implying

$$K(\phi^o, \tilde{\phi}_1) = K(\phi^o, \tilde{\phi}_2).$$

So $\tilde{\phi}_1 \simeq \tilde{\phi}_2$. ■

Corollary 4.5.5 *Assume conditions A1 and A2 hold. If* $\phi = (K^o, A, \pi, \theta) \in \Phi_{K^o}$ *and* $\phi \sim \phi^o$, *then* $\tilde{\phi} \simeq \tilde{\phi}^o$, *where* $\tilde{\phi} = (K^o, A, \theta)$ *and* $\tilde{\phi}^o = (K^o, A^o, \theta^o)$.

Proof :

By A1 and A2, $\pi_i^o > 0$, for $i = 1, \ldots, K^o$ and by Corollary 2.6.5, $\pi_i > 0$, for $i = 1, \ldots, K^o$. By Lemma 4.5.4, the conclusion of the corollary follows. ■

As direct consequences of Lemma 2.5.13 and Corollary 2.5.10, we have the following lemmas.

Lemma 4.5.6 *For any* $K \in \boldsymbol{N}$ *and* $\tilde{\phi}_1 \in \tilde{\Phi}_K^c$, *there is* $\tilde{\phi}_2 \in \tilde{\Phi}_{K+1}^c$ *such that* $\tilde{\phi}_1 \simeq \tilde{\phi}_2$.

Lemma 4.5.7 *For* $\tilde{\phi}_1 \in \tilde{\Phi}_K^c$, *there are infinitely many* $\tilde{\phi}_2 \in \tilde{\Phi}_{K+1}^c$ *such that* $\tilde{\phi}_1 \simeq \tilde{\phi}_2$.

Corollary 4.5.8 *Assume that conditions A1, A2 and A3 hold. Let* $\tilde{\phi}^o = (K^o, A^o, \theta^o)$.

(a). If $K < K^o$, *then there is no* $\tilde{\phi} \in \tilde{\Phi}_K^c$ *such that* $\tilde{\phi} \simeq \tilde{\phi}^o$.

(b). If $K = K^o$, *then there are at least finitely many* $\tilde{\phi} \in \tilde{\Phi}_K^c$ *such that* $\tilde{\phi} \simeq \tilde{\phi}^o$.

(c). If $K > K^o$, *then there are infinitely many* $\tilde{\phi} \in \tilde{\Phi}_K^c$ *such that* $\tilde{\phi} \simeq \tilde{\phi}^o$.

Proof :

Since K^o is minimum, then (a) follows. (b) holds since $\sigma(\tilde{\phi}^o) \simeq \tilde{\phi}^o$, for every permutation σ of $\{1, \ldots, K^o\}$ and (c) follows directly from Lemma 4.5.7. ■

By Lemma 4.5.6, we now can define an order $\preceq$ on $\{\tilde{\Phi}^c_K\}$.

Definition 4.5.9 *Define an order* $\preceq$ *on* $\{\tilde{\Phi}^c_K\}$ *by*

$$\tilde{\Phi}^c_K \preceq \tilde{\Phi}^c_L, \qquad K, L \in \boldsymbol{N}$$

if and only if for every $\tilde{\phi}_1 \in \tilde{\Phi}^c_K$, *there is* $\tilde{\phi}_2 \in \tilde{\Phi}^c_L$ *such that* $\tilde{\phi}_1 \simeq \tilde{\phi}_2$.

By Lemma 4.5.6, we have:

Lemma 4.5.10 *For every* $K \in \boldsymbol{N}$,

$$\tilde{\Phi}^c_K \preceq \tilde{\Phi}^c_{K+1}.$$

From Lemma 4.5.10, the new families of parameter for hidden Markov models are *nested* families.

Let $\tilde{\Phi}^c_K$, for some arbitrary but fixed $K \in \boldsymbol{N}$, be the parameter space for hidden Markov models. For $i, j = 1, \ldots, K$, let $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ be the coordinate projections on $\tilde{\Phi}^c_K$, which is defined by

$$\alpha_{ij}(\tilde{\phi}) = \alpha_{ij} \qquad \text{and} \qquad \theta_i(\tilde{\phi}) = \theta_i,$$

for $\tilde{\phi} = (K, A, \theta) \in \tilde{\Phi}^c_K$. Then for $n \in \boldsymbol{N}$ and $y = \{y_t\} \in \mathcal{Y}$, the log-likelihood function $L_n(\cdot, y)$ is defined on $\tilde{\Phi}^c_K$ by

$$\begin{aligned} L_n(\tilde{\phi}, y) &= \frac{1}{n} \log p_{\tilde{\phi}}(y_1, \ldots, y_n) \\ &= \frac{1}{n} \log \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \alpha^K_{x_1} f(y_1, \theta_{x_1}(\tilde{\phi})) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}(\tilde{\phi}) f(y_t, \theta_{x_t}(\tilde{\phi})) \end{aligned}$$

Suppose that a similar condition to A5 hold, that is,

$A5^*$. For each $i, j = 1, \ldots, K$, $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ are continuous on $\tilde{\Phi}_K^c$.

Then A5* together with A4 imply the *continuity* of $L_n(\cdot, y)$ on $\tilde{\Phi}_K^c$, for every $n \in \boldsymbol{N}$ and $y \in \mathcal{Y}$. Furthermore, from the proof of Lemma 4.2.1, it can be seen that for each $y \in \mathcal{Y}$, $\{L_n(\cdot, y) : n \in \boldsymbol{N}\}$ is an *equicontinuous* sequence on $\tilde{\Phi}_K^c$.

4.6 Kullback-Leibler Divergence and Parameters which are Equivalent with the True Parameter

Let $\tilde{\Phi}_K^c$ be the selected parameter space for hidden Markov models, for some $K \in \boldsymbol{N}$. If $K \geq K^o$, then by Corollary 4.5.8,

$$\left\{\tilde{\phi} \in \tilde{\Phi}_K^c : \tilde{\phi} \simeq \tilde{\phi}^o\right\} \neq \emptyset.$$

Recall that $\phi^o = (K^o, A^o, \pi^o, \theta^o)$ and $\tilde{\phi}^o = (K^o, A^o, \theta^o)$.

Let $\tilde{\phi} \in \tilde{\Phi}_K^c$. If $\tilde{\phi} \simeq \tilde{\phi}^o$, then by definition of $\simeq$, it is clear that $K(\phi^o, \tilde{\phi}) = 0$. On the otherhand, if $K(\phi^o, \tilde{\phi}) = 0$, is $\tilde{\phi} \simeq \tilde{\phi}^o$? This whole section is dedicated to the answer to this question.

In this section, we will adapt the work of [34] to our case. The adaptation is possible due to the existence of parameter $\varphi^o \in \Phi_K^c$ which satisfies :

(a). $\pi_i(\varphi^o) > 0, \quad$ for $i = 1, \ldots, K$

(b). $\pi(\varphi^o) = (\pi_i(\varphi^o))$ is a stationary probability distribution of the transition probability matrix $(\alpha_{ij}(\varphi^o))$

(c). $\varphi^o \sim \phi^o$

(d). $E_{\varphi^o}\left[|\log f(Y_1, \theta_i(\varphi^o))|\right] < \infty, \qquad$ for $i = 1, \ldots, K$.

From the proof of Lemma 2.5.9 and under conditions A1, A2 and A6, the existence of such φ^o is guaranteed. By (a), (c) and Lemma 4.5.4, $\tilde{\varphi}^o \simeq \tilde{\phi}^o$, where $\tilde{\varphi}^o = (K, A(\varphi^o), \theta(\varphi^o))$.

Since (a), (c) and (d) hold, by Theorem 4.4.2 and Lemma 4.5.4,

$$\begin{aligned} H(\phi^o) &= H(\varphi^o) \\ H(\phi^o, \tilde{\phi}) &= H(\varphi^o, \tilde{\phi}), \qquad \forall\, \tilde{\phi} \in \tilde{\Phi}_K^c. \end{aligned}$$

Then to prove $K(\phi^o, \tilde{\phi}) = 0$ implies $\tilde{\phi} \simeq \tilde{\phi}^o$ is equivalent showing $K(\varphi^o, \tilde{\phi}) = 0$ implies $\tilde{\phi} \simeq \tilde{\varphi}^o$. Thus throughout this section, the role of ϕ^o will be replaced by φ^o.

First, as in [34], the Fustenberg and Kesten [23] approach will be used to define a new probability space in which the process $\{Y_t\}$ and $\{P_{\varphi^o}(X_t = i \,|Y_{t-1}, Y_{t-2}, \ldots) : i = 1, \ldots, K,\ t \in \boldsymbol{N}\}$ are stationary.

The Fustenberg and Kesten approach requires a careful accounting of the probability spaces and measures involved. We begin with the process $\{Y_t\}$ which is defined on the probability space $(\mathcal{Y}^\infty, \mathcal{B}, P_{\varphi^o})$, where $\mathcal{Y}^\infty$ is the set of all realizations $y = \{y_t\}$ and $\mathcal{B}$ is the Borel σ-algebra of $\mathcal{Y}^\infty$. Let T be the shift operator on $\mathcal{Y}^\infty$ which is defined by

$$T\{y_t\} = \{y_{t+1}\}.$$

Since $\varphi^o \sim \phi^o$, then by condition A2, $\{Y_t\}$ is stationary under φ^o, with respect to φ^o, that is,

$$P_{\varphi^o}(T^{-1}(A)) = P_{\varphi^o}(A), \qquad A \in \mathcal{B},$$

where

$$T^{-1}(A) = \{y \in \mathcal{Y}^\infty : Ty \in A\}.$$

Furthermore, by condition A1, $\{Y_t\}$ is also ergodic under φ^o.

Let Ω be the set of sequences $\omega = \{(y_t, u^{(t)}, v^{(t)})\}$, where $\{y_t\}$ is a realization of $\{Y_t\}$, $u^{(t)}$ and $v^{(t)}$ are K-dimensional vectors satisfying :

$$\begin{aligned} & u_j^{(t)}, v_j^{(t)} \geq 0, \qquad j = 1, \ldots, K, \quad t = 1, 2, \ldots \\ & \|u^{(t)}\| = 0, \quad \text{or} \quad 1, \qquad t = 1, 2, \ldots \\ & \|v^{(t)}\| = 0, \quad \text{or} \quad 1, \qquad t = 1, 2, \ldots \end{aligned}$$

with

$$\|a\| = \sum_{i=1}^{K} |a_i|, \qquad \text{for} \quad a = (a_1, \ldots, a_K).$$

Now define $Y_t, U^{(t)}, V^{(t)}$ as the coordinate functions on Ω, that is,

$$Y_t(\omega) = y_t, \qquad U^{(t)}(\omega) = u^{(t)}, \qquad V^{(t)}(\omega) = v^{(t)},$$

for $\omega = \{(y_t, u^{(t)}, v^{(t)})\} \in \Omega$.

For $\tilde{\phi} \in \Phi_K^c$ with $K \geq K^o$, let Ω^o be the subset of Ω on which

$$u_j^{(1)} = v_j^{(1)} = \pi_j(\varphi^o), \qquad \text{for} \quad j = 1, \ldots, K \tag{4.82}$$

$$u_k^{(t+1)} = \frac{\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o)) \alpha_{jk}(\varphi^o)}{\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o))}, \qquad \text{for} \quad k = 1, \ldots, K, \quad t = 1, 2, \ldots \tag{4.83}$$

$$v_k^{(t+1)} = \frac{\sum_{j=1}^{K} v_j^{(t)} f(y_t, \theta_j(\tilde{\phi})) \alpha_{jk}(\tilde{\phi})}{\sum_{j=1}^{K} v_j^{(t)} f(y_t, \theta_j(\tilde{\phi}))}, \qquad \text{for} \quad k = 1, \ldots, K, \quad t = 1, 2, \ldots . \tag{4.84}$$

If $\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o)) = 0$, define $u_k^{(t+1)} = 0$, for $k = 1, \ldots, K$. Also if $\sum_{j=1}^{K} v_j^{(t)} f(y_t, \theta_j(\tilde{\phi})) = 0$, define $v_k^{(t+1)} = 0$, for $k = 1, \ldots, K$.

Remarks 4.6.1

(a). $u_j^{(1)} = v_j^{(1)} = \pi_j(\varphi^o) > 0, \quad \text{for} \quad j = 1, \ldots, K.$

$$\|u^{(1)}\| = \|v^{(1)}\| = \sum_{j=1}^{K} \pi_j(\varphi^o) = 1.$$

(b). For $t = 1, 2, \ldots$, if $\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o)) = 0$, then

$$u_k^{(t+1)} = 0, \qquad \text{for} \quad k = 1, \ldots, K$$

and

$$\|u^{(t+1)}\| = \sum_{k=1}^{K} u_k^{(t+1)} = 0,$$

if $\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o)) \neq 0$, then

$$\begin{aligned} \|u^{(t+1)}\| &= \sum_{k=1}^{K} u_k^{(t+1)} \\ &= \frac{\sum_{k=1}^{K} \sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o)) \alpha_{jk}(\varphi^o)}{\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o))} \\ &= \frac{\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o))}{\sum_{j=1}^{K} u_j^{(t)} f(y_t, \theta_j(\varphi^o))} \\ &= 1, \end{aligned}$$

as $\sum_{k=1}^{K} \alpha_{jk}(\varphi^o) = 1$, for $j = 1, \ldots, K$.

(c). Similarly,

$$v_k^{t+1} \geq 0, \qquad \text{for} \quad k = 1, \ldots, K, \quad t = 1, 2, \ldots$$

$$\|v^{(t+1)}\| = 0 \quad \text{or} \quad 1, \qquad \text{for} \quad t = 1, 2, \ldots$$

(d). So by (a), (b) and (c), $\Omega^o \subset \Omega$.

Notice that on Ω^o, each $\omega = \{(y_t, u^{(t)}, v^{(t)})\}$ is *uniquely* defined by $y = \{y_t\} \in \mathcal{Y}^\infty$. Hence Ω^o may be taken as the sample space for the $\{Y_t\}$. Define a function $r : \Omega^o \to \mathcal{Y}^\infty$ by

$$r\{(y_t, u^{(t)}, v^{(t)})\} = \{y_t\},$$

then r is a 1-1 *correspondence.*

Let

$$\mathcal{B}_{\Omega^o} = \{r^{-1}(A) \subset \Omega^o : A \in \mathcal{B}\}.$$

$\mathcal{B}_{\Omega^o}$ is a σ-algebra. Also let

$$\mathcal{B}_\Omega = \{A \subset \Omega : A \cap \Omega^o \in \mathcal{B}_{\Omega^o}\},$$

then $\mathcal{B}_\Omega$ is a σ-algebra and $\mathcal{B}_{\Omega^o} \subset \mathcal{B}_\Omega$.

The goal is to define a probability measure on $\mathcal{B}_\Omega$, under which $\{(Y_t, U^{(t)}, V^{(t)})\}$ is a stationary sequence, while $\{Y_t\}$ has the same distribution as it does under P_{φ^o}.

Since $r : \Omega^o \to \mathcal{Y}^\infty$ is a 1-1 correspondence, then we can *carry over* the measure P_{φ^o} to $\mathcal{B}_{\Omega^o}$ and *trivially extend* it to a measure $P'_{\varphi^o,\tilde{\phi}}$ on $\mathcal{B}_\Omega$. Define $P'_{\varphi^o,\tilde{\phi}}$ on $\mathcal{B}_\Omega$ as follows :

$$P'_{\varphi^o,\tilde{\phi}}(A) = P_{\varphi^o}(r(A \cap \Omega^o)), \qquad A \in \mathcal{B}_\Omega.$$

Observe that

$$P'_{\varphi^o,\tilde{\phi}}(A) = P'_{\varphi^o,\tilde{\phi}}(A \cap \Omega^o), \qquad A \in \mathcal{B}_\Omega$$

and

$$\begin{aligned} P'_{\varphi^o,\tilde{\phi}}(\Omega^o) &= P_{\varphi^o}(r(\Omega^o)) \\ &= P_{\varphi^o}(\mathcal{Y}^\infty) \\ &= 1. \end{aligned} \tag{4.85}$$

By (4.82), (4.83), (4.84) and (4.85), on the support of $P'_{\varphi^o,\tilde{\phi}}$,

$$U_j^{(1)} = V_j^{(1)} = \pi_j(\varphi^o), \qquad \text{for} \quad j = 1, \ldots, K \tag{4.86}$$

$$U_k^{(t+1)} = \frac{\sum_{j=1}^{K} U_j^{(t)} f(Y_t, \theta_j(\varphi^o))\alpha_{jk}(\varphi^o)}{\sum_{j=1}^{K} U_j^{(t)} f(Y_t, \theta_j(\varphi^o))}, \quad \text{for} \quad k = 1, \ldots, K, \quad t = 1, 2, \ldots \tag{4.87}$$

$$V_k^{(t+1)} = \frac{\sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi}))\alpha_{jk}(\tilde{\phi})}{\sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi}))}, \quad \text{for} \quad k = 1, \ldots, K, \quad t = 1, 2, \ldots \tag{4.88}$$

Lemma 4.6.2 *The process* $\{Y_t\}$ *has the same distribution under* $P'_{\varphi^o,\tilde{\phi}}$ *as under* P_{φ^o}.

Proof :

Let $A \in \mathcal{R}^n$, where $\mathcal{R}^n$ is a Borel σ-algebra of $\boldsymbol{R}^n$, then

$$\begin{aligned} & P'_{\varphi^o,\tilde{\phi}}\{\omega \in \Omega : (Y_1(\omega), \ldots, Y_n(\omega)) \in A\} \\ &= P'_{\varphi^o,\tilde{\phi}}\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega^o : (y_1, \ldots, y_n) \in A\} \\ &= P_{\varphi^o}\{\{(y_t)\} \in \mathcal{Y}^\infty : (y_1, \ldots, y_n) \in A\} \\ &= P_{\varphi^o}\{y \in \mathcal{Y}^\infty : (Y_1(y), \ldots, Y_n(y)) \in A\}. \end{aligned}$$

■

Let T_Ω be the *shift operator* defined on Ω by

$$T_\Omega\{(y_t, u^{(t)}, v^{(t)})\} = \{(y_{t+1}, u^{(t+1)}, v^{t+1})\}.$$

Lemma 4.6.3 *Assume condition A2 holds, then* $\{Y_t\}$ *is a stationary process with respect to* T_Ω *under* $P'_{\varphi^o,\tilde{\phi}}$.

Proof :

Let $A \in \mathcal{R}^n$ and

$$\begin{aligned} B &= \Big\{\omega \in \Omega : (Y_1(\omega), \ldots, Y_n(\omega)) \in A\Big\} \\ &= \Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega : (y_1, \ldots, y_n) \in A\Big\}, \end{aligned}$$

then

$$\begin{aligned} T_\Omega^{-1}(B) &= \Big\{\omega \in \Omega : T_\Omega(\omega) \in B\Big\} \\ &= \Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega : \{(y_{t+1}, u^{(t+1)}, v^{(t+1)})\} \in B\Big\} \\ &= \Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega : (y_2, \ldots, y_{n+1}) \in A\Big\}. \end{aligned}$$

$$\begin{aligned} P'_{\varphi^o,\tilde{\phi}}\Big(T_\Omega^{-1}(B)\Big) &= P_{\varphi^o,\tilde{\phi}}\Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega^o : (y_2, \ldots, y_{n+1}) \in A\Big\} \\ &= P_{\varphi^o}\Big\{\{(y_t)\} \in \mathcal{Y}^\infty : (y_2, \ldots, y_{n+1}) \in A\Big\} \\ &= P_{\varphi^o}\Big\{\{(y_t)\} \in \mathcal{Y}^\infty : (y_1, \ldots, y_n) \in A\Big\} \qquad (4.89) \\ &= P'_{\varphi^o,\tilde{\phi}}\Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega^o : (y_1, \ldots, y_n) \in A\Big\} \\ &= P'_{\varphi^o,\tilde{\phi}}(B \cap \Omega^o) \\ &= P'_{\varphi^o,\tilde{\phi}}(B), \end{aligned}$$

where (4.89) follows from the stationarity of $\{Y_t\}$ with respect to T under P_{φ^o}.

■

Notice that

$$u_k^{(2)} = \frac{\sum_{j=1}^K u_j^{(1)} f(y_1, \theta_j(\varphi^o))\alpha_{jk}(\varphi^o)}{\sum_{j=1}^K u_j^{(1)} f(y_1, \theta_j(\varphi^o))}, \qquad \text{for} \quad k = 1, \ldots, K$$

$$v_k^{(2)} = \frac{\sum_{j=1}^K v_j^{(1)} f(y_1, \theta_j(\tilde{\phi}))\alpha_{jk}(\tilde{\phi})}{\sum_{j=1}^K v_j^{(1)} f(y_1, \theta_j(\tilde{\phi}))}, \qquad \text{for} \quad k = 1, \ldots, K$$

need not equal $\pi_k(\varphi^o)$, when $u_k^{(1)} = v_k^{(1)} = \pi_k(\varphi^o)$, then in general the $\{(U^{(t)}, V^{(t)})\}$ process *is not stationary* under $P'_{\varphi^o,\tilde{\phi}}$. However, based on $P'_{\varphi^o,\tilde{\phi}}$, we can construct a probability measure, say $\tilde{P}_{\varphi^o,\tilde{\phi}}$ on $\mathcal{B}_\Omega$ such that :

(a). $\{(Y_t, U^{(t)}, V^{(t)})\}$ is a stationary process under $\tilde{P}_{\varphi^o,\tilde{\phi}}$

(b). The process $\{Y_t\}$ has the same distribution under $\tilde{P}_{\varphi^o,\tilde{\phi}}$ as under P_{φ^o},

as follows.

For each $k = 1, 2, \ldots$, define a measure $P'_{\varphi^o,\tilde{\phi}} T_\Omega^{-k+1}$ on $\mathcal{B}_\Omega$ by

$$P'_{\varphi^o,\tilde{\phi}} T_\Omega^{-k+1}(A) = P'_{\varphi^o,\tilde{\phi}}\{\omega \in \Omega : T_\Omega^{k-1}(\omega) \in A\},$$

for $A \in \mathcal{B}_\Omega$.

Lemma 4.6.4 *For each $n = 1, 2, \ldots$, let*

$$\tilde{P}^{(n)}_{\varphi^o,\tilde{\phi}} = \frac{1}{n} \sum_{k=1}^{n} P'_{\varphi^o,\tilde{\phi}} T_\Omega^{-k+1},$$

then there exists a subsequence $\{n_i\}$ and a probability measure $\tilde{P}_{\varphi^o,\tilde{\phi}}$ on $\mathcal{B}_\Omega$ such that :

(a). $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$ converges weakly to $\tilde{P}_{\varphi^o,\tilde{\phi}}$, in the sense that the finite dimensional joint distribution functions of the variables $Y_t, U^{(t)}, V^{(t)}$ with respect to $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$ converge to the corresponding joint distribution functions of the $Y_t, U^{(t)}, V^{(t)}$ with respect to $\tilde{P}_{\varphi^o,\tilde{\phi}}$ at each continuity point of the latter

(b). $\{(Y_t, U^{(t)}, V^{(t)})\}$ is a stationary process under $\tilde{P}_{\varphi^o,\tilde{\phi}}$

(c). The process $\{Y_t\}$ has the same distribution under $\tilde{P}_{\varphi^o,\tilde{\phi}}$ as under P_{φ^o}.

Proof :

The idea of the proof comes from Lemma 1 of [23].

For each $m, n = 1, 2, \ldots$, let $F_m^{(n)}$ be the joint distribution function of $Y_1, \ldots, Y_m; U^{(1)}, \ldots, U^{(m)}; V^{(1)}, \ldots, V^{(m)}$ with respect to $\tilde{P}^{(n)}_{\varphi^o, \tilde{\phi}}$. To show the existence of the subsequence $\{n_i\}$ and the probability measure $\tilde{P}_{\varphi^o, \tilde{\phi}}$ and to prove (a) it will suffice to show that for each m, there exists a subsequence $\{n_k\}$ and a distribution function F_m such that

$$\lim_{k \to \infty} F_m^{(n_k)}(x) = F_m(x),$$

for each continuity point x of F_m. Using diagonal procedure the required subsequence $\{n_i\}$ is obtained and we have

$$\lim_{i \to \infty} F_m^{(n_i)}(x) = F_m(x), \qquad \forall\, m = 1, 2, \ldots$$

for each continuity point x of F_m. Thus, the probability measure $\tilde{P}_{\varphi^o, \tilde{\phi}}$ now can be defined on $\mathcal{B}_\Omega$ by assigning F_m as the joint distribution function of $Y_1, \ldots, Y_m; U^{(1)}, \ldots, U^{(m)}; V^{(1)}, \ldots, V^{(m)}$ under $\tilde{P}_{\varphi^o, \tilde{\phi}}$.

By Helly selection theorem, for sequence of distribution functions $\{F_m^{(n)}\}$, there exists a subsequence $\{n_k\}$ and a function F_m, such that :

- $0 \leq F_m \leq 1$
- F_m is non-decreasing in each variable
- $\triangle_{a_1, b_1} \cdots \triangle_{a_{(2K+1)m}, b_{(2K+1)m}} F_m \geq 0$, for $(2K+1)m$-bounded rectangle $(a, b]$, where :

$$\begin{aligned} (a, b] &= (a_1, b_1] \times \cdots \times (a_{(2K+1)m}, b_{(2K+1)m}] \\ \triangle_{a_i, b_i} &= F_m(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_{(2K+1)m}) \\ &\quad - F_m(x_1, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_{(2K+1)m}) \end{aligned}$$

- F_m is continuous from above
- For each continuity point x of F_m,

$$\lim_{k \to \infty} F_m^{(n_k)}(x) = F_m(x). \tag{4.90}$$

We will prove that F_m is a distribution function. We need to show that

$$F_m(\infty) = 1, \qquad \infty = (\infty, \ldots, \infty) \tag{4.91}$$

and

$$F_m(x) = 0, \qquad \text{if at least one coordinate of } x \text{ is } -\infty. \tag{4.92}$$

We will prove that (4.91) holds and the proof for (4.92) can be verified similarly.

Let $a = (a_1, \ldots, a_m) \in \boldsymbol{R}^m$ and $b = (b^{(1)}, \ldots, b^{(m)}), c = (c^{(1)}, \ldots, c^{(m)}) \in \boldsymbol{R}^{Km}$. Let

$$\begin{aligned} A_{a,b,c} &= \Big\{\omega \in \Omega \,:\, Y_i(\omega) \le a_i,\ U_j^{(i)}(\omega) \le b_j^{(i)},\ V_j^{(i)}(\omega) \le c_j^{(i)}, \\ &\qquad\qquad i = 1, \ldots, m,\ j = 1, \ldots, K\Big\} \\ &= \Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega \,:\, y_i \le a_i,\ u_j^{(i)} \le b_j^{(i)},\ v_j^{(i)} \le c_j^{(i)}, \\ &\qquad\qquad i = 1, \ldots, m,\ j = 1, \ldots, K\Big\}, \end{aligned}$$

$$\begin{aligned} A_a &= \Big\{\omega \in \Omega : Y_1(\omega) \le a_1, \ldots, Y_m(\omega) \le a_m\Big\} \\ &= \Big\{\{(y_t, u^{(t)}, v^{(t)})\} \in \Omega : y_1 \le a_1, \ldots, y_m \le a_m\Big\} \end{aligned}$$

and

$$\begin{aligned} B_a &= \Big\{y \in \mathcal{Y}^\infty : Y_1(y) \le a_1, \ldots, Y_m(y) \le a_m\Big\} \\ &= \Big\{\{(y_t)\} \in \mathcal{Y}^\infty : y_1 \le a_1, \ldots, y_m \le a_m\Big\}. \end{aligned}$$

Notice that

$$r(A_a \cap \Omega^o) = B_a.$$

Since for $t = 1, 2, \ldots$

$$\begin{aligned} &u_j^{(t)}, v_j^{(t)} \ge 0, \qquad j = 1, \ldots, K \\ &\|u^{(t)}\| = 0, \quad \text{or} \quad 1 \\ &\|v^{(t)}\| = 0, \quad \text{or} \quad 1, \end{aligned}$$

then $0 \le u_j^{(t)}, v_j^{(t)} \le 1$ and hence

$$A_{a,b,c} = A_a \tag{4.93}$$

for b and c with $b_j^{(i)}, c_j^{(i)} \geq 1$, $i = 1, \ldots, m$, $j = 1, \ldots, K$. Let G_m be the joint distribution functions of $Y_1, \ldots, Y_m$ under P_{φ^o}, then

$$G_m(a) = P_{\varphi^o}(B_a) \tag{4.94}$$

$$F_m^{(n)}(a, b, c) = \tilde{P}^{(n)}_{\varphi^o, \tilde{\phi}}(A_{a,b,c}) \tag{4.95}$$

By (4.90), (4.95), (4.93), stationarity of $\{Y_t\}$ under $P'_{\varphi^o, \tilde{\phi}}$ and (4.94),

$$\begin{aligned}
F_m(\infty, \infty, \infty) &\geq F_m(\infty, 1, 1) \\
&= \lim_{a \to \infty} F_m(a, 1, 1) \\
&= \lim_{a \to \infty} \lim_{k \to \infty} F_m^{(n_k)}(a, 1, 1) \\
&= \lim_{a \to \infty} \lim_{k \to \infty} \tilde{P}^{(n_k)}_{\varphi^o, \tilde{\phi}}(A_{a,1,1}) \\
&= \lim_{a \to \infty} \lim_{k \to \infty} \tilde{P}^{(n_k)}_{\varphi^o, \tilde{\phi}}(A_a) \\
&= \lim_{a \to \infty} \lim_{k \to \infty} \frac{1}{n_k} \sum_{t=1}^{n_k} P'_{\varphi^o, \tilde{\phi}}\left(T_\Omega^{-t+1} A_a\right) \\
&= \lim_{a \to \infty} \lim_{k \to \infty} \frac{1}{n_k} \sum_{t=1}^{n_k} P'_{\varphi^o, \tilde{\phi}}(A_a) \\
&= \lim_{a \to \infty} P'_{\varphi^o, \tilde{\phi}}(A_a) \\
&= \lim_{a \to \infty} P_{\varphi^o}\left(r(A_a \cap \Omega^o)\right) \\
&= \lim_{a \to \infty} P_{\varphi^o}(B_a) \\
&= \lim_{a \to \infty} G_m(a) \\
&= G_m(\infty) \\
&= 1.
\end{aligned} \tag{4.96}$$

Since $0 \leq F_m \leq 1$ and by (4.96), then $F_m(\infty, \infty, \infty) = 1$. So (a) is proved.

To prove (b), let $A \in \mathcal{R}^{(2K+1)m}$ be a continuity set of the joint distribution of $Y_1, \ldots, Y_m; U^{(1)}, \ldots, U^{(m)}; V^{(1)}, \ldots, V^{(m)}$ under $\tilde{P}_{\varphi^o, \tilde{\phi}}$. Then by Theorem 29.1 of [11], page 390,

$$\tilde{P}_{\varphi^o, \tilde{\phi}}\left\{(Y_1, \ldots, Y_m; U^{(1)}, \ldots, U^{(m)}; V^{(1)}, \ldots, V^{(m)}) \in A\right\}$$

$$
\begin{aligned}
&= \lim_{i\to\infty} \tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m;U^{(1)},\ldots,U^{(m)};V^{(1)},\ldots,V^{(m)})\in A\Big\}\\
&= \lim_{i\to\infty} \frac{1}{n_i}\sum_{k=1}^{n_i} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_k,\ldots,Y_{m+k};U^{(k)},\ldots,U^{(m+k)};V^{(k)},\ldots,V^{(m+k)})\in A\Big\}\\
&= \lim_{i\to\infty} \frac{1}{n_i}\sum_{k=2}^{n_i+1} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_k,\ldots,Y_{m+k};U^{(k)},\ldots,U^{(m+k)};V^{(k)},\ldots,V^{(m+k)})\in A\Big\}\\
&\quad + \lim_{i\to\infty} \frac{1}{n_i} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m;U^{(1)},\ldots,U^{(m)};V^{(1)},\ldots,V^{(m)})\in A\Big\}\\
&\quad - \lim_{i\to\infty} \frac{1}{n_i} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_{n_i+1},\ldots,Y_{n_i+m+1};U^{(n_i+1)},\ldots,U^{(n_i+m+1)};\\
&\qquad\qquad V^{(n_i+1)},\ldots,V^{(n_i+m+1)})\in A\Big\}\\
&= \lim_{i\to\infty} \frac{1}{n_i}\sum_{k=2}^{n_i+1} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_k,\ldots,Y_{m+k};U^{(k)},\ldots,U^{(m+k)};V^{(k)},\ldots,V^{(m+k)})\in A\Big\}\\
&\quad +0-0\\
&= \lim_{i\to\infty} \tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}\Big\{(Y_2,\ldots,Y_{m+1};U^{(2)},\ldots,U^{(m+1)};V^{(2)},\ldots,V^{(m+1)})\in A\Big\}\\
&= \tilde{P}_{\varphi^o,\tilde{\phi}}\Big\{(Y_2,\ldots,Y_{m+1};U^{(2)},\ldots,U^{(m+1)};V^{(2)},\ldots,V^{(m+1)})\in A\Big\}.
\end{aligned}
$$

Hence $\{(Y_t,U^{(t)},V^{(t)})\}$ is stationary under $\tilde{P}_{\varphi^o,\tilde{\phi}}$.

To prove (c), let $B\in\mathcal{R}^m$ be a continuity set of the joint distribution of $Y_1,\ldots,Y_m$ under $\tilde{P}_{\varphi^o,\tilde{\phi}}$, then by Theorem 29.1 of [11], page 390 and by stationarity of $\{Y_t\}$ under $P'_{\varphi^o,\tilde{\phi}}$,

$$
\begin{aligned}
\tilde{P}_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m)\in B\Big\} &= \lim_{i\to\infty} \tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m)\in B\Big\}\\
&= \lim_{i\to\infty} \frac{1}{n_i}\sum_{k=1}^{n_i} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_k,\ldots,Y_{m+k})\in B\Big\}\\
&= \lim_{i\to\infty} \frac{1}{n_i}\sum_{k=1}^{n_i} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m)\in B\Big\}\\
&= \lim_{i\to\infty} P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m)\in B\Big\}\\
&= P'_{\varphi^o,\tilde{\phi}}\Big\{(Y_1,\ldots,Y_m)\in B\Big\}\\
&= P_{\varphi^o}\Big\{(Y_1,\ldots,Y_m)\in B\Big\} \qquad (4.97)
\end{aligned}
$$

The equation (4.97) follows since $\{Y_t\}$ has the same distribution under $P'_{\varphi^o,\tilde{\phi}}$ as under P_{φ^o}. ■

Remarks 4.6.5 Lemma 4.6.4 is similar to Lemma 4 of [34].

The next goal is to interpret the process $\{U^{(t)}\}$ under $P'_{\varphi^o,\widetilde{\phi}}$.

Recall that on the support of $P'_{\varphi^o,\widetilde{\phi}}$,

$$U_j^{(1)} = V_j^{(1)} = \pi_j(\varphi^o), \qquad \text{for} \quad j = 1, \ldots, K \tag{4.98}$$

and

$$U_k^{(t+1)} = \frac{\sum_{j=1}^{K} U_j^{(t)} f(Y_t, \theta_j(\varphi^o))\alpha_{jk}(\varphi^o)}{\sum_{j=1}^{K} U_j^{(t)} f(Y_t, \theta_j(\varphi^o))} \tag{4.99}$$

$$V_k^{(t+1)} = \frac{\sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\widetilde{\phi}))\alpha_{jk}(\widetilde{\phi})}{\sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\widetilde{\phi}))} \tag{4.100}$$

for $t = 1, 2, \ldots$ and $k = 1, \ldots, K$.

Lemma 4.6.6 *On the support of* $P'_{\varphi^o,\widetilde{\phi}}$,

$$U_j^{(t)} = P_{\varphi^o}(X_t = j | Y_{t-1}, \ldots, Y_1)$$

for $j = 1, \ldots, K$ *and* $t = 1, 2, \ldots$.

Proof :

For the proof we use mathematical induction. From (4.98), for $j = 1, \ldots, K$,

$$U_j^{(1)} = \pi_j(\varphi^o) = P_{\varphi^o}(X_1 = j).$$

Assume that for some t,

$$U_k^{(t)} = P_{\varphi^o}(X_t = k | Y_{t-1}, \ldots, Y_1), \qquad k = 1, \ldots, K. \tag{4.101}$$

We will prove that

$$U_k^{(t+1)} = P_{\varphi^o}(X_{t+1} = k|Y_t, \ldots, Y_1), \qquad k = 1, \ldots, K.$$

By (4.99) and (4.101),

$$\begin{aligned}
U_k^{(t+1)} &= \frac{\sum_{j=1}^K U_j^{(t)} f(Y_t, \theta_j(\varphi^o))\alpha_{jk}(\varphi^o)}{\sum_{j=1}^K U_j^{(t)} f(Y_t, \theta_j(\varphi^o))} \\
&= \frac{\sum_{j=1}^K P_{\varphi^o}(X_t = j|Y_{t-1}, \ldots, Y_1)p_{\varphi^o}(Y_t|X_t = j)P_{\varphi^o}(X_{t+1} = k|X_t = j)}{\sum_{j=1}^K P_{\varphi^o}(X_t = j|Y_{t-1}, \ldots, Y_1)p_{\varphi^o}(Y_t|X_t = j)} \\
&= \frac{\sum_{j=1}^K p_{\varphi^o}(X_t = j, Y_t, X_{t+1} = k|Y_{t-1}, \ldots, Y_1)}{\sum_{j=1}^K p_{\varphi^o}(X_t = j, Y_t|Y_{t-1}, \ldots, Y_1)} \\
&= \frac{p_{\varphi^o}(Y_t, X_{t+1} = k|Y_{t-1}, \ldots, Y_1)}{p_{\varphi^o}(Y_t|Y_{t-1}, \ldots, Y_1)} \\
&= P_{\varphi^o}(X_{t+1} = k|Y_t, \ldots, Y_1), \qquad (4.102)
\end{aligned}$$

where (4.102) follows from Lemma 2.3.2. ■

So under $P'_{\varphi^o,\tilde{\phi}}$,

$$U_j^{(t)} = P_{\varphi^o}(X_t = j|Y_{t-1}, \ldots, Y_1)$$

for $j = 1, \ldots, K$ and $t = 1, 2, \ldots$. The operation of shifting the time scale and taking limit to obtain $\tilde{P}_{\varphi^o,\tilde{\phi}}$ has the effect of converting $U_j^{(1)}$ into a conditional probability depending on infinitely many past value of $\{Y_t\}$. Precisely, $U_j^{(1)}$ represents

$$P_{\varphi^o}(X_1 = j|Y_0, Y_{-1}, \ldots). \qquad (4.103)$$

Therefore, the entropy

$$\begin{aligned}
H(\phi^o) &= H(\varphi^o) \\
&= E_{\varphi^o}\left[-\log \sum_{j=1}^K P_{\varphi^o}(X_1 = j|Y_0, Y_{-1}, \ldots) f(Y_1, \theta_j(\varphi^o))\right]
\end{aligned}$$

is seen to be equal to

$$\tilde{E}_{\varphi^o,\tilde{\phi}}\left[-\log \sum_{j=1}^K U_j^{(1)} f(Y_1, \theta_j(\varphi^o))\right],$$

which is true and proved in Lemma 4.6.8.

Before proving Lemma 4.6.8, we will need the following lemma.

Lemma 4.6.7 *For* $n = 1, 2, \ldots,$

$$(a).\ \sum_{t=1}^{n} \log\left(\sum_{j=1}^{K} U_j^{(t)} f(Y_t, \theta_j(\varphi^o))\right) = \log\left(\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1, \ldots, Y_n | X_1 = j)\right)$$

$$(b).\ \sum_{t=1}^{n} \log\left(\sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi}))\right) = \log\left(\sum_{j=1}^{K} V_j^{(1)} p_{\tilde{\phi}}(Y_1, \ldots, Y_n | X_1 = j)\right).$$

Proof :

Here we will only prove (a), the proof for (b) is similar. First, using mathematical induction, we prove that for $t = 1, 2, \ldots.$

$$\sum_{j=1}^{K} U_j^{(t)} f(Y_t, \theta_j(\varphi^o)) = \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1, \ldots, Y_t | X_1 = j)}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1, \ldots, Y_{t-1} | X_1 = j)}. \tag{4.104}$$

For $t = 2$, by definition,

$$\begin{aligned}
\sum_{k=1}^{K} U_k^{(2)} f(Y_2, \theta_k(\varphi^o)) &= \frac{\sum_{k=1}^{K}\sum_{j=1}^{K} U_j^{(1)} f(Y_1, \theta_j(\varphi^o)) \alpha_{jk}(\varphi^o) f(Y_2, \theta_k(\varphi^o))}{\sum_{j=1}^{K} U_j^{(1)} f(Y_1, \theta_j(\varphi^o))} \\
&= \frac{\sum_{j=1}^{K} U_j^{(1)} f(Y_1, \theta_j(\varphi^o)) \sum_{k=1}^{K} \alpha_{jk}(\varphi^o) f(Y_2, \theta_k(\varphi^o))}{\sum_{j=1}^{K} U_j^{(1)} f(Y_1, \theta_j(\varphi^o))} \\
&= \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1, Y_2 | X_1 = j)}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1 | X_1 = j)}.
\end{aligned}$$

Assume that for some t,

$$\sum_{j=1}^{K} U_j^{(t-1)} f(Y_{t-1}, \theta_j(\varphi^o)) = \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1, \ldots, Y_{t-1} | X_1 = j)}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1, \ldots, Y_{t-2} | X_1 = j)}$$

then

$$\begin{aligned}
&\sum_{k=1}^{K} U_k^{(t)} f(Y_t, \theta_k(\varphi^o)) \\
&\quad = \frac{\sum_{k=1}^{K}\sum_{j=1}^{K} U_j^{(t-1)} f(Y_{t-1}, \theta_j(\varphi^o)) \alpha_{jk}(\varphi^o) f(Y_t, \theta_k(\varphi^o))}{\sum_{j=1}^{K} U_j^{(t-1)} f(Y_{t-1}, \theta_j(\varphi^o))}
\end{aligned}$$

$$= \frac{\sum_{k=1}^{K}\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{t-1}|X_1=j)\alpha_{jk}(\varphi^o) f(Y_t,\theta_k(\varphi^o))}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{t-1}|X_1=j)}$$

$$= \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{t-1}|X_1=j)\sum_{k=1}^{K}\alpha_{jk}(\varphi^o) f(Y_t,\theta_k(\varphi^o))}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{t-1}|X_1=j)}$$

$$= \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_t|X_1=j)}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{t-1}|X_1=j)}.$$

Thus (4.104) is proved.

By (4.104),

$$\sum_{t=1}^{n}\log\sum_{j=1}^{K} U_j^{(t)} f(Y_t,\theta_j(\varphi^o))$$

$$= \log\left\{\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o)) \times \sum_{j=1}^{K} U_j^{(2)} f(Y_2,\theta_j(\varphi^o)) \times\cdots\times \sum_{j=1}^{K} U_j^{(n)} f(Y_n,\theta_j(\varphi^o))\right\}$$

$$= \log\left\{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1|X_1=j) \times \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,Y_2|X_1=j)}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1|X_1=j)} \times\cdots\times \frac{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_n|X_1=j)}{\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{n-1}|X_1=j)}\right\}$$

$$= \log\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_n|X_1=j),$$

thus (a) follows. ■

Lemma 4.6.8 *Assume conditions A1, A2, A4, A5* and A6 hold. Then*

$$H(\varphi^o) = \tilde{E}_{\varphi^o,\tilde{\phi}}\left[-\log\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right].$$

Proof :

Since

$$\min_{1\le j\le K}\log f(Y_1,\theta_j(\varphi^o)) \le \log\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o)) \le \max_{1\le j\le K}\log f(Y_1,\theta_j(\varphi^o))$$

and

$$\tilde{E}^{(n_i)}_{\varphi^o,\tilde{\phi}}\Big[|\log f(Y_1,\theta_j(\varphi^o))|\Big] = E_{\varphi^o}\Big[|\log f(Y_1,\theta_j(\varphi^o))|\Big] < \infty,$$

for every $j = 1,\ldots,K$ and $i = 1,2,\ldots$, then the sequence of laws of $\log \sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))$ with respect to $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$ is uniformly integrable. Also, the joint distribution of $(Y_1, U^{(1)})$ under $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$ converges weakly to the corresponding distribution under $\tilde{P}_{\varphi^o,\tilde{\phi}}$, then by Theorem 25.12 of [11], page 348,

$$\tilde{E}^{(n_i)}_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right] \longrightarrow \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right]$$

as $i \to \infty$, implying

$$\begin{aligned}
&\tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right] \\
&= \lim_{i\to\infty} \tilde{E}^{(n_i)}_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right] \\
&= \lim_{i\to\infty} \frac{1}{n_i}\sum_{t=1}^{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} U_j^{(t)} f(Y_t,\theta_j(\varphi^o))\right] \\
&= \lim_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\sum_{t=1}^{n_i}\log\sum_{j=1}^{K} U_j^{(t)} f(Y_t,\theta_j(\varphi^o))\right] \\
&= \lim_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} U_j^{(1)} p_{\varphi^o}(Y_1,\ldots,Y_{n_i}|X_1=j)\right] \qquad (4.105) \\
&= \lim_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K} \pi_j(\varphi^o) p_{\varphi^o}(Y_1,\ldots,Y_{n_i}|X_1=j)\right] \qquad (4.106) \\
&= \lim_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}[\log p_{\varphi^o}(Y_1,\ldots,Y_{n_i})] \\
&= \lim_{i\to\infty} \frac{1}{n_i} E_{\varphi^o}[\log p_{\varphi^o}(Y_1,\ldots,Y_{n_i})] \\
&= -H(\varphi^o),
\end{aligned}$$

where (4.105) follows from Lemma 4.6.7 and (4.106) holds since $U_j^{(1)} = \pi_j(\varphi^o)$, under $P'_{\varphi^o,\tilde{\phi}}$. ■

The result of Lemma 4.6.8 can be extended for $H(\varphi^o,\tilde{\phi})$ shown by the following lemma.

Lemma 4.6.9 *Assume conditions A1, A2, A4, A5* and A7 hold. Then for every* $\tilde{\phi} \in \tilde{\Phi}_K^c$, *with* $K \geq K^o$,

$$H(\varphi^o, \tilde{\phi}) = \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\right].$$

Proof :

The proof follows the proof of Lemma 5 of [34].

By the ergodic theorem, there exists a random variable Z such that

$$\tilde{P}_{\varphi^o,\tilde{\phi}}\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \log \sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi})) = Z\right\} = 1$$

and

$$\tilde{E}_{\varphi^o,\tilde{\phi}}[Z] = \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\right].$$

First, we will show that $\tilde{E}_{\varphi^o,\tilde{\phi}}[Z] \leq H(\varphi^o, \tilde{\phi})$. By part (b) of Lemma 4.6.7,

$$\begin{aligned}
\tilde{E}_{\varphi^o,\tilde{\phi}}[Z] &= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \log \sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi}))\right] \\
&= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\lim_{n\to\infty} \frac{1}{n} \log \sum_{j=1}^{K} V_j^{(1)} p_{\tilde{\phi}}(Y_1, \ldots, Y_n | X_1 = j)\right] \\
&\leq \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\lim_{n\to\infty} \frac{1}{n} \log \left(\max_{1\leq j\leq K} p_{\tilde{\phi}}(Y_1, \ldots, Y_n | X_1 = j)\right)\right] \\
&= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\lim_{n\to\infty} \frac{1}{n} \log q_{\tilde{\phi}}(Y_1, \ldots, Y_n)\right] \\
&= E_{\varphi^o}\left[\lim_{n\to\infty} \frac{1}{n} \log q_{\tilde{\phi}}(Y_1, \ldots, Y_n)\right] \\
&= E_{\varphi^o}[H(\varphi^o, \tilde{\phi})] \qquad (4.107) \\
&= H(\varphi^o, \tilde{\phi}),
\end{aligned}$$

where (4.107) follows from the proof of Theorem 4.4.3.

Next, we will show that $\tilde{E}_{\varphi^o,\tilde{\phi}}[Z] \geq H(\varphi^o, \tilde{\phi})$. Without loss of generality,

assume that $H(\varphi^o, \tilde{\phi}) > -\infty$. Let

$$A = \left\{ \omega \in \Omega \ : \ \log \sum_{j=1}^{K} V_j^{(1)}(\omega) f(Y_1(\omega), \theta_j(\tilde{\phi})) \leq 0 \right\}.$$

Since the joint distribution of $(Y_1, V^{(1)})$ under $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$ converges weakly to the corresponding distribution under $\tilde{P}_{\varphi^o,\tilde{\phi}}$, then by Theorem 25.11 of [11], page 347,

$$\begin{aligned} \int_A \left| \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi})) \right| d\tilde{P}_{\varphi^o,\tilde{\phi}} \\ \leq \liminf_{i\to\infty} \int_A \left| \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi})) \right| d\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}. \end{aligned}$$

implying

$$\begin{aligned} \int_A \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}_{\varphi^o,\tilde{\phi}} \\ \geq \limsup_{i\to\infty} \int_A \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}. \end{aligned} \tag{4.108}$$

Also, since

$$\begin{aligned} \left\{ \min_{1\leq j\leq K} \log f(Y_1, \theta_j(\tilde{\phi})) \right\}^+ &\leq \left\{ \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi})) \right\}^+ \\ &\leq \left\{ \max_{1\leq j\leq K} \log f(Y_1, \theta_j(\tilde{\phi})) \right\}^+ \end{aligned}$$

and by condition A6,

$$\tilde{E}^{(n_i)}_{\varphi^o,\tilde{\phi}} \left[\left\{ \log f(Y_1, \theta_j(\tilde{\phi})) \right\}^+ \right] = E_{\varphi^o} \left[\left\{ \log f(Y_1, \theta_j(\tilde{\phi})) \right\}^+ \right] < \infty,$$

for any n_i, then the sequence of laws of $\left\{ \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi})) \right\}^+$ with respect to $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$ is uniformly integrable, then by Theorem 25.12 of [11], page 348,

$$\begin{aligned} \lim_{i\to\infty} \int_{\Omega\setminus A} \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}} \\ = \int_{\Omega\setminus A} \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}_{\varphi^o,\tilde{\phi}}. \end{aligned} \tag{4.109}$$

Hence by (4.108), (4.109), Lemma 4.6.7 and Theorem 4.4.3

$$
\begin{aligned}
\tilde{E}_{\varphi^o,\tilde{\phi}}[Z] &= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\right] \\
&= \int_A \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}_{\varphi^o,\tilde{\phi}} \\
&\quad + \int_{\Omega\setminus A} \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}_{\varphi^o,\tilde{\phi}} \\
&\geq \limsup_{i\to\infty} \int_A \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}} \\
&\quad + \limsup_{i\to\infty} \int_{\Omega\setminus A} \log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\, d\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}} \\
&= \limsup_{i\to\infty} \tilde{E}^{(n_i)}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} V_j^{(1)} f(Y_1, \theta_j(\tilde{\phi}))\right] \\
&= \limsup_{i\to\infty} \frac{1}{n_i} \sum_{t=1}^{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi}))\right] \\
&= \limsup_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\sum_{t=1}^{n_i} \log \sum_{j=1}^{K} V_j^{(t)} f(Y_t, \theta_j(\tilde{\phi}))\right] \\
&= \limsup_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} V_j^{(1)} p_{\tilde{\phi}}(Y_1, \ldots, Y_{n_i} | X_1 = j)\right] \\
&= \limsup_{i\to\infty} \frac{1}{n_i} E'_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} \pi_j(\varphi^o) p_{\phi}(Y_1, \ldots, Y_{n_i} | X_1 = j)\right] \qquad (4.110) \\
&= \limsup_{i\to\infty} \frac{1}{n_i} E_{\varphi^o}\left[\log p_{\tilde{\phi}}(Y_1, \ldots, Y_{n_i})\right] \qquad (4.111) \\
&= H(\varphi^o, \tilde{\phi}).
\end{aligned}
$$

Equation (4.110) and (4.111) follow since $V_j^{(1)} = \pi_j(\varphi^o) > 0$, for $j = 1, \ldots, K$ under $P'_{\varphi^o,\tilde{\phi}}$. So the conclusion of the lemma follows. ■

Based on Lemma 4.6.8 and Lemma 4.6.9, we can answer the main question of this section.

Lemma 4.6.10 *Assume conditions A1, A2, A3, A4, A5*, A6 and A7 hold. If* $\tilde{\phi} \in \tilde{\Phi}^c_K$, *with* $K \geq K^o$ *and* $K(\varphi^o, \tilde{\phi}) = 0$, *then* $\tilde{\phi} \simeq \tilde{\varphi}^o$.

Proof :

The proof follows the proof of Lemma 6 of [34]

Let $\tilde{\phi} \in \tilde{\Phi}_K^c$, with $K \geq K^o$ and $n \in \boldsymbol{N}$. Let Q be the distribution of $U^{(1)}$ under $\tilde{P}_{\varphi^o,\tilde{\phi}}$. If B is a continuity set of Q and $A \in \mathcal{R}^n$, then

$$\begin{aligned}
&\tilde{P}_{\varphi^o,\tilde{\phi}}\left\{(Y_1,\ldots,Y_n) \in A, U^{(1)} \in B\right\} \\
&= \lim_{i\to\infty} \frac{1}{n_i} \sum_{t=1}^{n_i} P'_{\varphi^o,\tilde{\phi}}\left\{(Y_t,\ldots,Y_{t+n}) \in A, U^{(t)} \in B\right\} \\
&= \lim_{i\to\infty} \frac{1}{n_i} \sum_{t=1}^{n_i} \int_B \int_A \sum_{j=1}^{K} u_j^{(t)} p_{\varphi^o}(y_t,\ldots,y_{t+n}|j)\, d\mu(y_t)\cdots d\mu(y_{t+n}) dQ_t(u^{(t)}) \\
&= \lim_{i\to\infty} \int_B \int_A \sum_{j=1}^{K} u_j^{(1)} p_{\varphi^o}(y_1,\ldots,y_n|j)\, d\mu(y_1)\cdots d\mu(y_n) dQ^{(n_i)}(u^{(1)}) \\
&= \int_B \int_A \sum_{j=1}^{K^o} u_j^{(1)} p_{\varphi^o}(y_1,\ldots,y_n|j)\, d\mu(y_1)\cdots d\mu(y_n) dQ(u^{(1)}), \qquad (4.112)
\end{aligned}$$

where Q_t are the distributions of $U^{(t)}$ under $P'_{\varphi^o,\tilde{\phi}}$ and $Q^{(n_i)}$ are the distributions of $U^{(1)}$ under $\tilde{P}^{(n_i)}_{\varphi^o,\tilde{\phi}}$. The second equality follows, since $U_j^{(t)} = P_{\varphi^o}(X_t = j \,|Y_{t-1},\ldots,Y_1)$ under $P'_{\varphi^o,\tilde{\phi}}$. From (4.112), the conditional density of $Y_1,\ldots,Y_n$ given $U^{(1)}$ under $\tilde{P}_{\varphi^o,\tilde{\phi}}$ is

$$\sum_{j=1}^{K} u_j^{(1)} p_{\varphi^o}(y_1,\ldots,y_n|j). \qquad (4.113)$$

By Lemma 4.6.8, stationarity of $\{(Y_t, U^{(t)}, V^{(t)})\}$ under $\tilde{P}_{\varphi^o,\tilde{\phi}}$, Lemma 4.6.7 and (4.113)

$$\begin{aligned}
-nH(\varphi^o) &= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right] \\
&\quad + \cdots + \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right] \\
&= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} U_j^{(1)} f(Y_1,\theta_j(\varphi^o))\right] \\
&\quad + \cdots + \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log \sum_{j=1}^{K} U_j^{(n)} f(Y_n,\theta_j(\varphi^o))\right]
\end{aligned}$$

$$
\begin{aligned}
&= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\sum_{t=1}^{n}\log\sum_{j=1}^{K}U_j^{(t)}f(Y_t,\theta_j(\varphi^o))\right] \\
&= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K}U_j^{(1)}p_{\varphi^o}(Y_1,\ldots,Y_n|X_1=j)\right] \\
&= \int\int\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j)\times\log\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j) \\
&\qquad\qquad d\mu(y_1)\cdots d\mu(y_n)dQ(u^{(1)})\,. \quad (4.114)
\end{aligned}
$$

Similarly, by Lemma 4.6.9, stationarity of $\{(Y_t,U^{(t)},V^{(t)})\}$ under $\tilde{P}_{\varphi^o,\tilde{\phi}}$, Lemma 4.6.7 and (4.113),

$$
\begin{aligned}
nH(\varphi^o,\tilde{\phi}) &= \tilde{E}_{\varphi^o,\tilde{\phi}}\left[\log\sum_{j=1}^{K}V_j^{(1)}p_{\tilde{\phi}}(Y_1,\ldots,Y_n|j)\right] \\
&= \int\int\int\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j)\times\log\sum_{j=1}^{K}v_j^{(1)}p_{\tilde{\phi}}(y_1,\ldots,y_n|j) \\
&\qquad\qquad d\mu(y_1)\cdots d\mu(y_n)d\tilde{Q}(u,v)\,, \quad (4.115)
\end{aligned}
$$

where $\tilde{Q}$ is the distribution of $(U^{(1)},V^{(1)})$ under $\tilde{P}_{\varphi^o,\tilde{\phi}}$.

Since the marginal distribution of $\tilde{Q}$ corresponding to the first coordinate is Q, then by (4.114) and (4.115)

$$
\begin{aligned}
K(\varphi^o,\tilde{\phi}) &= -H(\varphi^o)-H(\varphi^o,\tilde{\phi}) \\
&= \frac{1}{n}\int\int\int\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j) \\
&\qquad\times\log\frac{\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j)}{\sum_{j=1}^{K}v_j^{(1)}p_{\tilde{\phi}}(y_1,\ldots,y_n|j)}\,d\mu(y_1)\cdots d\mu(y_n)d\tilde{Q}(u,v)\,.
\end{aligned}
\quad (4.116)
$$

The inner integral of (4.116), that is,

$$
\int\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j)\times\log\frac{\sum_{j=1}^{K}u_j^{(1)}p_{\varphi^o}(y_1,\ldots,y_n|j)}{\sum_{j=1}^{K}v_j^{(1)}p_{\tilde{\phi}}(y_1,\ldots,y_n|j)}d\mu(y_1)\cdots d\mu(y_n) \quad (4.117)
$$

for fixed u, v, is the Kullback-Leibler divergence between two mixtures of product densities. Hence $K(\varphi^o, \tilde{\phi}) \geq 0$ as (4.117) is non-negative by Lemma 4.3.3.

If $K(\varphi^o, \tilde{\phi}) = 0$, then (4.117) is zero for $\tilde{Q}$-almost every pair u, v, implying

$$\sum_{j=1}^{K} u_j^{(1)} p_{\varphi^o}(y_1, \ldots, y_n|j) = \sum_{j=1}^{K} v_j^{(1)} p_{\tilde{\phi}}(y_1, \ldots, y_n|j) \tag{4.118}$$

by Lemma 4.3.3. Equation (4.118) can be written in another form,

$$\begin{aligned} &\sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} u_{x_1}^{(1)} \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\varphi^o) \prod_{t=1}^{n} f(y_t, \theta_{x_t}(\varphi^o)) \\ &\quad = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} v_{x_1}^{(1)} \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\tilde{\phi}) \prod_{t=1}^{n} f(y_t, \theta_{x_t}(\tilde{\phi})) \end{aligned} \tag{4.119}$$

for $\tilde{Q}$-almost every pair u, v.

However, by condition A3 and Theorem 3.2.20, mixtures of product densities from family $\{f(.,\theta) : \theta \in \Theta\}$ are identifiable. Therefore, (4.119) implies

$$\begin{aligned} &\sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} u_{x_1}^{(1)} \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\varphi^o) \delta_{(\theta_{x_1}(\varphi^o),\ldots,\theta_{x_n}(\varphi^o))} \\ &\quad = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} v_{x_1}^{(1)} \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\tilde{\phi}) \delta_{(\theta_{x_1}(\tilde{\phi}),\ldots,\theta_{x_n}(\tilde{\phi}))} . \end{aligned} \tag{4.120}$$

for $\tilde{Q}$-almost every pair u, v.

Moreover, since under $\tilde{P}_{\varphi^o,\tilde{\phi}}$

$$U_j^{(1)} = P_{\varphi^o}(X_1 = j|Y_0, Y_{-1}, \ldots),$$

then

$$\begin{aligned} \tilde{E}_{\varphi^o,\tilde{\phi}}\left[U_j^{(1)}\right] &= \tilde{E}_{\varphi^o,\tilde{\phi}}[P_{\varphi^o}(X_1 = j|Y_0, Y_{-1}, \ldots)] \\ &= E_{\varphi^o}\left[E_{\varphi^o}\left[I_{\{X_1=j\}}|Y_0, Y_{-1}, \ldots\right]\right] \\ &= E_{\varphi^o}\left[I_{\{X_1=j\}}\right] \\ &= P_{\varphi^o}(X_1 = j) \\ &= \pi_j(\varphi^o) . \end{aligned} \tag{4.121}$$

Thus from (4.120) and (4.121)

$$\sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \pi_{x_1}(\varphi^o) \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\varphi^o) \delta_{(\theta_{x_1}(\varphi^o),\ldots,\theta_{x_n}(\varphi^o))}$$
$$= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \tilde{E}_{\varphi^o,\tilde{\phi}}[v_{x_1}^{(1)}] \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\tilde{\phi}) \delta_{\theta_{x_1}(\tilde{\phi}),\ldots,\theta_{x_n}(\tilde{\phi}))} . \quad (4.122)$$

which holds for any n. By Corollary 3.3.6, $\tilde{\phi} \simeq \tilde{\varphi}^o$. ■

Corollary 4.6.11 *Assume conditions A1, A2, A3, A4, A5*, A6 and A7 hold. If $\tilde{\phi} \in \tilde{\Phi}_K^c$, with $K \geq K^o$, then $K(\varphi^o, \tilde{\phi}) = 0$ if and only if $\tilde{\phi} \simeq \tilde{\varphi}^o$.*

Corollary 4.6.12 *Assume conditions A1, A2, A3, A4, A5*, A6 and A7 hold. Let $\tilde{\phi} \in \tilde{\Phi}_K^c$, for any $K \in \boldsymbol{N}$. If $\tilde{\phi} \not\simeq \tilde{\phi}^o$, then $K(\phi^o, \tilde{\phi}) > 0$.*

Proof :

If $\tilde{\phi} \in \tilde{\Phi}_K^c$, with $K \geq K^o$, and $\tilde{\phi} \not\simeq \tilde{\phi}^o$, then by Corollary 4.6.11, $K(\phi^o, \tilde{\phi}) = K(\varphi^o, \tilde{\phi}) > 0$.

If $\tilde{\phi} \in \tilde{\Phi}_K^c$ with $K < K^o$, then by Corollary 4.5.8, $\tilde{\phi} \not\simeq \tilde{\phi}^o$. Also by Lemma 4.5.6, there is $\tilde{\phi}_1 \in \tilde{\Phi}_K^c$, such that $\tilde{\phi}_1 \simeq \tilde{\phi}$. Since $\tilde{\phi} \not\simeq \tilde{\phi}^o$, then $\tilde{\phi}_1 \simeq \tilde{\phi}^o$ and by Corollary 4.6.11 $K(\phi^o, \tilde{\phi}_1) = K(\varphi^o, \tilde{\phi}_1) > 0$, implying $K(\phi^o, \tilde{\phi}) > 0$ as $\tilde{\phi}_1 \simeq \tilde{\phi}$.

■

4.7 Uniform Convergence of the Likelihood Process

The likelihood process for $\tilde{\phi} \in \tilde{\Phi}_K^c$ is defined as

$$
\begin{aligned}
L_n(\tilde{\phi}, Y) &= \frac{1}{n} \log p_{\tilde{\phi}}(Y_1, \ldots, Y_n) \\
&= \frac{1}{n} \log \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \alpha_{x_1}^{K} f(Y_1, \theta_{x_1}(\tilde{\phi})) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}(\tilde{\phi}) f(y_t, \theta_{x_t}(\tilde{\phi}))
\end{aligned}
$$

where $n \in \boldsymbol{N}$. By Theorem 4.4.3,

$$\lim_{n \to \infty} L_n(\tilde{\phi}, Y) = H(\phi^o, \tilde{\phi}),$$

with probability one under ϕ^o and pointwise in $\tilde{\phi} \in \tilde{\Phi}_K^c$. In this section, we will show that this convergence is uniform on $\tilde{\Phi}_K^c$.

Lemma 4.7.1 *Assume conditions A1, A2, A4, A5* and A7 hold. Then*

$$\lim_{n \to \infty} L_n(\tilde{\phi}, Y) = H(\phi^o, \tilde{\phi}),$$

with probability one under ϕ^o and uniformly on $\tilde{\Phi}_K^c$.

Proof :

From Corollary 4.2.2, $\{L_n(\cdot, Y)\}$ is an equicontinuous sequence on $\tilde{\Phi}_K^c$. Since $\tilde{\Phi}_K^c$ is compact and $L_n(\tilde{\phi}, Y)$ converges to $H(\phi^o, \tilde{\phi})$ pointwise in $\tilde{\phi} \in \tilde{\Phi}_K^c$, then by Lemma 39 of [44], page 168, $L_n(\tilde{\phi}, Y)$ converges to $H(\phi^o, \tilde{\phi})$ uniformly on $\tilde{\Phi}_K^c$. ■

Corollary 4.7.2 *Assume conditions A1, A2, A4, A5* and A7 hold. Then $H(\phi^o, \cdot)$ is continuous on $\tilde{\Phi}_K^c$.*

Proof :

Since $L_n(\cdot, Y)$ is continuous on $\tilde{\Phi}_K^c$ by A4 and A5* and $L_n(\cdot, Y)$ converges to $H(\phi^o, \cdot)$ uniformly on $\tilde{\Phi}_K^c$, then from [44], page 49, $H(\phi^o, \cdot)$ is continuous on $\tilde{\Phi}_K^c$. ■

Corollary 4.7.3 *Assume conditions A1, A2, A4, A5*, A6 and A7 hold. Then the Kullback-Leibler divergence $K(\phi^o, \cdot)$ is continuous on $\tilde{\Phi}_K^c$.*

Proof :

This is a direct consequence of Corollary 4.7.2 and Corollary 4.4.5. ■

Corollary 4.7.4 *Assume conditions A1, A2, A4, A5* and A7 hold. Let B a subset of $\tilde{\Phi}_K^c$, then*

$$\lim_{n\to\infty} \left\{ \sup_{\tilde{\phi}\in B} L_n(\tilde{\phi}, Y) \right\} = \sup_{\tilde{\phi}\in B} H(\phi^o, \tilde{\phi}),$$

with probability one under ϕ^o.

Proof :

Since B is a subset of $\tilde{\Phi}_K^c$, then by Lemma 4.7.1

$$\lim_{n\to\infty} L_n(\tilde{\phi}, Y) = H(\phi^o, \tilde{\phi}),$$

with probability one under ϕ^o and uniformly in $\tilde{\phi} \in B$. Then for given $\epsilon > 0$, there exists $N_o \in \boldsymbol{N}$, such that

$$\left| L_n(\tilde{\phi}, Y) - H(\phi^o, \tilde{\phi}) \right| < \epsilon, \qquad \forall n \geq N_o \quad \text{and} \quad \forall \tilde{\phi} \in B,$$

implying

$$\begin{aligned} \left| \sup_{\tilde{\phi}\in B} L_n(\tilde{\phi}, Y) - \sup_{\tilde{\phi}\in B} H(\phi^o, \tilde{\phi}) \right| &\leq \sup_{\tilde{\phi}\in B} \left| L_n(\tilde{\phi}, Y) - H(\phi^o, \tilde{\phi}) \right| \\ &\leq \epsilon, \qquad\qquad \forall n \geq N_o, \end{aligned}$$

which means that

$$\lim_{n\to\infty} \left\{ \sup_{\tilde{\phi}\in B} L_n(\tilde{\phi}, Y) \right\} = \sup_{\tilde{\phi}\in B} H(\phi^o, \tilde{\phi}),$$

with probability one under ϕ^o. ■

4.8 The Quasi True Parameter Set

The quasi true parameter set $\mathcal{N}$ is defined as a set of parameter in $\tilde{\Phi}_K^c$ which minimize the Kullback-Leibler divergence with respect to the true parameter ϕ^o, that is

$$\mathcal{N} = \left\{ \tilde{\phi}_1 : K(\phi^o, \tilde{\phi}_1) = \inf_{\tilde{\phi} \in \tilde{\Phi}_K^c} K(\phi^o, \tilde{\phi}) \right\}. \tag{4.123}$$

Since $K(\phi^o, \cdot)$ is continuous on $\tilde{\Phi}_K^c$ by Corollary 4.7.3 and $\tilde{\Phi}_K^c$ is compact, then the infimum of $K(\phi^o, \cdot)$ over $\tilde{\Phi}_K^c$ is attained by $\tilde{\phi}_1 \in \tilde{\Phi}_K^c$. Thus the infimum sign in (4.123) may be replaced by *minimum*.

Based on the results of sections 4.6 and 4.7, the quasi true parameter set $\mathcal{N}$ can be identified as follows.

Lemma 4.8.1 *Assume conditions A1, A2, A3. A4, A5*, A6 and A7 hold.*

(a). If $K < K^o$ *and* $\tilde{\phi} \in \mathcal{N}$, *then* $K(\phi^o, \tilde{\phi}) > 0$.

(b). If $K \geq K^o$ *and* $\tilde{\phi} \in \mathcal{N}$, *then* $K(\phi^o, \tilde{\phi}) = 0$ *and* $\mathcal{N} = \{\tilde{\phi} \in \tilde{\Phi}_K^c : \tilde{\phi} \simeq \tilde{\phi}^o\}$.

Remarks 4.8.2 From part (a) of Lemma 4.8.1, if $K < K^o$, then (4.123) asserts that the quasi true parameter set $\mathcal{N}$ is the set of parameters in $\tilde{\Phi}_K^c$ which are *closest* to ϕ^o.

Proof :
If $K < K^o$, then by Corollary 4.5.8, $\tilde{\phi} \not\simeq \tilde{\phi}^o$, for every $\tilde{\phi} \in \tilde{\Phi}_K^c$. Let $\tilde{\phi} \in \mathcal{N}$, since $\tilde{\phi} \in \tilde{\Phi}_K^c$, then $\tilde{\phi} \not\simeq \tilde{\phi}^o$, implying $K(\phi^o, \tilde{\phi}) > 0$ by Corollary 4.6.12. Thus (a) follows.

For (b), since $K(\phi^o, \tilde{\phi}) = 0$, if and only if $\tilde{\phi} \simeq \tilde{\phi}^o$, when $K \geq K^o$, by Corollary 4.6.11, then we have

$$\mathcal{N} = \{\tilde{\phi} \in \tilde{\Phi}_K^c : \tilde{\phi} \simeq \tilde{\phi}^o\}.$$

■

4.9 Consistency of the Maximum Likelihood Estimator

This section presents the main result of this chapter, which is to prove that the maximum likelihood estimator

$$\hat{\phi}_n(y) = \left\{ \tilde{\phi}_1 : L_n(\tilde{\phi}_1, y) = \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} L_n(\tilde{\phi}, y) \right\}$$

is convergent with probability one under ϕ^o to the quasi true parameter set

$$\mathcal{N} = \left\{ \tilde{\phi}_1 : K(\phi^o, \tilde{\phi}_1) = \inf_{\tilde{\phi} \in \tilde{\Phi}_K^c} K(\phi^o, \tilde{\phi}) \right\}.$$

Theorem 4.9.1 *Assume conditions A1, A2, A3, A4, A5*, A6 and A7 hold. Then*

$$\lim_{n \to \infty} \hat{\phi}_n = \mathcal{N},$$

with probability one under ϕ^o.

Proof :

Here we adapt the proof of Theorem 2.2.1 of [22], page 23.

We will prove that for every $\epsilon > 0$ and P_{ϕ^o}-almost all y, there exists $N(\epsilon, y) \in \boldsymbol{N}$, such that

$$\hat{\phi}_n(y) \subset \mathcal{N}_\epsilon, \qquad \text{for } n \geq N(\epsilon, y),$$

where

$$\mathcal{N}_\epsilon = \left\{ \tilde{\phi} \in \tilde{\Phi}_K^c : d(\tilde{\phi}, \mathcal{N}) < \epsilon \right\}$$

and $d(\tilde{\phi}, \mathcal{N})$ is the distance from $\tilde{\phi}$ to the set $\mathcal{N}$, defined by

$$d(\tilde{\phi}, \mathcal{N}) = \inf_{\tilde{\phi}_1 \in \mathcal{N}} \|\tilde{\phi} - \tilde{\phi}_1\|_K.$$

From corollary 4.4.5, the quasi true parameter set $\mathcal{N}$ can be expressed as

$$\begin{aligned}
\mathcal{N} &= \left\{ \tilde{\phi}_1 \in \tilde{\Phi}_K^c : K(\phi^o, \tilde{\phi}_1) = \inf_{\tilde{\phi} \in \tilde{\Phi}_K^c} K(\phi^o, \tilde{\phi}) \right\} \\
&= \left\{ \tilde{\phi}_1 \in \tilde{\Phi}_K^c : K(\phi^o, \tilde{\phi}_1) = \inf_{\tilde{\phi} \in \tilde{\Phi}_K^c} -H(\phi^o) - H(\phi^o, \tilde{\phi}) \right\} \\
&= \left\{ \tilde{\phi}_1 \in \tilde{\Phi}_K^c : H(\phi^o, \tilde{\phi}_1) = \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} H(\phi^o, \tilde{\phi}) \right\}. \qquad (4.124)
\end{aligned}$$

Given $\epsilon > 0$. Then $\mathcal{N}_\epsilon$ is an open subset of $\tilde{\Phi}_K^c$. Let $\mathcal{N}_\epsilon^*$ be the complement of $\mathcal{N}_\epsilon$ with respect to $\tilde{\Phi}_K^c$. $\mathcal{N}_\epsilon^*$ is closed and since $\tilde{\Phi}_K^c$ is compact, then $\mathcal{N}_\epsilon^*$ is compact.

For every $\tilde{\phi} \in \mathcal{N}_\epsilon^*$, choose $\lambda_{\tilde{\phi}} > 0$, such that

$$B(\tilde{\phi}, \lambda_{\tilde{\phi}}) \subset \mathcal{N}_{\frac{\epsilon}{2}}^*, \qquad (4.125)$$

where $B(\tilde{\phi}, \lambda_{\tilde{\phi}})$ is an open Euclidean ball centered at $\tilde{\phi}$ and of radius $\lambda_{\tilde{\phi}}$. Since $\mathcal{N}_\epsilon^*$ is compact, then there exists $\{\tilde{\phi}_1, \ldots, \tilde{\phi}_M\}$ such that

$$\mathcal{N}_\epsilon^* \subset \bigcup_{i=1}^{M} B(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i}). \qquad (4.126)$$

By (4.125) and (4.126),

$$\mathcal{N}_\epsilon^* \subset \bigcup_{i=1}^{M} B(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i}) \subset \bigcup_{i=1}^{M} \bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i}) \subset \mathcal{N}_{\frac{\epsilon}{2}}^*, \qquad (4.127)$$

where $\bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})$ is a closure $B(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})$.

Let

$$H = \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} H(\phi^o, \tilde{\phi}).$$

Let $y \in \mathcal{Y}$ such that by Corollary 4.7.4,

$$\sup_{\tilde{\phi} \in \bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})} L_n(\tilde{\phi}, y) \longrightarrow \sup_{\tilde{\phi} \in \bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})} H(\phi^o, \tilde{\phi}), \tag{4.128}$$

for $i = 1, \ldots, M$. By (4.127), there exists $\alpha_i > 0$, for $i = 1, \ldots, M$, such that

$$\sup_{\tilde{\phi} \in \bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})} H(\phi^o, \tilde{\phi}) = H - \alpha_i. \tag{4.129}$$

Therefore by (4.128) and (4.129), for every $i = 1, \ldots, M$, there exists $N_i \in \boldsymbol{N}$ such that

$$\sup_{\tilde{\phi} \in \bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})} L_n(\tilde{\phi}, y) < H - \frac{\alpha_i}{2}, \qquad \forall n \geq N_i. \tag{4.130}$$

Let

$$\alpha = \min_{1 \leq i \leq M} \alpha_i \qquad \text{and} \qquad N_o = \max_{1 \leq i \leq M} N_i,$$

then by (4.130), for $n \geq N_o$,

$$\sup_{\tilde{\phi} \in \mathcal{N}_\epsilon^*} L_n(\tilde{\phi}, y) \leq \sup_{\tilde{\phi} \in \bigcup_{i=1}^M \bar{B}(\tilde{\phi}_i, \lambda_{\tilde{\phi}_i})} L_n(\tilde{\phi}, y) < H - \frac{\alpha}{2}. \tag{4.131}$$

On the otherhand, Corollary 4.7.4 and (4.124) also implies that

$$\sup_{\tilde{\phi} \in \mathcal{N}_\epsilon} L_n(\tilde{\phi}, y) \longrightarrow \sup_{\tilde{\phi} \in \mathcal{N}_\epsilon \supset \mathcal{N}} H(\phi^o, \tilde{\phi}) = H.$$

Hence, there exists $N^o \in \boldsymbol{N}$, such that

$$\sup_{\tilde{\phi} \in \mathcal{N}_\epsilon} L_n(\tilde{\phi}, y) > H - \frac{\alpha}{2}, \qquad \forall n \geq N^o. \tag{4.132}$$

Let $N = \max\{N_o, N^o\}$, then (4.131) and (4.132) implies

$$\sup_{\tilde{\phi} \in \mathcal{N}_\epsilon *} L_n(\tilde{\phi}, y) < H - \frac{\epsilon}{2} < \sup_{\tilde{\phi} \in \mathcal{N}_\epsilon} L_n(\tilde{\phi}, y), \qquad \forall n \geq \boldsymbol{N}.$$

This means that

$$\hat{\phi}_n(y) \subset \mathcal{N}_\epsilon, \qquad \forall n \in \boldsymbol{N}.$$

So the theorem is proved. ■

Chapter 5

Estimation of the Order for Hidden Markov Models

The aim of this chapter is to study the problem of order estimation for hidden Markov models. To estimate the order, we adopt the *compensated log-likelihood* technique. The idea of using this technique comes from [22]. In [22], the technique is used to estimate the order of a hidden Markov model in which the observed process takes only finitely many values. The same technique has also been used in [2] for the estimation of the structural parameters of ARMA processes.

The compensated log-likelihood technique is based on the compensation of the log-likelihood function. A *compensator*, decreasing in K (size parameter), is added to the maximum log-likelihood, and the resulting compensated log-likelihood is maximized with respect to K. The problem is then to find a proper compensator which allows the strongly consistent estimation of the order.

In this chapter, the problem will be divided into two cases, finding compensators avoiding under estimation and compensators avoiding over estimation. Then by combining these cases, the strongly consistent estimator for the order

is obtained.

This chapter begins by introducing the compensated log-likelihood in section 5.1. In section 5.2, the sufficient condition for compensators avoiding under estimation is given. Section 5.3 concentrates on finding sufficient conditions for compensators avoiding over estimation. Finally, in section 5.4, we give an example of a compensator which allows the estimator of the order to be strongly consistent.

5.1 Compensated Log-likelihood

Suppose we are given a sequence of observations $\{y_1, \ldots, y_n\}$ to be modelled. As in Chapter 4, assume that the data sample $\{y_1, \ldots, y_n\}$ is the initial segment of a realization $\{y_t\}$, generated by a process $\{Y_t\}$, which is equivalent to the observation process of a hidden Markov model, with the unknown true parameter $\phi^o = (K^o, A^o, \pi^o, \theta^o)$. Our task now is to estimate the order K^o.

As the parametric models, we will use

$$\tilde{\Phi}^c = \bigcup_{K \in \boldsymbol{N}} \tilde{\Phi}^c_K ,$$

where

$$\tilde{\Phi}^c_K = \Big\{ \tilde{\phi} \ : \ \tilde{\phi} = (K, A, \theta), \text{ where } A \text{ and } \theta \text{ satisfy :}$$
$$A = (\alpha_{ij}), \quad \alpha_{ij} \geq 0, \quad \sum_{j=1}^{K} \alpha_{ij} = 1, \quad i, j = 1, \ldots, K$$
$$\theta = (\theta_i)^T, \quad \theta_i \in \Theta^c, \quad i = 1, \ldots, K \Big\}.$$

Recall that under $\tilde{\phi} \in \tilde{\Phi}^c_K$, $Y_1, \ldots, Y_n$ has the joint density function

$$p_{\tilde{\phi}}(y_1, \ldots, y_n) = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \alpha^K_{x_1} f(y_1, \theta_{x_1}(\tilde{\phi})) \prod_{t=2}^{n} \alpha_{x_{t-1}, x_t}(\tilde{\phi}) f(y_t, \theta_{x_t}(\tilde{\phi})),$$

where $\alpha^K = (\alpha_i^K)$ is the initial probability vector, with $\alpha_i^K > 0$, for $i = 1, \ldots, K$, $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ are the coordinate projections on $\tilde{\Phi}_K^c$, for $i, j = 1, \ldots, K$.

Let $\tilde{\phi}^o = (K^o, A^o, \theta^o)$. From Corollary 4.5.8, $\tilde{\Phi}_K^c$ contains no parameter equivalent to $\tilde{\phi}^o$, if $K < K^o$, and at least finitely many parameters equivalent to $\tilde{\phi}^o$, if $K = K^o$. If $K > K^o$, there are infinitely many parameters in $\tilde{\Phi}_K^c$ equivalent to $\tilde{\phi}^o$.

Throughout this section we will assume that the following conditions hold.

A1. The transition probability matrix A^o is *irreducible.*

A2. π^o is a *stationary* probability distribution of A^o.

A3. The family of finite mixtures on $\{f(\cdot, \theta) : \theta \in \Theta\}$ is *identifiable*

A4. $f(\cdot, \cdot) > 0$ and continuous on $\mathcal{Y} \times \Theta$. For each y, $f(y, \cdot)$ *vanishes* at infinity.

A5′. For each $K \in \boldsymbol{N}$ and $i, j = 1, \ldots, K$, $\alpha_{ij}(\cdot)$ and $\theta_i(\cdot)$ are continuous functions on $\tilde{\Phi}_K^c$

A6. $E_{\phi^o}\left[\left|\log f(Y_1, \theta_i^o)\right|\right] < \infty$, for $i = 1, \ldots, K^o$.

A7. For every $\theta \in \Theta$, $E_{\phi^o}\left[(\log f(Y_1, \theta))^+\right] < \infty$.

As in Chapter 4, for $y = \{y_t\}$, define the log-likelihood functions $L_{K,n}(\cdot, y)$ on $\tilde{\Phi}_K^c$ by

$$L_{K,n}(\tilde{\phi}, y) = \frac{1}{n} \log p_{\tilde{\phi}}(y_1, \ldots, y_n), \qquad n \in \boldsymbol{N}.$$

For $K, n \in \boldsymbol{N}$, let

$$L_{K,n}^*(y) = \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} L_{K,n}(\tilde{\phi}, y).$$

Suppose we attempt to use the *classical maximum likelihood* technique to estimate the order. Then we estimate the order K^o by random variable $\widehat{K}_n$, such that $1 \leq \widehat{K}_n \leq B$ and

$$\widehat{K}_n(y) = \min\left\{\widehat{K} \ : \ L^*_{\widehat{K},n}(y) = \max_{1\leq K\leq B} L^*_{K,n}(y)\right\},$$

where B is a positive integer imposed by concrete computing limitation. However, according to [2], for large n, the estimator $\widehat{K}_n$ is always equal to B and if we do not impose any bound B, the situation is worse, since the estimator $\widehat{K}_n$ tends to ∞, as $n \to \infty$. Hence, $\widehat{K}_n$ is *not consistent.*

Therefore, to the log-likelihood should be added a *compensator*, which will decrease the likelihood, when the size increases. This will discourage the selection of model with an *excessive* size.

Definition 5.1.1 *A* ***compensator*** *is any deterministic sequence of functions* $\delta_n : \boldsymbol{N} \to \boldsymbol{R}^+$, *such that* $\delta_n(K) \leq \delta_n(\widehat{K})$, *if* $K \leq \widehat{K}$. *A* ***compensated log-likelihood*** *is defined by*

$$C_{K,n}(y) = L^*_{K,n}(y) - \delta_n(K)$$

for $y \in \mathcal{Y}^\infty$ *and* $K, n \in \boldsymbol{N}$. *The* ***estimator of the order*** $\widehat{K}_n$ *is then defined by*

$$\widehat{K}_n(y) = \min\left\{\widehat{K} \ : \ C_{\widehat{K},n}(y) = \max_{1\leq K\leq B} C_{K,n}(y)\right\}, \tag{5.1}$$

where B *is a positive integer imposed by concrete computing limitation.*

The problem is now to find a proper compensator which allows $\widehat{K}_n$ to be *strongly consistent*, that is,

$$\widehat{K}_n \longrightarrow K^o, \qquad \text{with probability one under } \phi^o,$$

as $n \to \infty$.

5.2 Compensators Avoiding under Estimation

Based on the results of Chapter 4, we obtain the sufficient condition for compensators to avoid under estimation in the following theorem. The idea of the theorem and its proof comes from [22]. The same sufficient condition for the same type of hidden Markov model with ours, is also obtained by [46]. However, [46] used stronger assumptions and different approach in calculating the Kullback-Leibler divergence, which is based on the information of m-dimensional joint distribution of the observed process , for some integer $m \geq 2K^o$.

Theorem 5.2.1 (Compensators avoiding under estimation) *Assume conditions A1, A2, A3, A4, A5', A6 and A7 hold. If*

$$\lim_{n\to\infty} \delta_n(K) = 0, \qquad \textit{for every } K \in \boldsymbol{N},$$

then

$$\liminf_{n\to\infty} \widehat{K}_n \geq K^o,$$

with probability one under ϕ^o.

Proof :

Suppose that for every $K \in \boldsymbol{N}$,

$$\lim_{n\to\infty} \delta_n(K) = 0. \tag{5.2}$$

Let K be any positive integer such that $K < K^o$. By (5.2)

$$\begin{aligned}
&\lim_{n\to\infty} \Big\{ - C_{K,n}(y) + C_{K^o,n}(y) \Big\} \\
&\quad = \lim_{n\to\infty} \Big\{ - L^*_{K,n}(y) + L^*_{K^o,n}(y) + \delta_n(K) - \delta_n(K^o)(y) \Big\} \\
&\quad = \lim_{n\to\infty} \Big\{ - L^*_{K,n}(y) + L^*_{K^o,n}(y) \Big\} \\
&\quad = \lim_{n\to\infty} \Big\{ - L^*_{K,n}(y) + L_{K^o,n}(\tilde{\phi}^o, y) - L_{K^o,n}(\tilde{\phi}^o, y) + L^*_{K^o,n}(y) \Big\}, \qquad (5.3)
\end{aligned}$$

provides the limit at the right hand side exists.

By Corollary 4.7.4, Corollary 4.4.5 and part (a) of Lemma 4.8.1,

$$\begin{aligned}
\lim_{n\to\infty}\Big\{ - L^*_{K,n}(y) + L_{K^o,n}(\tilde{\phi}^o, y)\Big\} &= \lim_{n\to\infty}\Big\{ - \sup_{\tilde{\phi}\in\tilde{\Phi}^c_K} L_{K,n}(\tilde{\phi}, y) + L_{K^o,n}(\tilde{\phi}^o, y)\Big\} \\
&= \lim_{n\to\infty}\Big\{ \inf_{\tilde{\phi}\in\tilde{\Phi}^c_K} L_{K^o,n}(\tilde{\phi}^o, y) - L_{K,n}(\tilde{\phi}, y)\Big\} \\
&= \inf_{\tilde{\phi}\in\tilde{\Phi}^c_K}\Big\{ \lim_{n\to\infty} L_{K^o,n}(\tilde{\phi}^o, y) - L_{K,n}(\tilde{\phi}, y)\Big\} \\
&= \inf_{\tilde{\phi}\in\tilde{\Phi}^c_K} K(\phi^o, \tilde{\phi}) \\
&= \gamma(K) > 0, \qquad (5.4)
\end{aligned}$$

with probability one under ϕ^o.

Similarly, by Corollary 4.7.4, Corollary 4.4.5 and part (b) of Lemma 4.8.1,

$$\begin{aligned}
\lim_{n\to\infty}\Big\{L_{K^o,n}(\tilde{\phi}^o, y) - L^*_{K^o,n}(y)\Big\} &= \lim_{n\to\infty}\Big\{L_{K^o,n}(\tilde{\phi}^o, y) - \sup_{\tilde{\phi}\in\tilde{\Phi}^c_{K^o}} L_{K,n}(\tilde{\phi}, y)\Big\} \\
&= \lim_{n\to\infty}\Big\{ \inf_{\tilde{\phi}\in\tilde{\Phi}^c_{K^o}} L_{K^o,n}(\tilde{\phi}^o, y) - L_{K^o,n}(\tilde{\phi}, y)\Big\} \\
&= \inf_{\tilde{\phi}\in\tilde{\Phi}^c_{K^o}}\Big\{ \lim_{n\to\infty} L_{K^o,n}(\tilde{\phi}^o, y) - L_{K^o,n}(\tilde{\phi}, y)\Big\} \\
&= \inf_{\tilde{\phi}\in\tilde{\Phi}^c_{K^o}} K(\phi^o, \tilde{\phi}) \\
&= 0, \qquad (5.5)
\end{aligned}$$

with probability one under ϕ^o.

From (5.3), (5.4) and (5.5), for $K < K^o$

$$\lim_{n\to\infty}\Big\{ - C_{K,n}(y) + C_{K^o,n}(y)\Big\} = \gamma(K) > 0, \qquad (5.6)$$

with probability one under ϕ^o.

Suppose there is a subsequence $\widehat{K}_{n_i}(y)$ such that

$$\widehat{K}_{n_i}(y) \longrightarrow L, \qquad \text{as } i \to \infty, \qquad (5.7)$$

where $1 \leq L < K^o$. Since $\widehat{K}_{n_i}(y) \in \boldsymbol{N}$, for every n, then from (5.7), there is $M \in \boldsymbol{N}$, such that

$$\widehat{K}_{n_i}(y) = L, \qquad \forall i \geq M,$$

implying

$$\begin{aligned} \limsup_{i\to\infty} \Big\{ - C_{\widehat{K}_{n_i}(y),n_i}(y) + C_{K^o,n_i}(y) \Big\} &= \limsup_{i\to\infty} \Big\{ - C_{L,n_i}(y) + C_{K^o,n_i}(y) \Big\} \\ &= \gamma(L) > 0, \end{aligned} \tag{5.8}$$

by (5.6).

However, by definition of $\widehat{K}_n$,

$$\Big\{ - C_{\widehat{K}_n(y),n}(y) + C_{K^o,n}(y) \Big\} \leq 0, \qquad \text{for every } n \in \boldsymbol{N},$$

implying

$$\limsup_{i\to\infty} \Big\{ - C_{\widehat{K}_{n_i}(y),n_i}(y) + C_{K^o,n_i}(y) \Big\} \leq 0,$$

which contradicts to (5.8)

Therefore, every convergent subsequence of $\widehat{K}_n(y)$ must converge to a limit, which is greater or equal to K^o. Hence we have

$$\liminf_{n\to\infty} \widehat{K}_n(y) \geq K^o,$$

with probability one under ϕ^o. ■

5.3 Compensators Avoiding over Estimation

This section aims to find sufficient conditions for the compensators to avoid over estimation. The crucial problem in finding these sufficient conditions is to determine the *almost sure rate of growth of the maximized log-likelihood ratio*

$$\log \frac{p_{ML_K}(y_1, \ldots, y_n)}{p_{\phi^o}(y_1, \ldots, y_n)} \tag{5.9}$$

for every $K \geq K^o$, where

$$p_{ML_K}(y_1, \ldots, y_n) = \sup_{\tilde{\phi} \in \tilde{\Phi}^c_K} p_{\tilde{\phi}}(y_1, \ldots, y_n).$$

The *almost sure rate*, denoted by $O_{a.s.}$, is defined as follows.

Definition 5.3.1 *Let $\{Z_t : t \in \boldsymbol{N}\}$ be a sequence of random variables and $\{\alpha_t\}$ a sequence of positive real numbers. We say that that $Z_t = O_{a.s.}(\alpha_t)$ if there exists a positive random variable C almost surely finite such that*

$$|Z_t| \leq C\alpha_t, \qquad \forall t \in \boldsymbol{N}.$$

Using Csiszar lemma as a basic tool, we obtain that for any $K \in \boldsymbol{N}$,

$$\log \frac{p_{ML_K}(y_1, \ldots, y_n)}{p_{\phi^o}(y_1, \ldots, y_n)} = O_{a.s.}(\log n). \tag{5.10}$$

Based on this, the compensators avoiding over estimation can then be constructed.

For convenience, this section will be divided into four subsections. In the first subsection we introduce Csiszar lemma. This lemma initially holds for processes taking values on a finite set. However, in subsection 5.3.2, it can be shown that the conclusion of the lemma also holds for some processes which takes values on an infinite set. In subsection 5.3.3, we apply the results of subsections 5.3.1 and 5.3.2 to hidden Markov model. From subsection 5.3.3, we then obtain the rate of growth of the maximized log-likelihood ratio, which is presented in subsection 5.3.4. Finally, in subsection 5.3.5, the sufficient conditions for compensators avoiding over estimation are given.

5.3.1 Csiszar lemma

This subsection studies the Csiszar lemma, which will be a very useful tool for determining the rate of growth of the maximized log-likelihood ratio.

In order to prove Csiszar lemma, the following integral will be needed.

Lemma 5.3.2 *Let*

$$V = \left\{ p = (p_1, \ldots, p_K) \in \boldsymbol{R}^K \ : \ p_i \geq 0, \ i = 1, \ldots, K, \ \sum_{i=1}^{K} p_i = 1 \right\}$$

and

$$\alpha_i > -1, \qquad \textit{for } i = 1, \ldots, K.$$

Then

$$\int_V \prod_{i=1}^{K} p_i^{\alpha_i} \, dp = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i + 1)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i + K\right)}.$$

Proof :

The proof is based on the fact that for $x, y > 0$,

$$\int_0^1 t^{x-1}(1-t)^{y-1} \, dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{5.11}$$

see for example, [21], page 49.

Let $\alpha_i > -1$, for $i = 1, \ldots, K$, then by (5.11),

$$\begin{aligned}
\int_V \prod_{i=1}^{K} p_i^{\alpha_i} \, dp \ &= \ \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{i=1}^{K-3} p_i} \int_0^{1-\sum_{i=1}^{K-2} p_i} P_1^{\alpha_1} p_2^{\alpha^2} \cdots p_{K-2}^{\alpha_{K-2}} p_{K-1}^{\alpha_{K-1}} \\
&\quad \times \Big(1 - \sum_{i=1}^{K-1} p_i\Big)^{\alpha_K} \, dp_{K-1} \, dp_{K-2} \cdots dp_2 \, dp_1 \\
&= \ \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{i=1}^{K-3} p_i} \int_0^1 p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_{K-2}^{\alpha_{K-2}} \\
&\quad \times \Big(1 - \sum_{i=1}^{K-2} p_i\Big)^{\alpha_{K-1}+\alpha_K+1} t^{\alpha_{K-1}} (1-t)^{\alpha_K} \, dt \, dp_{K-2} \cdots dp_2 \, dp_1
\end{aligned}$$

$$= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{i=1}^{K-3} p_i} p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_{K-2}^{\alpha_{K-2}} \times \Big(1 - \sum_{i=1}^{K-2} p_i\Big)^{\alpha_{K-1}+\alpha_K+1} dp_{K-2} \cdots dp_2\, dp_1 \times \int_0^1 t^{\alpha_{K-1}} (1-t)^{\alpha_K}\, dt \tag{5.12}$$

$$= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{i=1}^{K-4} p_i} \int_0^{1-\sum_{i=1}^{K-3} p_i} P_1^{\alpha_1} p_2^{\alpha^2} \cdots p_{K-3}^{\alpha_{K-3}} p_{K-2}^{\alpha_{K-2}} \times \Big(1 - \sum_{i=1}^{K-2} p_i\Big)^{\alpha_{K-1}+\alpha_K+1} dp_{K-2}\, dp_{K-3} \cdots dp_2\, dp_1 \times \frac{\Gamma(\alpha_{K-1}+1)\Gamma(\alpha_K+1)}{\Gamma(\alpha_{K-1}+\alpha_K+2)}. \tag{5.13}$$

The equation (5.12) is obtained by letting

$$p_{K-1} = 1 - \sum_{i=1}^{K-2} p_i.$$

Using the same technique, calculate the integral in (5.13) with respect to $p_{K-2}, p_{K-3}, \ldots, p_1$ respectively, then we have

$$\begin{aligned} \int \prod_{i=1}^{K} p_i\, dp &= \frac{\Gamma(\alpha_1+1)\Gamma(\sum_{i=1}^K \alpha_i + K - 1)}{\Gamma(\sum_{i=1}^K \alpha_i + K)} \cdot \frac{\Gamma(\alpha_2+1)\Gamma(\sum_{i=1}^K \alpha_i + K - 2)}{\Gamma(\sum_{i=1}^K \alpha_i + K - 1)} \\ &\quad \cdots \frac{\Gamma(\alpha_{K-2}+1)\Gamma(\alpha_{K-1}+\alpha_K+2)}{\Gamma(\alpha_{K-2}+\alpha_{K-1}+\alpha_K+3)} \cdot \frac{\Gamma(\alpha_{K-1}+1)\Gamma(\alpha_K+1)}{\Gamma(\alpha_{K-1}+\alpha_K+2)} \\ &= \frac{\prod_{i=1}^K \Gamma(\alpha_i+1)}{\Gamma(\sum_{i=1}^K \alpha_i + K)}. \end{aligned}$$

■

Now consider an independent identically distributed process $\{Z_t : t \in \boldsymbol{N}\}$, with values in $\{1, \ldots, K\}$ and distribution P. Let

$$P(Z_t = i) = p_i, \qquad i = 1, \ldots, K,$$

then the joint density function of $Z_1, \ldots, Z_n$,

$$P(z_1, \ldots, z_n) = \prod_{i=1}^{n} P(Z_t = z_t) = \prod_{i=1}^{K} p_i^{n_i},$$

where n_i is the number of times i occurs in $z_1, \ldots, z_n$. This probability is a maximum if

$$p_i = \frac{n_i}{n}, \qquad i = 1, \ldots, K.$$

Hence the maximum likelihood estimate is given by

$$P_{ML}(z_1, \ldots, z_n) = \prod_{i=1}^{K} \left(\frac{n_i}{n}\right)^{n_i}.$$

Define a mixture distribution Q such that

$$Q(z_1, \ldots, z_n) = \int_V \Big(\prod_{i=1}^{K} p_i^{n_i}\Big) \cdot \nu(p)\, dp,$$

where

$$\nu(p) = \frac{\Gamma\Big(\sum_{i=1}^{K} \alpha_i + K\Big)}{\prod_{i=1}^{K} \Gamma(\alpha_i + 1)} \cdot \prod_{i=1}^{K} p_i^{\alpha_i},$$

with

$$\alpha_i > -1, \qquad \text{for } i = 1, \ldots, K.$$

Since $p_i \geq 0$, for $i = 1, \ldots, K$ and $\sum_{i=1}^{K} p_i = 1$, then by lemma 5.3.2

$$\int_V \nu(p)\, dp = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i + 1)}{\prod_{i=1}^{K} \Gamma(\alpha_i + 1)} \int_V \prod_{i=1}^{K} p_i^{\alpha_i}\, dp = 1.$$

Also by lemma 5.3.2,

$$\begin{aligned}
Q(z_1, \ldots, z_n) &= \frac{\Gamma\Big(\sum_{i=1}^{K} \alpha_i + K\Big)}{\prod_{i=1}^{K} \Gamma(\alpha_i + 1)} \int_V \prod_{i=1}^{K} p_i^{n_i} \cdot \prod_{i=1}^{K} p_i^{\alpha_i}\, dp \\
&= \frac{\Gamma\Big(\sum_{i=1}^{K} \alpha_i + K\Big)}{\prod_{i=1}^{K} \Gamma(\alpha_i + 1)} \int_V \prod_{i=1}^{K} p_i^{n_i + \alpha_i}\, dp \\
&= \frac{\Gamma\Big(\sum_{i=1}^{K} \alpha_i + K\Big)}{\prod_{i=1}^{K} \Gamma(\alpha_i + 1)} \cdot \frac{\prod_{i=1}^{K} \Gamma(n_i + \alpha_i + 1)}{\Gamma(\sum_{i=1}^{K} n_i + \alpha_i + K)} \\
&= \frac{\Gamma\Big(\sum_{i=1}^{K} \alpha_i + K\Big)}{\Gamma\Big(\sum_{i=1}^{K} n_i + \alpha_i + K\Big)} \prod_{i=1}^{K} \frac{\Gamma(n_i + \alpha_i + 1)}{\Gamma(\alpha_i + 1)}. \qquad (5.14)
\end{aligned}$$

Lemma 5.3.3 (Csiszar [13]) *If Q is defined by (5.14), with $\alpha_i = -\frac{1}{2}$, for $i = 1, \ldots, K$, then*

$$\begin{aligned} \log \frac{P_{ML}(z_1, \ldots, z_n)}{Q(z_1, \ldots, z_n)} &\leq \log \frac{\Gamma\left(n + \frac{K}{2}\right) \cdot \Gamma\left(\frac{1}{2}\right)}{\Gamma\left(n + \frac{1}{2}\right) \cdot \Gamma\left(\frac{K}{2}\right)} \\ &\leq \frac{K-1}{2} \log n - \log \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} + \epsilon_n, \end{aligned}$$

where $\lim_{n \to \infty} \epsilon_n = 0$.

Proof :

Let $\alpha_i = -\frac{1}{2}$, for $i = 1, \ldots, K$, then by (5.14)

$$Q(z_1, \ldots, z_n) = \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(n + \frac{K}{2}\right)} \prod_{i=1}^{K} \frac{\Gamma\left(n_i + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}. \tag{5.15}$$

To prove the first inequality, it must be shown that

$$P_{ML}(z_1, \ldots, z_n) \leq Q(z_1, \ldots, z_n) \cdot \frac{\Gamma\left(n + \frac{K}{2}\right) \cdot \Gamma\left(\frac{1}{2}\right)}{\Gamma\left(n + \frac{1}{2}\right) \cdot \Gamma\left(\frac{K}{2}\right)},$$

that is by (5.15)

$$\begin{aligned} \prod_{i=1}^{K} \left(\frac{n_i}{n}\right)^{n_i} &\leq \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(n + \frac{K}{2}\right)} \cdot \left(\prod_{i=1}^{K} \frac{\Gamma\left(n_i + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} \right) \cdot \frac{\Gamma\left(n + \frac{K}{2}\right) \cdot \Gamma\left(\frac{1}{2}\right)}{\Gamma\left(n + \frac{1}{2}\right) \cdot \Gamma\left(\frac{K}{2}\right)} \\ &= \frac{\prod_{i=1}^{K} \left(n_i - \frac{1}{2}\right)\left(n_i - \frac{3}{2}\right) \cdots \frac{1}{2}}{\left(n - \frac{1}{2}\right)\left(n - \frac{3}{2}\right) \cdots \frac{1}{2}}. \end{aligned} \tag{5.16}$$

However for any $m \in \boldsymbol{N}$,

$$\begin{aligned} \left(m - \frac{1}{2}\right)\left(m - \frac{3}{2}\right) \cdots \frac{1}{2} &= \left(\frac{2m-1}{2}\right)\left(\frac{2m-3}{2}\right) \cdots \frac{1}{2} \\ &= \frac{(2m-1)(2m-3) \cdots 1}{2^m} \\ &= \frac{(2m-1)(2m-3) \cdots 1}{2^m} \cdot \frac{2m(2m-2) \cdots 2}{2m(2m-2) \cdots 2} \\ &= \frac{(2m)!}{2^m} \cdot \frac{1}{2(m) \cdot 2(m-1) \cdots 2 \cdot 1} \\ &= \frac{(2m)!}{2^m} \cdot \frac{1}{2^m \cdot m!} \\ &= \frac{2m(2m-1) \cdots (m+1)}{2^{2m}}. \end{aligned}$$

So (5.16) can be expressed as

$$\begin{aligned}\prod_{i=1}^{K}\left(\frac{n_i}{n}\right)^{n_i} &\leq \prod_{i=1}^{K}\frac{2n_i(2n_i-1)\cdots(n_i+1)}{2^{2n_i}}\cdot\frac{2^n}{2n(2n-1)\cdots(n+1)}\\ &= \frac{\prod_{i=1}^{K}2n_i(2n_i-1)\cdots(n_i+1)}{2n(2n-1)(n+1)},\end{aligned}$$

or in a long form

$$\begin{aligned}&\underbrace{\overbrace{\left(\frac{n_1}{n}\cdots\frac{n_1}{n}\right)}^{n_1\text{ terms}}\overbrace{\left(\frac{n_2}{n}\cdots\frac{n_2}{n}\right)}^{n_2\text{ terms}}\cdots\overbrace{\left(\frac{n_K}{n}\cdots\frac{n_K}{n}\right)}^{n_K\text{ terms}}}_{n\text{ terms}}\\ &\leq \frac{\overbrace{\{2n_1(2n_1-1)\cdots(n_1+1)\}}^{n_1\text{ terms}}\cdots\overbrace{\{2n_K(2n_K-1)\cdots(n_K+1)\}}^{n_K\text{ terms}}}{\underbrace{2n(2n-1)(2n-2)\cdots(n+1)}_{n\text{ terms}}}.\end{aligned}\tag{5.17}$$

So (5.17) will be proved if we can show that it is possible to assign to each $l=1,\ldots,n$, in a one to one manner, a pair (i,j), $1\leq i\leq K$, $1\leq j\leq n_i$, such that

$$\frac{n_i}{n}\leq\frac{n_i+j}{n+l}.\tag{5.18}$$

For any given l and i, (5.18) holds if and only if

$$\begin{aligned}j &\geq (n+l)\cdot\frac{n_i}{n}-n_i\\ &= n_i\cdot\frac{l}{n}.\end{aligned}$$

Hence the number of j that satisfy (5.18) is greater than $n_i-n_i\frac{l}{n}$ and the total number of pairs (i,j), $1\leq i\leq K$, $1\leq j\leq n_i$ satisfying (5.18) is greater than

$$\sum_{i=1}^{K}\left(n_i-n_i\cdot\frac{l}{n}\right)=n-l.$$

It follows that, if we assign to $l=n$, any (i,j) satisfying (5.18), that is, i may be chosen arbitrary and $j=n_i$, then recursively assign to each $l=n-1,n-2,\ldots,$ etc, a pair (i,j) satisfying (5.18) that were not assign previously. We never get

stuck, at each step, there will be at least one free pair (i,j), because the total number of pairs (i,j) satisfying (5.18) is greater than $(n-l)$, that is the number of pairs already assigned. So the first inequality is proved.

The proof for the second inequality, use the Stirling's formula for Γ-function, that is,

$$\Gamma(z) \approx \sqrt{2\pi} e^{-z} z^{z-\frac{1}{2}}, \tag{5.19}$$

for $z \in W(\delta) = \{z \in C : z \neq 0, -\pi + \delta \leq Arg z \leq \pi - \delta\}$, where $0 < \delta < \pi$ (see for example, Stromberg, K.R., [49], page 468).

By (5.19),

$$\begin{aligned}
\log \frac{\Gamma\left(n+\frac{K}{2}\right)}{\Gamma\left(n+\frac{1}{2}\right)} &\approx \log \frac{\sqrt{2\pi}\, e^{-n-\frac{K}{2}} \left(n+\frac{K}{2}\right)^{n+\frac{K}{2}-\frac{1}{2}}}{\sqrt{2\pi}\, e^{-n-\frac{1}{2}} \left(n+\frac{1}{2}\right)^{n}} \\
&= \log \left\{ e^{-\left(\frac{K-1}{2}\right)} \left(\frac{n+\frac{K}{2}}{n+\frac{1}{2}}\right)^{n} \left(n+\frac{K}{2}\right)^{\frac{K-1}{2}} \right\} \\
&= -\left(\frac{K-1}{2}\right) + n \log \left(\frac{n+\frac{K}{2}}{n+\frac{1}{2}}\right) + \frac{K-1}{2} \log \left(n+\frac{K}{2}\right).
\end{aligned} \tag{5.20}$$

Expand $\log\left(n+\frac{K}{2}\right)$ using Taylor's formula,

$$\log\left(n+\frac{K}{2}\right) = \log n + \frac{K}{2} \cdot \frac{1}{n} - \frac{1}{2!} \cdot \left(\frac{K}{2}\right)^2 \cdot \frac{1}{n^2} + R(3), \tag{5.21}$$

where

$$R(3) = \frac{1}{3!} \cdot \left(\frac{K}{2}\right)^3 \cdot \frac{2}{\zeta^3}, \qquad \text{for } n \leq \zeta \leq n + \tfrac{K}{2}.$$

Let

$$\epsilon_n = \left\{ \frac{K}{2} \cdot \frac{1}{n} - \frac{1}{2!} \cdot \left(\frac{K}{2}\right)^2 \cdot \left(\frac{1}{n^2}\right) + R(3) \right\} \left(\frac{K-1}{2}\right) + n \log \left(\frac{n+\frac{K}{2}}{n+\frac{1}{2}}\right). \tag{5.22}$$

Then it is clear that

$$\lim_{n\to\infty} \epsilon_n = 0.$$

Thus by (5.20), (5.21) and (5.22),

$$\begin{aligned}
\log \frac{\Gamma\left(n+\frac{K}{2}\right)\cdot\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(n+\frac{1}{2}\right)\cdot\Gamma\left(\frac{K}{2}\right)} &= \log \frac{\Gamma\left(n+\frac{K}{2}\right)}{\Gamma\left(n+\frac{1}{2}\right)} - \log \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} \\
&\approx -\left(\frac{K-1}{2}\right) + \left(\frac{K-1}{2}\right)\log n + \epsilon_n - \log \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} \\
&\leq \left(\frac{K-1}{2}\right)\log n - \log \frac{\Gamma\left(\frac{K}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} + \epsilon_n.
\end{aligned}$$

■

5.3.2 Extension of Csiszar lemma

In this subsection we try to extend the Csiszar lemma to processes which take values on an infinite set.

Let $\{Z_t : t \in \boldsymbol{N}\}$ be an independent identically distributed random process. Let $f(\cdot,\theta)$ be the density function of Z_t under the probability measure P_θ, for $\theta \in \Theta^c$. Then the joint density function of $Z_1,\ldots,Z_n$

$$p_\theta(z_1,\ldots,z_n) = \prod_{i=1}^{n} f(z_i,\theta).$$

Suppose that θ^* is a maximum likelihood estimator, then the maximum likelihood estimate is given by

$$p_{ML}(z_1,\ldots,z_n) = \prod_{i=1}^{n} f(z_i,\theta^*).$$

Next two lemmas give the examples of distribution functions having similar property as in Csiszar lemma.

Lemma 5.3.4 *If $\{Z_t : t \in \boldsymbol{N}\}$ is an independent identically distributed Poisson process, then there exists a probability measure Q with the corresponding*

density function q such that

$$\log \frac{p_{ML}(z_1, \ldots, z_n)}{q(z_1, \ldots, z_n)} \leq \log n - C + \epsilon_n,$$

where C is a constant and $\lim_{n\to\infty} \epsilon_n = 0$.

Proof :

Let

$$f(z, \theta) = \frac{e^{-\theta}\theta^z}{z!}, \qquad z \in \{0, 1, 2, \ldots\}, \qquad \theta \in (0, \infty).$$

Then the joint density function of $Z_1, \ldots, Z_n$ is

$$\begin{aligned} p_\theta(z_1, \ldots, z_n) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{z_i}}{z_i!} \\ &= \frac{e^{-n\theta}\theta^{(\sum_{i=1}^n z_i)}}{z_1! z_2! \cdots z_n!}. \end{aligned} \tag{5.23}$$

Equation (5.23) is maximum if

$$\theta = \frac{\sum_{i=1}^n z_i}{n}.$$

Hence the maximum likelihood estimate is given by

$$p_{ML}(z_1, \ldots, z_n) = \frac{e^{-(\sum_{i=1}^n z_i)}(\sum_{i=1}^n z_i)^{(\sum_{i=1}^n z_i)} n^{-(\sum_{i=1}^n z_i)}}{z_1! z_2! \cdots z_n!}.$$

Define probability measure Q through its joint density function q, which is defined by

$$q(z_1, \ldots, z_n) = \int_0^\infty p_\theta(z_1, \ldots, z_n)\, \nu(\theta)\, d\theta,$$

where

$$\nu(\theta) = e^{-\theta}.$$

It is clear that

$$\int_0^\infty \nu(\theta)\, d\theta = \int_0^\infty e^{-\theta}\, d\theta = 1.$$

Then by lemma 4.1.2,

$$\begin{aligned} q(z_1,\ldots,z_n) &= \int_0^\infty \frac{e^{-n\theta}\theta^{(\sum_{i=1}^n z_i)}}{z_1!z_2!\cdots z_n!}\cdot e^{-\theta}\,d\theta \\ &= \int_0^\infty \frac{e^{-(n+1)\theta}\theta^{(\sum_{i=1}^n z_i)}}{z_1!z_2!\cdots z_n!}\,d\theta \\ &= \frac{\Gamma(\sum_{i=1}^n z_i+1)(n+1)^{-(\sum_{i=1}^n z_i+1)}}{z_1!z_2!\cdots z_n!}. \end{aligned}$$

Using Stirling's formula ([1], page 257),

$$\begin{aligned} \frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} &= \frac{e^{-(\sum_{i=1}^n z_i)}(\sum_{i=1}^n z_i)^{(\sum_{i=1}^n z_i)}n^{-(\sum_{i=1}^n z_i)}}{\Gamma(\sum_{i=1}^n z_i+1)(n+1)^{-(\sum_{i=1}^n z_i+1)}} \\ &= \frac{e^{-(\sum_{i=1}^n z_i)}(\sum_{i=1}^n z_i)^{(\sum_{i=1}^n z_i)}n^{-(\sum_{i=1}^n z_i)}}{(\sum_{i=1}^n z_i)!\,(n+1)^{-(\sum_{i=1}^n z_i+1)}} \\ &\le \frac{e^{-(\sum_{i=1}^n z_i)}(\sum_{i=1}^n z_i)^{(\sum_{i=1}^n z_i)}n^{-(\sum_{i=1}^n z_i)}}{\sqrt{2\pi}\,(\sum_{i=1}^n z_i)^{(\sum_{i=1}^n z_i)}\sqrt{\sum_{i=1}^n z_i}\;e^{-(\sum_{i=1}^n z_i)}} \\ &\quad\times\frac{1}{(n+1)^{-(\sum_{i=1}^n z_i)}(n+1)^{-1}} \\ &\le \frac{1}{\sqrt{2\pi}}\left(\frac{n}{n+1}\right)^{-(\sum_{i=1}^n z_i)}(n+1). \end{aligned}$$

Hence

$$\log\frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} \le -\left(\sum_{i=1}^n z_i\right)\log\left(\frac{n}{n+1}\right) - \log\sqrt{2\pi} + \log(n+1). \quad (5.24)$$

Expand $\log(n+1)$ using Taylor's expansion,

$$\log(n+1) = \log n + \frac{1}{n} - \frac{1}{2n^2} + R(3), \quad (5.25)$$

where

$$R(3) = \frac{2}{3!}\cdot\frac{1}{\zeta^3}, \qquad \text{where } n \le \zeta \le n+1.$$

Let

$$\epsilon_n = -\left(\sum_{i=1}^n z_i\right)\log\left(\frac{n}{n+1}\right) + \frac{1}{n} - \frac{1}{2n^2} + R(3), \quad (5.26)$$

then it is obvious that

$$\lim_{n\to\infty} \epsilon_n = 0.$$

Thus by (5.24), (5.25) and (5.26),

$$\log \frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} \leq \log n - \log\sqrt{2\pi} + \epsilon_n\,.$$

■

Lemma 5.3.5 *If $\{Z_t : t \in \boldsymbol{N}\}$ is an independent identically distributed normal with fixed (known) variance process, then there exists a probability measure Q with the corresponding density q such that*

$$\log \frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} \leq \frac{1}{2}\log n - C + \epsilon_n\,,$$

where C is a constant and $\lim_{n\to\infty} \epsilon_n = 0$.

Proof :

Let

$$f(z,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\theta}{\sigma}\right)^2}, \qquad \theta \in \boldsymbol{R}, \quad z \in \boldsymbol{R} \quad \text{and} \quad \sigma > 0 \text{ known.}$$

Then the joint density fuction of $Z_1,\ldots,Z_n$ is

$$\begin{aligned} p_\theta(z_1,\ldots,z_n) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z_i-\theta}{\sigma}\right)^2} \\ &= (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(z_i-\theta)^2\right)} \\ &= (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left\{\left(\sum_{i=1}^{n} z_i^2\right)-2\theta\left(\sum_{i=1}^{n} z_i\right)+n\theta^2\right\}} \\ &= (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} z_i^2\right)} e^{\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} z_i\right)\theta-\frac{n}{2\sigma^2}\theta^2}\,. \quad (5.27) \end{aligned}$$

The equation (5.27) is maximum if

$$\theta = \frac{1}{n}\sum_{i=1}^{n} z_i\,,$$

hence the maximum likelihood estimate is given by

$$\begin{aligned} p_{ML}(z_1,\ldots,z_n) &= (\sigma\sqrt{2\pi})^{-n}\, e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)}\, e^{\frac{1}{n\sigma^2}\left(\sum_{i=1}^n z_i\right)^2 - \frac{1}{2n\sigma^2}\left(\sum_{i=1}^n z_i\right)^2} \\ &= (\sigma\sqrt{2\pi})^{-n}\, e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)}\, e^{\frac{1}{2n\sigma^2}\left(\sum_{i=1}^n z_i\right)^2}. \end{aligned}$$

Define probability measure Q through its joint density function q as follows,

$$q(z_1,\ldots,z_n) = \int_{-\infty}^{\infty} p_\theta(z_1,\ldots,z_n)\,\alpha(\theta)\,d\theta,$$

where

$$\alpha(\theta) = \frac{1}{\sqrt{\pi}}\, e^{-\theta^2}.$$

From [39], page 344, for

$$\int_{-\infty}^{\infty} e^{-px^2-qx}\,dx = \sqrt{\frac{\pi}{p}}\exp\left(\frac{q^2}{4p}\right), \qquad Re\,p > 0. \tag{5.28}$$

Using (5.28), it is obvious that

$$\int_{-\infty}^{\infty} \alpha(\theta)\,d\theta = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-\theta^2}\,d\theta = 1.$$

By (5.28)

$$\begin{aligned} q(z_1,\ldots,z_n) &= \int_{-\infty}^{\infty} (\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)} e^{\frac{1}{\sigma^2}\left(\sum_{i=1}^n z_i\right)\theta - \frac{n}{2\sigma^2}\theta^2} \frac{1}{\sqrt{\pi}} e^{-\theta^2}\,d\theta \\ &= \frac{1}{\sqrt{\pi}}(\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)} \int_{-\infty}^{\infty} e^{\frac{1}{\sigma^2}\left(\sum_{i=1}^n z_i\right)\theta - \left(\frac{n}{2\sigma^2}+1\right)\theta^2}\,d\theta \\ &= \frac{1}{\sqrt{\pi}}(\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)} \sqrt{\frac{\pi}{\frac{n}{2\sigma^2}+1}}\, \exp\left(\frac{\frac{1}{\sigma^4}\left(\sum_{i=1}^n z_i\right)^2}{4\left(\frac{n}{2\sigma^2}+1\right)}\right) \\ &= \frac{1}{\sqrt{\pi}}(\sigma\sqrt{2\pi})^{-n}\sqrt{\frac{2\sigma^2\pi}{n+2\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)} e^{\frac{1}{\sigma^2(2n+4\sigma^2)}\left(\sum_{i=1}^n z_i\right)^2}. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} &= \frac{(\sigma\sqrt{2\pi})^{-n} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)} e^{\frac{1}{2n\sigma^2}\left(\sum_{i=1}^n z_i\right)^2}}{\frac{1}{\sqrt{\pi}}(\sigma\sqrt{2\pi})^{-n}\sqrt{\frac{2\sigma^2\pi}{n+2\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n z_i^2\right)} e^{\frac{1}{\sigma^2(2n+4\sigma^2)}\left(\sum_{i=1}^n z_i\right)^2}} \\ &= \frac{\sqrt{n+2\sigma^2}}{\sqrt{2\sigma^2}}\exp\left\{\left(\frac{1}{2n\sigma^2} - \frac{1}{\sigma^2(2n+4\sigma^2)}\right)\left(\sum_{i=1}^n z_i\right)^2\right\} \end{aligned}$$

and

$$\log \frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} = \frac{1}{2}\log(n+2\sigma^2) - \frac{1}{2}\log(2\sigma^2) + \Big(\frac{1}{2n\sigma^2} - \frac{1}{\sigma^2(2n+4\sigma^2)}\Big)\Big(\sum_{i=1}^{n} z_i\Big)^2. \quad (5.29)$$

By Taylor's expansion,

$$\log(n+2\sigma^2) = \log n + \frac{2\sigma^2}{n} - \frac{4}{2!}\cdot\frac{\sigma^4}{n^2} + R(3), \quad (5.30)$$

where

$$R(3) = \frac{8\sigma^6}{3!}\cdot\frac{2}{\zeta^3}, \qquad \text{with } n \le \zeta \le n+2\sigma^2.$$

Let

$$\epsilon_n = \frac{1}{2}\Big(\frac{2\sigma^2}{n} - \frac{4}{2!}\cdot\frac{\sigma^4}{n^2} + R(3)\Big) + \Big(\frac{1}{2n\sigma^2} - \frac{1}{2n\sigma^2+4\sigma^4}\Big)\Big(\sum_{i=1}^{n} z_i\Big)^2, \quad (5.31)$$

then it is clear that

$$\lim_{n\to\infty} \epsilon_n = 0.$$

By (5.29), (5.30) and (5.31)

$$\log \frac{p_{ML}(z_1,\ldots,z_n)}{q(z_1,\ldots,z_n)} \le \frac{1}{2}\log n - \frac{1}{2}\log(2\sigma^2) + \epsilon_n.$$

So the proof is complete. ■

5.3.3 Application to hidden Markov models

Recall that under $\tilde{\phi} \in \tilde{\Phi}_K^c$, the Markov chain $X_1,\ldots,X_n$ has the joint density function

$$P_{\tilde{\phi}}(X_1 = x_1,\ldots,X_n = x_n) = P_{\tilde{\phi}}(x_1,\ldots,x_n) = \alpha_{x_1}^K \prod_{t=2}^{n} \alpha_{x_{t-1},x_t}(\tilde{\phi}). \quad (5.32)$$

Since the initial probability distribution $\alpha^K = (\alpha_i^K)$ is *fixed* for every $\tilde{\phi} \in \tilde{\Phi}_K^c$, then for optimization purpose, this probability can be ignored and hence (5.32) can be rewritten in the form

$$P_{\tilde{\phi}}(x_1, \ldots, x_n) = \prod_{i=1}^{K} \prod_{j=1}^{K} (\alpha_{ij}(\tilde{\phi}))^{n_{ij}}, \tag{5.33}$$

where n_{ij} is the number of times the pair (i, j) occurs in adjacent places in $x_1, \ldots, x_n$. Let n_i be the number of the occurences of i in $x_1, \ldots, x_{n-1}$. Notice that probability in (5.33) is maximized when

$$\alpha_{ij}(\tilde{\phi}) = \frac{n_{ij}}{n_i}, \qquad i, j = 1, \ldots, K.$$

Hence the maximum likelihood estimate in $\tilde{\Phi}_K^c$ is given by

$$P_{ML_K}(x_1, \ldots, x_n) = \prod_{i=1}^{K} \prod_{j=1}^{K} \left(\frac{n_{ij}}{n_i}\right)^{n_{ij}}.$$

A consequence of Csiszar lemma (Lemma 5.3.3) is the following lemma.

Lemma 5.3.6 *There exists a probability measure Q such that*

$$\log \frac{P_{ML_K}(x_1, \ldots, x_n)}{Q(x_1, \ldots, x_n)} \leq \frac{K(K-1)}{2} \log n + KC,$$

where C is a constant.

Proof :

Define a mixture density Q such that

$$Q(x_1, \ldots, x_n) = \prod_{i=1}^{K} \int \prod_{j=1}^{K} \left(\frac{n_{ij}}{n_i}\right)^{n_{ij}} \nu(n_{i\cdot})\, dn,$$

where

$$\nu(n_{ij}) = \frac{\Gamma(\frac{K}{2})}{(\Gamma(\frac{1}{2}))^K} \cdot \prod_{j=1}^{K} \left(\frac{n_{ij}}{n_i}\right)^{-\frac{1}{2}}.$$

Then by Csiszar lemma,

$$\begin{aligned}\log \frac{P_{ML_K}(x_1,\ldots,x_n)}{Q(x_1,\ldots,x_n)} &= \sum_{i=1}^{K} \log \frac{\prod_{j=1}^{K} (\frac{n_{ij}}{n_i})^{n_{ij}}}{\int \prod_{j=1}^{K} (\frac{n_{ij}}{n_i})^{n_{ij}} \nu(n_{i\cdot})\, dn} \\ &\leq \sum_{i=1}^{K} \Big\{ \frac{K-1}{2} \log n_i + C_i \Big\} \\ &\leq \frac{K(K-1)}{2} \log n + KC,\end{aligned}$$

where C_i are constants and $C = \max_{1\leq i\leq K} C_i$. ■

By definition of hidden Markov models, given a realization $\{x_t\}$ of the Markov chain $\{X_t\}$, the process $\{Y_t\}$ is a sequence of conditionally independent random variables. Recall that under $\tilde{\phi} \in \tilde{\Phi}_K^c$, Y_t given x_t has the conditional density $f(.,\theta_{x_t}(\tilde{\phi}))$. Hence the conditional density of $Y_1,\ldots,Y_n$ given $x_1,\ldots,x_n$ can be expressed as

$$\begin{aligned}p_{\tilde{\phi}}(y_1,\ldots,y_n|x_1,\ldots,x_n) &= \prod_{t=1}^{n} p_{\tilde{\phi}}(y_t|x_t) \\ &= \prod_{t=1}^{n} f(y_t,\theta_{x_t}(\tilde{\phi})) \\ &= \prod_{i=1}^{K} \prod_{t\in N_i} f(y_t,\theta_i(\tilde{\phi})),\end{aligned}$$

where

$$N_i = \{1 \leq t \leq n : X_t = i\}, \qquad i = 1,\ldots,K.$$

In this case, we would like the family of densities $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta^c\}$ having the *Csiszar property*, that is, there exists a probability measure Q with the corresponding density function q such that

$$\log \frac{\sup_{\theta\in\Theta^c} \prod_{t=1}^{n} f(y_t,\theta)}{q(y_1,\ldots,y_n)} \leq \log n + \text{Constant}.$$

Lemma 5.3.4 and Lemma 5.3.5 shows that the family of Normal distributions with fixed and known variance, and Poisson distributions have this property.

Let,

$$p_{ML_K}(y_1,\ldots,y_n|x_1,\ldots,x_n) = \sup_{\tilde{\phi}\in\tilde{\Phi}_K^c} p_{\tilde{\phi}}(y_1,\ldots,y_n|x_1,\ldots,x_n),$$

then we have next lemma which is similar to Lemma 5.3.6.

Lemma 5.3.7 *If the family of densities $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta^c\}$ has the Csiszar property, then there exists a probability measure Q with the corresponding density function q such that*

$$\log \frac{p_{ML_K}(y_1,\ldots,y_n|x_1,\ldots,x_n)}{q(y_1,\ldots,y_n|x_1,\ldots,x_n)} \leq K\log n + KC,$$

where C is a constant.

Proof :

Suppose that the family of densities $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta^c\}$ has the Csiszar property, then for every $i = 1,\ldots,K$, there is a probability density q_i such that

$$\log \frac{\sup_{\tilde{\phi}\in\tilde{\Phi}_K^c} \prod_{t\in N_i} f(y_t,\theta_i(\tilde{\phi}))}{q_i(y_1,\ldots,y_n)} \leq \log n_i + C_i, \tag{5.34}$$

where C_i, $i = 1,\ldots,K$ are constants. Define a probability measure Q through its density function q such that

$$q(y_1,\ldots,y_n|x_1,\ldots,x_n) = \prod_{i=1}^{K} q_i(y_1,\ldots,y_n).$$

Then by (5.34),

$$\begin{aligned}
\log \frac{p_{ML_K}(y_1,\ldots,y_n|x_1,\ldots,x_n)}{q(y_1,\ldots,y_n|x_1,\ldots,x_n)} &= \log \frac{\sup_{\tilde{\phi}\in\tilde{\Phi}_K^c} \prod_{i=1}^{K}\prod_{t\in N_i} f(y_t,\theta_i(\tilde{\phi}))}{\prod_{i=1}^{K} q_i(y_1,\ldots,y_n)} \\
&\leq \sum_{i=1}^{K} \log \frac{\sup_{\tilde{\phi}\in\tilde{\Phi}_K^c} \prod_{t\in N_i} f(y_t,\theta_i(\tilde{\phi}))}{q_i(y_1,\ldots,y_n)} \\
&\leq \sum_{i=1}^{K} \log n_i + C_i \\
&\leq K\log n + KC,
\end{aligned}$$

where $C = \max_{1 \le i \le K} C_i$. Hence, the theorem is proved. ■

Using Lemma 5.3.6 and Lemma 5.3.7 we prove a similar result for the observed process $\{Y_t\}$.

Lemma 5.3.8 *Assume that the family of densities $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta^c\}$ has the Csiszar property. Then there exists a probability measure $\tilde{Q}$ with corresponding density function $\tilde{q}$ such that*

$$\log \frac{p_{ML_K}(y_1, \ldots, y_n)}{\tilde{q}(y_1, \ldots, y_n)} \le \frac{K(K+1)}{2} \log n + KC$$

where

$$p_{ML_K}(y_1, \ldots, y_n) = \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} p_{\tilde{\phi}}(y_1, \ldots, y_n)$$

and C is a constant.

Proof :

$$\begin{aligned}
& p_{ML_K}(y_1, \ldots, y_n) \\
&= \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} p_{\tilde{\phi}}(y_1, \ldots, y_n). \\
&= \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} P_{\tilde{\phi}}(x_1, \ldots, x_n) \cdot p_{\tilde{\phi}}(y_1, \ldots, y_n | x_1, \ldots, x_n) \\
&\le \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} P_{\tilde{\phi}}(x_1, \ldots, x_n) \cdot \sup_{\tilde{\phi} \in \tilde{\Phi}_K^c} p_{\tilde{\phi}}(y_1, \ldots, y_n | x_1, \ldots, x_n) \\
&= \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} P_{ML_K}(x_1, \ldots, x_n) \cdot p_{ML_K}(y_1, \ldots, y_n | x_1, \ldots, x_n). \quad (5.35)
\end{aligned}$$

By lemma 5.3.6, there exists a probability measure Q_1 such that

$$P_{ML_K}(x_1, \ldots, x_n) \le Q_1(x_1, \ldots, x_n) n^{\frac{K(K-1)}{2}} e^{KC_1}, \quad (5.36)$$

where C_1 is a constant. Also from lemma 5.3.7, there exists a probability measure Q_2 with corresponding density q_2 such that

$$p_{ML_K}(y_1, \ldots, y_n | x_1, \ldots, x_n) \le q_2(y_1, \ldots, y_n | x_1, \ldots, x_n) n^K e^{KC_2}, \quad (5.37)$$

where C_2 is a constant.

Define a probability measure $\tilde{Q}$ through its density function $\tilde{q}$, where

$$\tilde{q}(y_1,\ldots,y_n) = \sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} Q_1(x_1,\ldots,x_n)\cdot q_2(y_1,\ldots,y_n|x_1,\ldots,x_n). \quad (5.38)$$

Then by (5.35), (5.36), (5.37) and (5.38),

$$\begin{aligned} p_{ML_K}(y_1,\ldots,y_n) &\leq \sum_{x_1=1}^{K}\cdots\sum_{x_n=1}^{K} Q(x_1,\ldots,x_n)\cdot q_2(y_1,\ldots,y_n|x_1,\ldots,x_n) \\ &\quad \times n^{\frac{K(K+1)}{2}} e^{K(C_1+C_2)} \\ &= \tilde{q}(y_1,\ldots,y_n)\, n^{\frac{K(K+1)}{2}} e^{K(C_1+C_2)}. \end{aligned}$$

Thus

$$\log\frac{p_{ML_K}(y_1,\ldots,y_n)}{\tilde{q}(y_1,\ldots,y_n)} \leq \frac{K(K+1)}{2}\log n + K(C_1+C_2).$$

■

5.3.4 Rate of growth of the maximized log-likelihood ratio

Based on the results of subsection 5.3.3, we obtain the rate of growth of the maximized log-likelihood ratio, in the following lemma. The idea of the proof of this lemma comes from [22].

Lemma 5.3.9 *Assume that the family of densities $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta^c\}$ has the Csiszar property, then for any $K \in \boldsymbol{N}$,*

$$\limsup_{n\to\infty} (\log n)^{-1} \log\frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)} \leq \frac{K(K+1)}{2} + 2$$

P_{ϕ^o}-almost sure.

Proof :

By Lemma 5.3.8, there exists a probability measure $\tilde{Q}$, with corresponding density function $\tilde{q}$ such that,

$$\log \frac{p_{ML_K}(y_1, \ldots, y_n)}{\tilde{q}(y_1, \ldots, y_n)} \leq \frac{K(K+1)}{2} \log n + KC,$$

where C is a constant. Hence

$$\limsup_{n \to \infty} (\log n)^{-1} \log \frac{p_{ML_K}(y_1, \ldots, y_n)}{\tilde{q}(y_1, \ldots, y_n)} \leq \frac{K(K+1)}{2}. \tag{5.39}$$

For every $n \in \boldsymbol{N}$, let

$$\begin{aligned} A_n &= \left\{ \{y_t\} \in \mathcal{Y}^\infty : (\log n)^{-1} \log \frac{\tilde{q}(y_1, \ldots, y_n)}{p_{\phi^\circ}(y_1, \ldots, y_n)} > 2 \right\} \\ &= \left\{ \{y_t\} \in \mathcal{Y}^\infty : \tilde{q}(y_1, \ldots, y_n) > n^2 \, p_{\phi^\circ}(y_1, \ldots, y_n) \right\}, \end{aligned}$$

then

$$\begin{aligned} P_{\phi^\circ}(A_n) &= \int_{A_n} p_{\phi^\circ}(y_1, \ldots, y_n) \, dy_1 \cdots dy_n \\ &< \int_{A_n} \frac{1}{n^2} \, \tilde{q}(y_1, \ldots, y_n) \, dy_1 \cdots dy_n \\ &\leq \frac{1}{n^2}. \end{aligned}$$

Thus,

$$\sum_{n=1}^{\infty} P_{\phi^\circ}(A_n) \leq \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

and hence by Borel-Cantelli lemma,

$$P_{\phi^\circ}\left(\limsup_{n \to \infty} A_n \right) = 0$$

implying

$$\limsup_{n \to \infty} (\log n)^{-1} \log \frac{\tilde{q}(y_1, \ldots, y_n)}{p_{\phi^\circ}(y_1, \ldots, y_n)} \leq 2, \tag{5.40}$$

P_{ϕ°-almost sure.

By (5.39) and (5.40),

$$\begin{aligned}
&\limsup_{n\to\infty} (\log n)^{-1} \log \frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)} \\
&\qquad \leq \limsup_{n\to\infty} (\log n)^{-1} \log \frac{p_{ML_K}(y_1,\ldots,y_n)}{\widetilde{q}(y_1,\ldots,y_n)} \\
&\qquad\quad + \limsup_{n\to\infty} (\log n)^{-1} \log \frac{\widetilde{q}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)} \\
&\qquad \leq \frac{K(K+1)}{2} + 2,
\end{aligned}$$

P_{ϕ^o}-almost sure. ■

As a direct consequence of Lemma 5.3.9, we have the next corollary.

Corollary 5.3.10 *Assume that the family of densities* $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta^c\}$ *has Csiszar property, then for any* $K \in \boldsymbol{N}$,

$$\log \frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)} = O_{a.s.}(\log n).$$

The following lemma is very simple, but later it will play an important role in the proof of Theorem 5.3.12.

Lemma 5.3.11 *For* $K \geq K^o$,

$$\liminf_{n\to\infty} (\log n)^{-1} \log \frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)} \geq 0,$$

with probability one under ϕ^o.

Proof :

From Corollary 4.5.8, for every $K \geq K^o$, there exists $\tilde{\phi} \in \tilde{\Phi}^c_K$, such that $\tilde{\phi} \simeq \tilde{\phi}^o$. Since $\{Y_t\}$ has the same law under $\tilde{\phi}^o$ and ϕ^o and

$$p_{ML_K}(y_1,\ldots,y_n) = \sup_{\tilde{\phi}\in\tilde{\Phi}^c_K} p_{\tilde{\phi}}(y_1,\ldots,y_n),$$

then it is clear that

$$p_{ML_K}(y_1, \ldots, y_n) \geq p_{\phi^o}(y_1, \ldots, y_n), \tag{5.41}$$

for any $n \in \boldsymbol{N}$. Then the conclusion of the lemma follows trivially. ■

5.3.5 Compensators avoiding over estimation

In this subsection, the sufficient conditions for the compensators avoiding over estimation are given. These conditions are similar to [22], which hold for hidden Markov model, in which the observed process takes values on a finite set.

Theorem 5.3.12 *Suppose that conditions A1, A2, A3, A4, A5', A6 and A7 hold and the family of densities $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta^c\}$ has Csiszar property. If the compensator is of the form*

$$\delta_n(K) = \varphi(n)h(K) \tag{5.42}$$

where φ satisfies

$$\liminf_{n\to\infty} \left(\frac{\log n}{n}\right)^{-1} \varphi(n) > 1 \tag{5.43}$$

and h satisfies

$$h(\widehat{K}) - h(K) \geq \frac{\widehat{K}(\widehat{K}+1)}{2} + 2, \qquad \text{for } \widehat{K} > K \geq 1, \tag{5.44}$$

then

$$\limsup_{n\to\infty} \widehat{K}(n) \leq K_o,$$

with probability one under ϕ^o.

Proof :

Suppose the compensator $\delta_n(K)$ satisfies the hypotheses of the lemma. Let K

be any positive integer such that $K > K_o$, then by Lemma 5.3.9 and Lemma 5.3.11,

$$\begin{aligned}
&\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\left(\frac{1}{n}\log\frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{ML_{K^o}}(y_1,\ldots,y_n)}\right)\\
&\quad=\limsup_{n\to\infty}(\log n)^{-1}\left(\log\frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)}-\log\frac{p_{ML_{K^o}}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)}\right)\\
&\quad\le\limsup_{n\to\infty}(\log n)^{-1}\log\frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)}\\
&\qquad-\liminf_{n\to\infty}(\log n)^{-1}\log\frac{p_{ML_{K^o}}(y_1,\ldots,y_n)}{p_{\phi^o}(y_1,\ldots,y_n)}\\
&\quad\le\left(\frac{K(K+1)}{2}+2\right)+0\\
&\quad=\frac{K(K+1)}{2}+2. \qquad (5.45)
\end{aligned}$$

By hypothesis (5.44) and (5.43),

$$\begin{aligned}
&\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\varphi(n)\big(h(K^o)-h(K)\big)\\
&\qquad\le\limsup_{n\to\infty}-\left(\frac{\log n}{n}\right)^{-1}\varphi(n)\left(\frac{K(K+1)}{2}+2\right)\\
&\qquad=-\left(\frac{K(K+1)}{2}+2\right)\cdot\liminf_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\varphi(n)\\
&\qquad<-\left(\frac{K(K+1)}{2}+2\right). \qquad (5.46)
\end{aligned}$$

Hence by hypothesis (5.42), (5.45) and (5.46),

$$\begin{aligned}
&\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\big(C_{K,n}(y)-C_{K^o,n}(y)\big)\\
&\quad=\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\big(L^*_{K,n}(y)-\varphi(n)h(K)-L^*_{K^o,n}(y)+\varphi(n)h(K^o)\big)\\
&\quad=\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\left(\frac{1}{n}\log\frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{ML_{K^o}}(y_1,\ldots,y_n)}+\varphi(n)\big(h(K^o)-h(K)\big)\right)\\
&\quad\le\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\left(\frac{1}{n}\log\frac{p_{ML_K}(y_1,\ldots,y_n)}{p_{ML_{K^o}}(y_1,\ldots,y_n)}\right)\\
&\qquad+\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\varphi(n)\big(h(K^o)-h(K)\big)
\end{aligned}$$

$$
\begin{aligned}
&< \left(\frac{K(K+1)}{2}+2\right) - \left(\frac{K(K+1)}{2}+2\right) \\
&= 0.
\end{aligned}
$$

So we have

$$\limsup_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\left(C_{K,n}(y) - C_{K^o,n}(y)\right) < 0, \qquad \text{for } K > K^o. \tag{5.47}$$

Suppose for $y \in \mathcal{Y}^\infty$, there is a subsequence $\widehat{K}_{n_i}(y)$ such that

$$\widehat{K}_{n_i}(y) \longrightarrow L, \qquad \text{as } i \to \infty, \tag{5.48}$$

where $L > K^o$. Since $\widehat{K}_n(y) \in \boldsymbol{N}$, for every n, then from (5.48), there is $M \in \boldsymbol{N}$, such that

$$\widehat{K}_{n_i}(y) = L, \qquad \forall i \geq M,$$

implying

$$
\begin{aligned}
&\limsup_{i\to\infty}\left(\frac{\log n_i}{n_i}\right)^{-1}\left(C_{\widehat{K}_{n_i}(y),n_i}(y) - C_{K^o,n_i}(y)\right) \\
&\qquad = \limsup_{i\to\infty}\left(\frac{\log n_i}{n_i}\right)^{-1}\left(C_{L,n_i}(y) - C_{K^o,n_i}(y)\right) \\
&\qquad < 0,
\end{aligned} \tag{5.49}
$$

by (5.47).

However, by definition of $\widehat{K}_n$,

$$C_{\widehat{K}_n(y),n}(y) - C_{K^o,n}(y) \geq 0, \qquad \text{for every } n \in \boldsymbol{N},$$

implying

$$\limsup_{i\to\infty}\left(\frac{\log n_i}{n_i}\right)^{-1}\left(C_{\widehat{K}_{n_i}(y),n_i}(y) - C_{K^o,n_i}(y)\right) \geq 0,$$

which contradicts with (5.49).

Therefore, every convergent subsequence of $\widehat{K}_n(y)$ must converge to a limit which is less or equal to K^o. Hence, it follows

$$\limsup_{n\to\infty}\widehat{K}_n(y) \leq K^o,$$

with probability one under ϕ^o. ■

5.4 Consistent Estimation of the Order

Finally, in this section, an example of a compensator which avoids both under estimation and over estimation are given. The idea of this compensator comes from [22].

Theorem 5.4.1 *Suppose that conditions A1, A2, A3, A4, A5', A6 and A7 hold and the family of densities* $\mathcal{F} = \{f(\cdot,\theta) : \theta \in \Theta^c\}$ *has Csiszar property. Then the compensator*

$$\delta_n(K) = \varphi(n)h(K),$$

where

$$\varphi(n) = 2\,\frac{\log n}{n}$$

and

$$h(K) = K^2(K+1)^2$$

produces a strongly consistent estimator $\widehat{K}_n$ *of* K^o.

Proof :

It is clear that, for each $n \in \boldsymbol{N}$,

$$\delta_n(K) \leq \delta_n(\widehat{K}), \qquad \text{for } K \leq \widehat{K}$$

and for every $K \in \boldsymbol{N}$,

$$\begin{aligned}\lim_{n\to\infty} \delta_n(K) &= \lim_{n\to\infty} 2K^2(K+1)^2\left(\frac{\log n}{n}\right)\\ &= 2K^2(K+1)^2\left(\lim_{n\to\infty}\frac{1}{n}\right)\\ &= 0.\end{aligned}$$

So by Theorem 5.2.1,

$$\liminf_{n\to\infty} \widehat{K}_n \geq K^o, \tag{5.50}$$

with probability one under ϕ^o.

It is obvious that

$$\begin{aligned}\liminf_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1}\varphi(n) &= \liminf_{n\to\infty}\left(\frac{\log n}{n}\right)^{-1} 2\left(\frac{\log n}{n}\right)\\ &= 2 > 1.\end{aligned}$$

Now we only have to show that for $\widehat{K} > K \geq 1$,

$$h(\widehat{K}) - h(K) \geq \frac{\widehat{K}(\widehat{K}+1)}{2} + 2$$

or

$$2\widehat{K}^2(\widehat{K}+1)^2 - 2K^2(K+1)^2 - \widehat{K}(\widehat{K}+1) \geq 4.$$

Let $\widehat{K} = K + k$, for some $k \geq 1$, then

$$\begin{aligned}&2\widehat{K}^2(\widehat{K}+1)^2 - 2K^2(K+1)^2 - \widehat{K}(\widehat{K}+1)\\ &\quad= 2(K+k)^2(K+k+1)^2 - 2K^2(K+1)^2 - (K+k)(K+k+1)\\ &\quad= 2(K^2+2kK+k^2)\left\{(K+1)^2 + 2k(K+1) + k^2\right\}\\ &\qquad -2K^2(K+1)^2 - (K+k)(K+k+1)\\ &\quad= 2K^2(K+1)^2 + 2K^2\left\{2k(K+1)+k^2\right\}\\ &\qquad +2(2kK+k^2)(K+1)^2 + 2(2kK+k^2)\left\{2k(K+1)+k^2\right\}\\ &\qquad -2K^2(K+1)^2 - (K+k)(K+k+1)\\ &\quad= 4K^2k(K+1) + 2K^2k^2\\ &\qquad +(2K+k)\left\{2k(K+1)^2 + 4k^2(K+1) + 2k^3\right\}\\ &\qquad -(K+k)(K+k+1)\\ &\quad\geq 4K^2k(K+1) + 2K^2k^2\\ &\quad\geq 4.\end{aligned}$$

So by Theorem 5.3.12,

$$\limsup_{n\to\infty} \widehat{K}_n \leq K^o, \tag{5.51}$$

with probability one under ϕ^o.

From (5.50) and (5.51),

$$\liminf_{n\to\infty} \widehat{K}_n = \limsup_{n\to\infty} \widehat{K}_n = K^o,$$

with probability one under ϕ^o, implying $\lim_{n\to\infty} \widehat{K}_n$ exists and

$$\lim_{n\to\infty} \widehat{K}_n = K^o,$$

with probability one under ϕ^o. ■

Bibliography

[1] M. Abramowitz and Stegun, editors. *Handbook of mathematical functions with formulas, graphs and mathematical tables*, volume 55 of *National bureau of standards applied mathematics series*. National bureau of standard applied mathematics, Washington DC, 1966.

[2] R. Azencott and D. Dacunha-Castelle. *Series of irregular observation : forecasting and model building*. Applied probability. Springer Verlag, New York, 1986.

[3] J.S Baras and L. Finesso. *Consistent estimation of the order of hidden Markov chains*, volume 184 of *lecture notes in control and inform. science*, pages 26–39. Springer Verlag, 1991. Stochastic theory and adaptive control.

[4] L.E. Baum and J.A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73:360–363, 1967.

[5] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

[6] L.E. Baum, T.A. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

[7] J.J. Benedetto. *Real Variable and Integration*. Teubner, Stuttgart, 1976.

[8] B.R. Bhat. *Modern Probability Theory.* Wiley Eastern Limited, New Delhi, second edition, 1988.

[9] P.J. Bickel and Y. Ritov. Inference in hidden markov models i: local asymptotic normality in the stationary case. *Bernoulli*, 2(3):199–228, 1996.

[10] P.J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum likelihood estimator for general hidden markov models. submitted to Ann. Stat, 1997.

[11] P. Billingsley. *Probability and Measure.* John Willey, New York, 1986.

[12] D. Blackwell and L. Koopmans. On the identifiability problems for functions of finite markov chains. *Ann. Math. Stat.*, 28:1011–1015, 1957.

[13] I. Csiszar. Information theoretic methods in statistics. Notes for course ENEE 728F, Spring 1990.

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc.*, 39:1–38, 1977. Ser B.

[15] R. Durrett. *Probability : Theory and Examples.* Duxbury Press, Belmont, second edition, 1995.

[16] R.E. Edwards. *Fourier Series.* Springer Verlag, New York, second edition, 1979.

[17] M. Eisenberg. *Topology.* Holt, Rinehart and Winston, New York, 1974.

[18] R.J. Elliott, L. Aggoun, and J.B. Moore. *Hidden markov models : estimation and control*, volume 29 of *Application of Mathematics.* Springer Verlag, New York, 1993.

[19] R.J. Elliott and W.C. Hunter. Financial signal processing. Seminar paper, Department of Applied Mathematics, University of Adelaide, 1995.

[20] B.S. Everitt and D.J. Hand. *Finite mixture distributions.* Chapman and Hall, London, 1981.

[21] O.J. Farrel and B. Ross. *Solved problem in analysis.* Dover, New York, 1971.

[22] L. Finesso. *Consistent estimation of the order for Markov and hidden Markov chains.* PhD thesis, Graduate school of the university Maryland, 1990.

[23] H. Fustenberg and H. Kesten. Products of random matrices. *Ann. Math. Statist*, 31:457–469, 1960.

[24] F.R. Gantmacher. *Applications of the Theory of Matrices.* Interscience, New York, 1959.

[25] R. Goodman. *Introduction to stochastic models.* Benjamin/Cummings, California, 1988.

[26] J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[27] J.D. Hamilton. Analysis of time series subject to change in regime. *Journal of Econometrics*, 45:39–70, 1990.

[28] J.D. Hamilton. Estimation, inference and forecasting of time series subject to changes in regime. In G.S. Maddala, C.R. Rao, and H.D. Vinod, editors, *Handbook of statistics*, volume 11, pages 231–260. Elsevier Science B.V., 1993.

[29] J.D. Hamilton. State space models. In R.F. Engle and D.L. Mc Fadden, editors, *Handbook of Econometrics*, volume IV, pages 3041–3080. Elsevier Science B.V., 1994.

[30] J.D. Hamilton. *Time Series Analysis.* Princeton University Press, Princeton, New Jersey, 1994.

[31] S. Karlin and H.M. Taylor. *A First Course in Stochastic Processes.* Academic Press, New York, 1975.

[32] J.G. Kemeny and J.L. Snell. *Finite Markov Chains.* Springer Verlag, New York, 1976.

[33] J.F.C. Kingman. *Subadditive process*, volume 539 of *lecture notes in Math.*, pages 167–223. Springer Verlag, 1976. Ecole d'Ete Probabilites de Saint Flour V.

[34] B.G. Leroux. Maximum likelihood estimation for hidden markov models. *Stochastic processes and their applications*, 40:127–143, 1992.

[35] B.G. Leroux and M.L. Putterman. Maximum-penalized-likelihood estimation for independent and markov dependent mixture models. *Biometrics*, 48:545–558, 1992.

[36] T.A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of Royal Society B*, 44(2):226–233, 1982.

[37] G.J. Mc. Lachlan and K.E. Basford. *Mixture models.* Marcel dekker, New York, 1988.

[38] T. Petrie. Probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 40:97–115, 1969.

[39] A.P. Prudnikov, Yu.A. Brychkov, and O.I. Marichev. *Integrals and Series.* Gordon and Breach science, New York, volume 1: elementary functions edition, 1986.

[40] L.R. Rabiner. A tutorial on a hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[41] L.R. Rabiner and B.H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–16, 1986.

[42] L.R. Rabiner, B.H. Juang, S.E. Levinson, and M.M. Sondhi. Recognition of isolated digits using hidden markov models with continuous mixture densities. *AT & T Technical Journal*, 64(6):1211–1234, 1985.

[43] L.R. Rabiner, B.H. Juang, S.E. Levinson, and M.M. Sondhi. Some properties of continuous hidden markov model representations. *AT & T Technical Journal*, 64(6):1251–1270, 1985.

[44] H.L. Royden. *Real Analysis.* Mac Millan, New York, third edition, 1988.

[45] T. Ryden. Consistent and assimptotically normal parameter estimates for hidden markov models. *The Annals of Statistics*, 22:1884–1895, 1994.

[46] T. Ryden. Estimating the order of hidden markov models. *Statistics*, 26:345–354, 1995.

[47] T. Ryden. On recursive estimation for hidden markov models. *Stochastic Processes and their Applications*, 66:79–96, 1996.

[48] A.N. Shiryayev. *Probability.* Springer-Verlag, New York, 1984.

[49] K.R. Stromberg. *An introduction to classical real analysis.* Wadsworth, Belmont, California, 1981.

[50] H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 34:1265–1269, 1963.

[51] H. Teicher. Identifiability of mixtures of product measures. *Ann. Math. Statist.*, 38:1300–1302, 1967.

[52] D.M. Titterington, A.F.M. Smith, and V.E. Makov. *Statistical analysis of finite mixture distributions.* John Wiley, New York, 1985.

[53] Peter Walters. *An Introduction to Ergodic Theory.* Springer Verlag, New York, 1982.

[54] S.J. Yakowitz and J.D. Spragins. On the identifiability of finite mixtures. *Ann. Math. Stat.*, 39(1):209–214, 1968.

Errata

Explanation: p/n means page p, line n from top . The first word/phrase is to be replaced by the second following the colon (:) but sometimes a simple addition is indicated. Mathematical typos are in **bold face**. The errors were mainly of a grammatical nature which was due in part to the fact that English was the second language of the author. This list has been presented in a compact form and lists the errors pointed out by one of the examiners.

vii/13 patient : patience 1/5 with : where 1/8 hidden : it is hidden 1/12 widespread : widely used 2/6 model : models 2/16 fact to be :fact, noticed in.. 3/21 in size : in the size 4/2 on : in 4/3 and also the last : (omit) 4/7 inspiring : inspired 4/7 is dedicated to solve : investigates 4/10 in [46], we will: those of [46], we shall 4/13 contains .. aim : reviews literature and gives the aims 5/8 for completeness and : (omit) 5/13 So the : The 5/14 can be : are 6/3 Such : Such a 6/4 a true: the true 6/5 this true parameter : these true parameters 7/4 with (2.2) : to (2.2) 7/7 then α_{ij} : Then the α_{ij} 7/14 Thus A : That is, A 11/14 the block : block 16/5 A pair : The pair 18/1 we will : we shall 18/15 expressed as : written 20/9 sometimes : sometimes the sequences 26/2 then : (omit) 26/3 By knowing : Knowing 26/4 parameters: the parameters 26/5 then: (omit) 26/9 $i = i, .. : i = 1, ..$ 27/2-5 $E[M(y)] = A$ 28/10 $\pi A = A : \pi A = \pi$ 35/10 Next Lemma : The next Lemma 35/13 are : are a 35/13 and : and a 36/14 (see 35/13) 42/12 that : that the 43/1 such: such a 43/2 parameter : parameter set 43/5 parameter : parameter set 43/8 parameter : parameter set 43/12 parameter : parameter set 43/16 As a straight : As a 43/20 true : a true 43/20 parameter : parameter set 44/1 parameter : parameter set 44/7 contradicting with : contradicting 44/8 parameter : parameter set 44/18 since : .Since 44/20 contradicting with : contradicting 44/20 it must be : (omit) 45/1 parameter : parameter set 45/10 important being ergodicity. Ergodicity is essential for the limit 48/11 The Kolmogorov consistency theorem .. gives the existence 48/15 now we have : we have 49/9 This is the Borel ..51/5 equivalent with : equivalent to 54/2 parameter : parameter set 54/22 in the implication form : (omit) 58/15 of a finite mixture 60/14 contradicting with : contradicting 60/19 N. Then $\hat{N} \geq N$. 61/5 To prove the Lemma it is sufficient to show 61/19 such that : such that each 64/6 define it Laplace transform 67/8 results of : results on 67/17 in one: in the one 68/23 bases : basis 69/3 of bases : in a basis 69/5 space.The 69/9 is linearly independent 70/6 There exists 70/19 Otherwise, suppose 72/19 theorem, mathematical 72/21 for some n.72/22 that then the class 73/17 The dominated convergence theorem applied to (3.22), ensures 74/6 by the monotone convergence theorem, (3.19) then 74/10 (3.25) implies 74/14 hypothesis, the calss of 74/15 implies 74/19 Therefore, the class 74/20 mixtures, we have 75/5 results hold, when .. are replaced by... 76/7 omit: to allow the above possibilities 76/15 In the expression 77/2 We shall 77/14 then by part : by part 77/19 We shall 78/14 written as : written 78/21 This is equivalent with 79/1 then by : by 79/15 The sufficiency is obvious. We shall 80/3 parameter : parameter set 80/4 parameter : parameter set 80/5 Then 80/9The next lemma an example of such a parameter set that can be a true parameter set 80/15 Then the size .. that is, there is no 80/21 contradicting the fact 81/1 which are equivalent to the true parameter set. 81/4 parameter : parameter set 81/21 Since the 82/5 parameter : parameter set 82/9 the parameters 85/16 the parameters 86/1 However, by 86/7 In the case 86/12 the parameters 87/12 identified as having the form 91/10 This gives 96/3 contradicting the 97/1 However, for $i = 1, \ldots, K_2$, $\hat{\pi}_i > 0$, so 97/3 From (3.97) 97/8 If there exists an 98/1 contradicting the fact .. is minimal 99/1 the parameters .. take the forms: 99/18 the parameters .. take the forms: 100/7 The ... contradicting the fact 100/15 the parameters 101/13 then we have: we have 101/17 focus on the 102/2 contradiction, we must have for 103/9 and is impossible 103/11 to obtain 103/12 is to be modelled 103/14 some stochastic 104/19 same as : equal to 104/20 So in this : In this 105/1 and Φ_k will be this class 107/1 the Kullback- 107/3 we shall 107/7 In section 107/17 the comparing 107/20 we shall 108/2 Therefore, our 108/3 giving a topology on the 108/4 We also give some 108/7 from the general to the hidden

110/6 compactification of 110/8 dense in .. Define the norm 117/6 We shall 118/1 then by: by 118/19 The equation : Equation 119/9 Similarly, exchanging 119/15 Such a process 119/19 So the: The 119/20 hold,then: hold. Then 120/18 almost surely 121/4 almost surely 121/16 almost surely 123/21 having an 124/4 having an 126/1 The Levy 126/5 theorem for : theorem 127/14 the Caesaro .. the stationary 128/7 by the 128/12 we shall 129/12 by the definition 129/14 By the 134/17 Kingman's 135/9 which : This 136/15 which : This 137/15 we shall 141/11 parameters 142/15 we shall 143/3 such a 143/11 space on 150/7 using a 150/13 By the Helly .. for the sequence 151/1 We shall 151/5 We shall 155/1 We shall 155/10 Therefore, under 155/13 taking the limit 155/14 values of 155/20 to equal 156/1 This is proved in 156/2 we shall 156/7 we shall 158/20 as is shown in the following lemma. 159/10 we shall 159/19 we shall 168/2 parameters 168/7 attained for 169/17 we shall 170/9 Given : Give 170/10 compact, then : compact, 170/19 is the closure of 171/22 The theorem 173/9 to be modelled :(omit) 173/13 task now is : task is 173/14 As parametric .., we shall 173/20 have joint 175/11 models 176/2 obtain sufficient 179/8 Using the 179/13 introduce the 181/9 calculating 181/10 respectively, we have 184/19 assigned previously. The recursive step is always possible. There will be at least..185/4 For proof of .., we use Stirling's 186/5 Extension of the 186/6 we extend 186/10 density function of .. is 187/10 is a maximum 188/13 Expanding 189/6 variance process : varaiance 189/6 and corresponding density 189/18 is a maximum 192/7 that the probability 192/11 of the Csiszar 193/1 by the 193/6 By the 193/19 show 194/11 property. Then 196/15 propert. Then 196/17 almost surely 197/18 This implies 197/20 almost surely 198/6 almost surely 198/9 has the Csiszar property. Then 199/6 for a hidden 199/9 has the Csiszar 201/15 by the definition 202/3 estimation is given 202/6 has the Csiszar 205/21 Variables