THESIS REPORT
Master's Degree

SYSTEMS
RESEARCH
CENTER

*Supported by the
National Science Foundation
Engineering Research Center
Program (NSFD CD 8803012),
Industry and the University*

Order Determination for Probabilistic Functions of Finite Markov Chains

by: L. Finesso
Advisor: J. Baras

M.S. 87-4
Formerly TR 87-160

ABSTRACT

Title of Thesis: Order Determination for Probabilistic Functions of Finite Markov Chains

Lorenzo Finesso, Master of Science, 1986

Thesis directed by: Dr. John Baras, Professor, Department of Electrical Engineering.

Let $\{Y_t\}$ be a stationary stochastic process with values in the finite set $\mathbb{Y}$.

We model $\{Y_t\}$ as a probabilistic function of a finite state Markov Chain $\{X_t\}$ i.e. X_t is such that:

$$P[\ Y_t \mid X^t,\ Y^{t-1}\ ] = P[\ Y_t \mid X_t\ ]$$

Define the cardinality of the state space of $\{X_t\}$ as the *order* of the model. The problem is to determine the order given the observations $\{y_1, y_2, \ldots, y_T\}$. We show that under mild conditions on the probability distribution function $P_Y(.)$ of $\{Y_t\}$ the order is identifiable and can be consistently determined from the data.

ORDER DETERMINATION FOR PROBABILISTIC FUNCTIONS OF FINITE MARKOV CHAINS

by

Lorenzo Finesso

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Master of Science
(Fall 1986)

INTRODUCTION

In 1966 Ulf Grenander posed, and partially answered, in a little-known paper, the following question: "can we look inside an unreliable automaton?".

Grenander's paper is one of the few devoted to the identification of finite state stochastic systems. Many people in the '60 contributed to the creation of a finite state stochastic system theory and a considerable corpus of knowledge is presented in e.g. [Paz] and [Carlyle].

Despite this fact very little work has been done in identification. Our aim here is to investigate for a restricted class of such models (probabilistic functions of Markov Chains) a particular aspect of the identification problem (order determination).

Roughly the problem can be described as follows. We observe a piece of a trajectory of a finite valued, stationary stochastic process $\{Y_t\}$ and try to model it as a probabilistic function of a finite Markov Chain. The underlying assumption is that the mechanism generating $\{Y_t\}$ can be reasonably well explained in the following way. There is an (unobserved) finite stationary Markov Chain $\{X_t\}$ such that:

$$P[X_{t+1} , Y_{t+1} | X^t , Y^t] = P[X_{t+1} , Y_{t+1} | X_t] \qquad (1)$$

It follows that the evolution of $\{Y_t\}$ can be probalistically characterized by the transition matrix of $\{X_t\}$ and the conditional probability distributions $P[Y_t | X_t]$.
Without loss of generality we can assume that the Markov Chain $\{X_t\}$ is such that $P[X_{t+1}, Y_{t+1} | X_t] = P[Y_{t+1} | X_{t+1}] P[X_{t+1} | X_t]$ it then follows from (1) that $\forall\ t_1 < t_2 < \ldots < t_k$

$$P[Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}] = \Sigma P[Y_{t_k} | X_{t_k}] P[X_{t_k} | X_{t_{k-1}}] \ldots P[Y_{t_1} | X_{t_1}] P[X_{t_1}]$$

To identify this model on the basis of the data $\{y_1, y_2, \ldots$ $\ldots, y_T\}$ means therefore:

- determine the cardinality of the state space of $\{X_t\}$ (which we call the *order* of the model)

- determine the parameters i.e. the transition matrix of $\{X_t\}$ and the conditional distributions $P[Y_t|X_t]$.

Work has been done in the past in the area of parameters estimation for these models (when the order is supposed known a priori).
[Baum] and [Petrie] in a series of papers analysed the asymptotic distribution of maximum likelihood estimators showing consistency. Moreover they produced an algorithm for the efficient computation of the maximum likelihood estimators.

In recent years people working in automatic speech recognition have been using these models (called Hidden Markov Models in their jargon) producing many studies on the numerical aspects of Baum algorithm. For a review see the paperby [Levinson et al.]

To the best of our knowledge nothing has been done for the problem of order determination. A first basic question is the following.

Suppose that $\{Y_t\}$ is actually generated by a probabilistic function of Markov Chain of order n (such an assumption is probably only seldom met in practice). Is it possible, on the basis of the data to determine n ?

Our results show that, if the probability distribution function (pdf) $P_Y(.)$ of the process $\{Y_t\}$ belongs to a large subclass of the class of pdf generated as functions of Markov Chains then n is identifiable. Moreover it can be estimated in a consistent way from the data.

THE MODEL

In this chapter we will define the class of models chosen to represent the observed finite valued process $\{Y_t\}$.

Notations

Let $\{Z_t\}$ be a stationary finite valued process (FVP) defined on the probability space (Ω,F,P).

We denote by $\mathbb{Z}:=\{\zeta_1,\zeta_2,..,\zeta_r\}$ the set of values of Z_t, and $\mathbb{Z}^{\circledast}$ the free monoid generated by $\mathbb{Z}$.

For $s\in\mathbb{Z}^{\circledast}$ let $|s|:=\text{length}(s)$. Define the function

$P : \mathbb{Z}^{\circledast} \to [0,1]$ as follows:

$P(\pm) = 1$

if $s=\varepsilon_1\varepsilon_2\varepsilon_3\cdots\varepsilon_k$ (where $\varepsilon_i\in\mathbb{Z}$) then

$P(s) = P[Z_t=\varepsilon_1, Z_{t+1}=\varepsilon_2, \ldots Z_{t+k}=\varepsilon_k]$

Since $\{Z_t\}$ is supposed stationary $P(s)$ does not depend on t.

The function $P(.)$ is called the probability distribution function (pdf) of $\{Z_t\}$.

Finite Stochastic Systems

For reasons to be shortly explained, finite stochastic systems are a natural class of models for FVP.

Definition σ.1 (SFSS)

A pair $(\{X_t\},\{Y_t\})$ of stochastic processes defined on some probability space (Ω,F,P) is said to be a stationary finite stochastic system (SFSS) if the following conditions are met:

i) $\{X_t\}$ the state process, is a FVPwith value in $\mathbb{X}=\{1,2,..,k\}$

The cardinality k of $\mathbb{X}$ is called the order of the system *ii)* $\{Y_t\}$ *the output process, is a* FVP *with value in* $\mathbb{Y}=\{1,2,..r\}$

iii) $(\{X_t\},\{Y_t\})$ *are jointly stationary*

iv) $P[Y_{t+1},X_{t+1}|X^t,Y^t] = P[Y_{t+1},X_{t+1}|X_t]$

This definition was introduced in [Picci]. An important consequence of $\alpha.1$ is the following:

<u>Lemma</u> $\alpha.2$

The state process $\{X_t\}$ *of a* SFSS *is a Markov Chain.*

<u>proof</u>:

From iv) in the definition we have:

$$P[X_{t+1}=j|X^t,Y^t] = P[X_{t+1}=j|X_t]$$

Taking conditional expectations with respect to X^t we have:

$$P[X_{t+1}=j|X^t] = P[X_{t+1}=j|X_t]$$

which is the desired Markov property. □

Also it follows immediately from the definition that a SFSS is completely specified by the set of matrices $\{M(\varepsilon),\varepsilon\in\mathbb{Y}\}$ where:

$$[M(\varepsilon)]_{i,j} = P[Y_{t+1}=\varepsilon,X_{t+1}=j|X_t=i] \quad i,j=1,2,..,k$$

and by the initial (invariant) probability distribution π for the Markov Chain $\{X_t\}$.

This means that all probabilities of the form $P[Y_1^t=s,X_0^t=w]$ with $s\in\mathbb{Y}^{\circledast}$, $w\in\mathbb{X}^{\circledast}$ can be computed explicitely in terms of $\{M(\varepsilon),\ \varepsilon\in\mathbb{Y}\}$ and π

In particular, for future reference, we observe that:

<u>Lemma</u> $\alpha.3$

The probability distribution function of the output process $\{Y_t\}$ *of a* SFSS *is given by:*

$$P(\varepsilon_1\varepsilon_2\varepsilon_3...\varepsilon_t) = \pi M(\varepsilon_1)M(\varepsilon_2)...M(\varepsilon_t)e$$

where $e=(1,1,..,1)\in\mathbb{R}^k$.

The Realization Problem

How can SFSS be used to model a FVP $\{Y_t\}$?
The answer depends crucially on what the available data about $\{Y_t\}$ are.

The ideal situation is when $\{Y_t\}$ is probabilistically completely known. We are given the probability space (Ω,F,P) and the maps Y_t : $(\Omega,F,P) \to \mathbb{Y}$ defining $\{Y_t\}$.

When this is the case, we can aim at perfect modelling and try to find a SFSS $(\{X_t\},\{Y'_t\})$ defined on the same $(\Omega F,P)$ and such that $Y'_t = Y_t$ $(\forall t)$ a.e.P.

This is the still open problem og strong stochastic realization for FVP.

Another interesting situation is when $\{Y_t\}$ is statistically completely known. We are given the pdf

$p : \mathbb{Y}^{\circledast} \to [0,1]$

of the FVP to model and the problem is to find a SFSS $(\{X_t\},\{Y_t\})$ such that $P(s) = p(s)$ $(\forall s \in \mathbb{Y}^{\circledast})$, where $P(.)$ is as in Lemma α.3

This is the weak stochastic realization problem for FVP.
The status of (weak) realization theory for FVP is rather unsatisfactory. The main result here is the *characterization* of the class of pdf of FVP that admit weak realizations of finite order, see [Heller], [Picci].

On the other hand the characterization of minimal realizations (i.e. of minimum order), the relations between them and how to construct a realization are still misterious aspects of the theory, see [Picci, van Schuppen].

We want to briefly explain why it is desirable to build a Realization Theory for FVP.

Observe that once $\{Y_t\}$ has been realized through a SFSS, it can be thought as a probabilistic function of the Markov Chain $\{X_t\}$ i.e. the present value Y_t is a probabilistic function of the preceding state X_{t-1} only and *not* of all the past history Y^t, from iv):

$$P[Y_t | X^{t-1}, Y^{t-1}] = P[Y_t | X_{t-1}]$$

Statisticians would say that X_t is a sufficient statistic for Y_t. A stochastic to deterministic transformation [Petrie] will make this point clearer.

Define the new state space $\mathbb{S} = \mathbb{X} \times \mathbb{Y}$ and the function

$$f : \mathbb{S} \to \mathbb{Y} \quad \text{as} \quad f(i,\varepsilon) = \varepsilon$$

The process $\{X'_t\} = \{(X_t, Y_t)\}$ is then Markov (follows from iv) and the process

$$Y'_t = f(X'_t)$$

has the same pdf of $\{Y_t\}$.

Realizable processes can therefore be interpreted as deterministic functions of Markov Chains (at least in the weak sense).

In Identification the first step to take is the choice of a class of models from which one is to be selected to represent the data. It is at this stage that Realization Theory of FVP (if there were one) would play its role giving the rationale for the choice of the class.

For a discussion of the relevance of Stochastic Realization Theory in Identification (in the contest of linear systems)

see [Kalman] and [Finesso,Picci].
Unfortunately, the rather poor development of the theory for FVP will force us to approach the Identification problem from the classical point of view.

In the classical approach the <u>model</u> of $\{Y_t\}$ is just its unknown pdf $P_Y(.)$ which must be inferred on the basis of the available data $\{y_1, y_2, \ldots, y_T\}$. The problem is one of statistical inference with non-independent samples.

To make the inference problem tractable a parametric family $\{P_\theta; \theta \in \Theta\}$ of pdf is chosen a priori from which one is to be selected.

Observe that any function:

$p : \mathbb{Y}^{\circledast} \rightarrow [0,1]$

satisfying:

i) $p(\emptyset) = 1$

ii) $p(s) = \sum_\varepsilon p(s\varepsilon) \qquad (\forall s \in \mathbb{Y}^{\circledast})$

iii) $p(s) = \sum_\varepsilon p(\varepsilon s) \qquad (\forall s \in \mathbb{Y}^{\circledast})$

is the pdf of a stationary FVP (Kolmogorov Theorem).
A useful by-product of the formulation of the realization problem is a convenient parametric family of pdf for FVP. We have in fact the following:

<u>Proposition</u> c.4

Let $k \in \mathbb{N}$ *be given, toghether with a set of* $r = \# \mathbb{Y}$ *matrices* $M(\varepsilon_1), M(\varepsilon_2), \ldots, M(\varepsilon_r)$ *in* $\mathbb{R}_+^{k \times k}$ *and a stochastic vector* $\pi \in \mathbb{R}^k$ *If* $A = \sum_\varepsilon M(\varepsilon)$ *is a stochastic matrix and* $\pi = \pi A$ *then the function* $P : \mathbb{Y}^{\circledast} \rightarrow [0,1]$ *defined as*

$P(\varepsilon_1 \varepsilon_2 \ldots \varepsilon_t) = \pi M(\varepsilon_1) M(\varepsilon_2) \ldots M(\varepsilon_t) e$

is a pdf of a stationary FVP *with value in* $\mathbb{Y}$.

proof :

direct check of the consistency conditions given before.

An assumption on the structure of the matrices $M(\cdot)$ will give us a more economical parametrization.

Assumption A.1

$\forall \varepsilon \in Y$ *there exists* $B_\varepsilon = \mathrm{diag}\{b_{1\varepsilon}, b_{2\varepsilon}, \ldots b_{k\varepsilon}\}$ *where* $b_{j\varepsilon} \in [0,1]$ $(\forall j, \varepsilon)$ *and* $\sum_\varepsilon b_{j\varepsilon} = 1$ $(\forall j)$ *such that* $M(\varepsilon) = AB_\varepsilon$

Assumption A.1 looses its misterious aspect when interpreted in the contest of SFSS. In the language of SFSS, A.1 corresponds to the factorization hypothesis:

$$P[Y_{t+1}=\varepsilon, X_{t+1}=j \mid X_t=i] = P[Y_{t+1}=\varepsilon \mid X_{t+1}=j]P[X_{t+1}=j \mid X_t=i]$$

Introducing the following notation:

$$b_{j\varepsilon} = P[Y_{t+1}=\varepsilon \mid X_{t+1}=j] \qquad B = \|b_{j\varepsilon}\|_{\substack{j=1,..,k\\ \varepsilon=1,..,r}}$$

$$a_{ij} = P[X_{t+1}=j \mid X_t=i] \qquad A = \|a_{ij}\|_{i,j=1,..,k}$$

then the factorization equation becomes:

$$M(\varepsilon) = AB_\varepsilon$$

In Stochastic automata literature a system satisfying this condition is called a Moore machine [Paz].

Given a general SFSS it is always possible to convert it to an equivalent Moore machine [Carlyle], there is therefore no loss of generality in making assmption A.1.

We are now in position to define the parametric family of pdf chosen to represent FVP.

Definition $\alpha.5$

$$\mathbb{P} = \{ P_{\theta} : \mathbb{Y}^{\circledast} \to [0,1] \;;\; \theta \in \Theta \}$$

where

$$\Theta = \{ \; A \in \mathbb{R}^{k\times k}, \; B \in \mathbb{R}^{k\times r}, \; \pi \in \mathbb{R}^{k}, \; k \in \mathbb{N} \quad \text{and}$$

A is a stochastic matrix

B is a stochastic matrix $\quad (\; B_{\varepsilon} = \mathrm{diag}\{ b_{1\varepsilon}, \ldots, b_{k\varepsilon} \})$

π is a stochastic vector and $\pi = \pi A \quad \}$

and

$$P_{\theta} \; : \; \varepsilon_1 \varepsilon_2 \; \ldots \; \varepsilon_t \to \pi A B_{\varepsilon_1} A B_{\varepsilon_2} \; \ldots A B_{\varepsilon_t} e$$

We will denote the generic element of Θ as $\theta = (k, A, B, \pi)$.

Taking $\mathbb{P}$ as parametric family of pdf we limit ourselves to the consideration of FVP that are functions of finite Markov Chains (FMC).

IDENTIFIABILITY

In this chapter we define a subfamily $\overline{\mathbb{P}} \subset \mathbb{P}$ for which we can prove an identifiability result.

The Petrie Family

Suppose to have available a sample path of the FVP $\{Y_t\}$ whose pdf $P_Y(.) \in \mathbb{P}$.

Applying some standard estimation procedure (e.g. max likelihood) we would like to determine "the" value $\theta_0 \in \Theta$ such that $P_{\theta_0} \sim P_Y$. Unfortunately this is in general impossible.

Definition β.1

Two parameters $\theta_1, \theta_2 \in \Theta$ *are called indistinguishable if* $\theta_1 \neq \theta_2$ *and* $P_{\theta_1} \sim P_{\theta_2}$

Clearly it is impossible on the basis of the data to discriminate between indistinguishable parameters.

It is possible to construct examples that show that Θ contains indistinguishable pairs. The following assumption will restrict enough Θ to eliminate the problem.

Assumption A.2

A *and* B *as defined in a.5 are such that*

i) $a_{ij} > 0 \quad (\forall i,j)$

ii) there exists $r \in \mathbb{Y}$ such that B_r has distinct diagonal elements.

Actually it is A.2 plus regularity (see below) that will give us an identifiable family.

Comments on A.2

i) can probably be relaxed to A irreducible aperiodic of which it is a special case.

ii) plays a crucial role in the proof of the identifiability result (theorem B.10), we will thus spend some words on its interpretation.

Since $b_{j\varepsilon} = P[Y_t=\varepsilon |\ X_t=j]$ it follows that $Y_t \perp\!\!\!\perp X_t$ if and only if the matrix B has all its rows equal.

Clearly the situation in which $Y_t \perp\!\!\!\perp X_t$ is meaningless from the modelling point of view, since it corresponds to no modelling at all: $P[Y_t=\varepsilon\ |X_t=j] = P[Y_t=\varepsilon]\quad (\forall j,\varepsilon)$

i.e. X_t has no influence on the process Y_t.

The weakest assumption to impose on the matrix B to rule out the case $Y_t \perp\!\!\!\perp X_t$ is that at least two rows of B are distinct (as vectors in $\mathbb{R}^r$) or equivalently that at least one column of B is not a positive multiple of the vector e.

Assumption ii) is stronger than that and hence we have a (partial) probabilistic interpretation for it.

What does iv) add to $Y_t \not\!\perp\!\!\!\perp X_t$?

We think that the right interpretation is systemistic.

Assumption iv) is the weakest assumption on the matrix B under which the map associating states $j \in \mathbb{X}$ to transition probabilities $p[Y_t|X_t=j]$ is injective.

We will not go into further details here, but this is likely to be related to some notion of observability for the models under consideration.

Definition β.2

$$\mathbb{P}' = \{ P_\theta \in \mathbb{P} \; ; \; \theta \in \Theta' \}$$

where

$$\Theta' = \{ \theta \in \Theta ; \; A, \; B \text{ satisfy } A.2 \}$$

We call $\mathbb{P}'$ the Petrie family.

Notice that if $\theta = (k, A, B, \pi) \in \Theta'$ then π is uniquely determined (in virtue of A.2i). We can therefore drop π from the list of parameters when working in Θ'.

The Regular Family

Definition β.3

Let $P_\theta \in \mathbb{P}$. *Define the set of compound sequence matrices of* P_θ *relative to* $\varepsilon \in \mathbb{Y}$ *as:*

$$\mathbb{C}_{\theta,\varepsilon} = \{ M \in \mathbb{R}^{m \times m} \; ; \; M_{ij} = P_\theta(s_i \varepsilon t_j) \; , \; m \in \mathbb{N}, \quad s_i, t_j \in \mathbb{Y}^{\circledast} \}$$

Notice that to completely specify an element of $\mathbb{C}_{\theta,\varepsilon}$ we must give the order m and the 2m words of $\mathbb{Y}^{\circledast}$: $s_1, s_2, \ldots, s_m ; t_1, t_2, \ldots t_m$.

The generic element of $\mathbb{C}_{\theta,\varepsilon}$ will be thus denoted as

$$P_\varepsilon(s_1, s_2, \ldots s_m ; t_1, t_2 \ldots t_m) = \| P_\theta(s_i \varepsilon t_j) \|_{i,j=1,\ldots,m}$$

Definition β.4

Let $P_\theta \in \mathbb{P}$. *Define the rank of* P_θ *relative to* $\varepsilon \in \mathbb{Y}$ *as*

$$r_\theta(\varepsilon) = \sup_{M \in \mathbb{C}_{\theta,\varepsilon}} (\text{rank } M)$$

In general $r_\theta(\varepsilon)$ can assume any value between 1 and .

It is easily checked that $r_\theta(\varepsilon) = 1$ for Markov chains.

Proposition β.5

Let $P_\theta \in \mathbb{P}$ *and* $\theta = (n, A, B, \pi)$. *Then* $r_\theta(\varepsilon) \leq n$.

To prove proposition β.5 we need to develop a more powerful notation. Let $P_\theta \in \mathbb{P}$ and $\theta = (n, A, B, \pi)$. If $s = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_t \in \mathbb{Y}^\circledast$ then from proposition α.4 we have $P_\theta(s) = \pi M(\varepsilon_1) M(\varepsilon_2) \ldots M(\varepsilon_t) e$ where $M(\varepsilon) = AB_\varepsilon$. The following is an abuse of notation into which we will indulge. Define the matrix valued function:

$$M(.) \ : \ \mathbb{Y}^\circledast \rightarrow \mathbb{R}^{n \times n} \quad \text{as:}$$

$$M(s) = M(\varepsilon_1) M(\varepsilon_2) \ldots M(\varepsilon_t) \qquad \text{for} \quad s = \varepsilon_1 \varepsilon_2 \ldots \varepsilon_t$$

In terms of SFSS we have the following interpretation:

$$[M(s)]_{ij} = P[Y_1^t = s, X_t = j \mid X_0 = i]$$

As a direct consequence of its definition $M(.)$ enjoys of the composition property: $M(st) = M(s)M(t)$ and:

$$P_\theta(st) = \pi M(st) e = \pi M(s) M(t) e \qquad (\forall s, t \in \mathbb{Y}^\circledast).$$

Introduce the vector valued functions:

$$g(.) \ : \ \mathbb{Y}^\circledast \rightarrow \mathbb{R}^n \qquad (\text{ a } \underline{\text{row}} \text{ vector})$$

$$h(.) \ : \ \mathbb{Y}^\circledast \rightarrow \mathbb{R}^n \qquad (\text{ a } \underline{\text{column}} \text{ vector})$$

defined as follows:

$$g(s) = \pi M(s)$$

$$h(s) = M(s) e$$

Functions $g(.)$ and $g(.)$ have the following interpretation in terms of SFSS:

$$[g(s)]_j = P \ [Y_1^t = s, X_t = j] \qquad (j = 1, 2, \ldots, n)$$

$$[h(s)]_i = P \ [Y_1^t = s \mid X_0 = i] \qquad (i = 1, 2, \ldots, n)$$

With this new notation:

$$P \ (st) = \pi M(s) M(t) e = g(s) h(t) \qquad (\forall s, t \in \mathbb{Y}^\circledast)$$

proof of prop. β.5

Observe that $P_\theta(s_i \varepsilon t_j) = \pi M(s_i)M(\varepsilon)M(t_j)e = g(s_i)M(\varepsilon)h(t_j)$

Define:

$$G(s_1,..,s_m) = \begin{bmatrix} \text{---} & g(s_1) & \text{---} \\ \text{---} & g(s_2) & \text{---} \\ \cdots & \cdots & \cdots \\ \text{---} & g(s_m) & \text{---} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

and:

$$H(t_1,..,t_m) = \begin{bmatrix} | & | & & | \\ h(t_1) & h(t_2) & .. & h(t_m) \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times m}$$

Then:

$$P_\theta(s_1,s_2,..s_m;t_1,t_2,..,t_m) = G(s_1,s_2,..,s_m)M(\varepsilon)H(t_1,t_2,..,t_m)$$

Since $M(\varepsilon) \in \mathbb{R}^{n \times n}$, Sylvester's inequality gives $r_\theta(\varepsilon) \leq n$.

Definition β.6

Let $P_\theta \in \mathbb{P}$ *and* $\theta = (n,A,B,\pi)$

P_θ *is said regular if* $r_\theta(\varepsilon) = n \quad (\forall \varepsilon \in \mathbb{Y})$.

This definition translates to the case of stochastic functions of Markov Chains the notion of regulatity introduced in [Gilbert]. Regular pdf enjoy interesting properties. If P_θ is regular then $\forall \varepsilon \in \mathbb{Y}$ there exist sequences $s_1^\varepsilon, s_2^\varepsilon,..,s_n^\varepsilon; t_1^\varepsilon, t_2^\varepsilon,..,t_n^\varepsilon$ such that:

$$P\ (s_1^\varepsilon,..,s_n^\varepsilon;t_1^\varepsilon,..,t_n^\varepsilon) = G(s_1^\varepsilon,..,s_n^\varepsilon)M(\varepsilon)H(t_1^\varepsilon,..,t_n^\varepsilon)$$

has rank n.

Therefore $G(s_1^\varepsilon,..,s_n^\varepsilon)$, $M(\varepsilon)$, and $H(t_1^\varepsilon,..,t_n^\varepsilon)$ are invertible.

A first consequence is that a set of 2n words in $\mathbb{Y}^{\circledast}$ can be determined that achieve maximum rank in $\mathbb{C}_{\theta,\varepsilon}$ $\forall\varepsilon\in\mathbb{W}$.

Lemma β.7

If $P_\theta\in\mathbb{P}$ *is of order* n *and regular then there exist sequences* $(s_1,s_2,\ldots,s_n;t_1,t_2,\ldots,t_n)$ *such that* $P_\varepsilon(s_1,\ldots,s_n;t_1,\ldots,t_n)$ *is non-singular* $\forall\varepsilon\in\mathbb{W}$

proof:

Let $\varepsilon,\mu\in\mathbb{W}$ $(\varepsilon\neq\mu)$. With the preceding notation

$P_\mu(s_1^\varepsilon,\ldots,s_n^\varepsilon;t_1^\varepsilon,\ldots,t_n^\varepsilon) = G(s_1^\varepsilon,\ldots,s_n^\varepsilon)M(\mu)H(t_1^\varepsilon,\ldots,t_n^\varepsilon)$

from the last observation regularity implies that $G(s_1^\varepsilon,\ldots,s_n^\varepsilon)$, $M(\mu)$ and $H(t_1^\varepsilon,\ldots,t_n^\varepsilon)$ are non-singular, hence $P_\mu(\ldots)$ is non-singular. We can therefore choose e.g. $\{s_i;t_j\}\equiv\{s_i^1,t_j^1\}$

Another consequence of regularity follows from $M(\varepsilon)$ non-singular $\forall\varepsilon\in\mathbb{W}$.

Since $M(\varepsilon) = AB_\varepsilon$ and $B_\varepsilon = \text{diag}\{b_{1\varepsilon},b_{2\varepsilon},\ldots,b_{n\varepsilon}\}$ we conclude that, under regularity of P_θ :

i) A is invertible

ii) $b_{j\varepsilon}>0$ $(\forall j,\varepsilon)$

The following is a sufficient condition for regularity that follows immediately from the preceding discussion.

Lemma β.8

Let $P_\theta\in\mathbb{P}$ *and of order* n.

If $b_{j\varepsilon}>0$ $(\forall j,\varepsilon)$ *and* $r_\theta(\varepsilon) = n$ *for some* $\varepsilon\in\mathbb{W}$ *then* P_θ *is regular.*

Definition β.9

$\mathbb{P}'' = \{P_\theta \in \mathbb{P} \ ; \ P_\theta \text{ is regular } \}$

$\overline{\mathbb{P}} \ = \ \mathbb{P}' \cap \mathbb{P}''$

How big is $\overline{\mathbb{P}}$? Consider the parameter space Θ of $\mathbb{P}$. For fixed n, Θ consists of the set of all stochastic matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times r}$, $\pi \in \mathbb{R}^{1\times n}$, which is isomorphic to the euclidean $n(n+r-2) - 1$ dimensional space. It is shown in [Petrie] that the subset $\overline{\Theta} \subset \Theta$ corresponding to $\overline{\mathbb{P}}$ is open and of full Lebesgue measure in $\mathbb{R}^{n(n+r-2)-1}$.

Notice that $\overline{\Theta}$ is in bijective correspondence with the parametrization of $\mathbb{P}''$ through the $M(\varepsilon)$ matrices (since $b_{j\varepsilon} > 0$ in $\overline{\Theta}$).

We will thus feel free to abuse a little the notation and write $\theta = (k,A,B)$ or, where convenient, $(k, M(\varepsilon) \ \varepsilon \in \mathbb{Y})$ when referring to elements of $\overline{\Theta}$.

Identifiability results

The main result of this section is theorem β.10 on the identifiability of $\overline{\mathbb{P}}$. Again special notation has to be developed for its proof. Let P_θ be regular of rank n and let $\{s_i, t_j\}_1^n$ be sets of sequences as in theorem β.7. Remember that $g(s) = \pi M(s)$ $(\forall s \in \mathbb{Y}^{\circledast})$, a row vector in $\mathbb{R}^n$. Observe that $\{g(s_i)\}_1^n$ is a basis for the space $\mathbb{G} = \text{span}\{g(s) \ s \in \mathbb{Y}^{\circledast}\}$ in fact $\dim \mathbb{G} \leq n$ and $\{g(s_i)\}_1^n$ are independent since $G(s_1, \ldots, s_n)$ is non-singular. Therefore :

$$\forall \ s \in \mathbb{Y}^{\circledast} \quad \exists \ \{a_k(s)\}_1^n \quad \text{such that} \quad g(s) = \sum_{k=1}^{n} a_k(s) g(s_k)$$

Consider the special sequences s of the form $s = s_i\varepsilon$ We have:

$$g(s_i\varepsilon) = \sum_k a_k(s_i\varepsilon) g(s_k)$$

But $g(s_i\varepsilon) = \pi M(s_i\varepsilon) = \pi M(s_i) M(\varepsilon) = g(s_i) M(\varepsilon)$

Define the matrix:

$$A_\varepsilon = \|a_j(s_i\varepsilon)\| \quad i,j=1,..,n$$

then:

$$G(s_1,..,s_n) = A_\varepsilon G(s_1,..,s_n)$$

(from now on we will write G for $G(s_1,..,s_n)$, H for $H(t_1,..,t_n)$, and P_ε for $P_\varepsilon(s_1,..,s_n;t_1,..,t_n)$).

It is crucial to observe that A_ε <u>only depends on</u> P_θ <u>and not on</u> θ <u>itself</u>.

In fact multiplying (1) on the right by $h(t)$ we get:

$$P_\theta(st) = g(s)h(t) = \sum_k a_k(s)g(s_k)h(t) = \sum_k a_k(s)P\ (s_k t)$$

Applying the same technique for $s=s_i\varepsilon$ and $t=t_j$ $(i,j=1,..,n)$ we obtain the set of equations:

$$A_\varepsilon GH = P_\varepsilon$$

and prove the claim.

We make the following:

<u>Observation</u>:

If $\theta = (n,M(\varepsilon))$, $\bar{\theta} = (\bar{n},\bar{M}(\varepsilon)) \in \mathbb{P}''$ and $P_\theta(.) = P_{\bar{\theta}}(.)$ then $n = \bar{n}$.

In fact from the equivalence of the two pdf we conclude that the sets of compound sequence matrices $\mathbb{C}_{\theta,\varepsilon}$ and $\mathbb{C}_{\bar{\theta},\varepsilon}$ coincide. Therefore $r_\theta(\cdot) = r_{\bar{\theta}}(\varepsilon)$ $(\forall\varepsilon\in\mathbb{Y})$. From regularity we conclude that $n = \bar{n}$.

<u>Theorem</u> 3.10

$\bar{\mathbb{P}}$ is identifiable modulo permutations of the state space $\mathbb{X}$.

<u>proof</u>:

From the preceding observation it follows that if $P_\theta(.) = P_{\bar{\theta}}(.)$ then $n = \bar{n}$. What remains to be proved is that $\bar{A}$ and $\bar{B}$ differ from A and B for a permutation of the state space $\mathbb{X}$.

From the equivalence of the pdf we get $A_\varepsilon = \bar{A}_\varepsilon$ and:

$$G\,M(\varepsilon)\,H = \bar{G}\,\bar{M}(\varepsilon)\,\bar{H}$$

from which we conclude that $\bar{G}$ and $\bar{H}$ are invertible.

From the definition of A_ε we get:

$$\bar{G}\,\bar{M}(\varepsilon) = A_\varepsilon\,\bar{G}$$

$$G\,M(\varepsilon) = A_\varepsilon\,G$$

Hence $A_\varepsilon = G\,M(\varepsilon)\,G^{-1}$ and

$$\bar{G}\,\bar{M}(\varepsilon) = G\,M(\varepsilon)\,G^{-1}\,\bar{G}$$

or, in symmetric form:

$$(G^{-1}\,\bar{G}\,)\,\bar{M}(\varepsilon) = M(\varepsilon)\,(G^{-1}\,\bar{G}) \qquad (1)$$

Substitute to $M(\varepsilon)$ its value AB_ε and add over ε :

$$(G^{-1}\,\bar{G})\,\bar{A} = A\,(G^{-1}\,\bar{G})$$

which substituted in (1) gives:

$$(G^{-1}\,\bar{G})\,\bar{B}_\varepsilon = B_\varepsilon\,(G^{-1}\,\bar{G}) \qquad (2)$$

Define $X = G^{-1}\,\bar{G}$. To complete the proof we need to show that X is a permutation matrix.

First observe that $Xe = e$ (i.e. X is a stochastic matrix)

Since $P_\theta \in \mathbb{P}'$ there is $\varepsilon_0 \in \mathbb{Y}$ such that B_{ε_0} has distinct elements on the diagonal. Equation (2) now reads:

$$B_{\varepsilon_0}\,X = X\,\bar{B}_{\varepsilon_0}$$

which means that column j of X (denoted $x_{\cdot j}$) satisfies:

$$B_{\varepsilon_0}\,x_{\cdot j} = \bar{b}_{j\varepsilon_0}\,x_{\cdot j}$$

Since B_{0} is diagonal, $\bar{b}_{j\varepsilon_0} = b_{k\varepsilon_0}$ for some k, and $x_{\cdot j} = e_k$ (the k-th unit vector of $\mathbb{R}^n$). We conclude that X has exactly a one in each column, and since it is a stochastic matrix it has exactly a one in each row. It is therefore a permutation matrix.

ORDER DETERMINATION

In this chapter we will study how to determine the order of $P_Y(.)$ from the data.

Preliminaries

For convenience we restate here the problem and the assumptions. A piece of a sample path of a FMC $\{Y_t\}$ with values in $\mathbb{Y} = \{1,..,r\}$ is observed. Denote the data $\{y_1^T\}$ where T can be arbitrarily large. We assume that the pdf $P_Y(.)$ of $\{Y_t\}$ is such that:

$P_Y \in \overline{\mathbb{P}}$

Let $P_Y(.) = P_{\theta_0}(.)$ where $\theta_0 = (n,A^0,B^0)$

On the basis of the data determine n.

The generic element of $\overline{\Theta}$ is denoted by $\theta = (k,A,B)$, and the parameter set can be decomposed as: $\overline{\Theta} = \bigcup_k \overline{\Theta}_k$, where

$\overline{\Theta}_k = \{\theta \in \overline{\Theta} \ ; \ \theta = (k,A,B)\}$

Sometimes we denote the generic element of $\overline{\Theta}_k$ as $\theta_k = (A,B)$.

For typographical convenience we will denote the measure induced by θ on $\mathbb{Y}^{\circledast}$ by $P(.|\theta)$ or by $P_\theta(.)$

Let $\underline{\delta} = \inf\{ a^0_{ij}, b^0_j\}$. It follows from the assumptions that $\underline{\delta} > 0$ and clearly $n\,\underline{\delta} \leq 1$ (since A^0 stochastic), therefore: $n \leq |\,\underline{\delta}^{-1}|$.

Now let $\delta > 0$ be given and define

$\overline{\Theta}_\delta = \{\theta \in \overline{\Theta} \ ; \ a_{ij} \geq \delta \ , \ b_{j\varepsilon} \geq \delta \ \ \forall\ i,j,\varepsilon\ \}$

and $K = |\,\delta^{-1}\,|$

If $\theta = (k,A,B) \in \overline{\Theta}_\delta$ then $k \leq K$. Notice that for $\delta < \underline{\delta}$, $\theta_0 = (n,A^0,B^0) \in \overline{\Theta}_\delta$ (i.e. $n \leq K$).

We will assume to know a lower bound on $\underline{\delta}$ (and hence an upper bound on n, i.e. K)

For the time being we limit our attention to $\overline{\Theta}_\delta$ for $\delta \leq \delta\_$

With obious meaning of the symbols we have:

$$\overline{\Theta}_\delta = \bigcup_{k=1}^{K} \overline{\overline{\Theta}}_{\delta,k}$$

A Kullback type inequality

Define the following random variables on $\mathbb{Y}^{\oplus}$:

$$f_T(\theta,Y(.)) = P_\theta[Y_0 | Y_{-T-1}^{-1}] \qquad \theta \in \overline{\overline{\Theta}}_\delta$$

(by convention $f_1(\theta,Y(.)) = P_\theta[Y_0]$).

Next lemma, stated and proved for $\overline{\Theta}_{\delta,n}$ in [Baum] is instrumental for the developments of this section.

Lemma γ.1

$$f(\theta,Y(.)) = \lim_T f_T(\theta,Y(.))$$

exists for every $Y(.) \in \mathbb{Y}^{\otimes}$ *and is a continuous function of when restricted to* $\overline{\Theta}_{\delta,k}$ $\forall k \in (1,K)$.

proof:

see corollary 2.5 in [Baum]. To extend the proof to the present case where the order of θ can be different from n, a change is needed in corollary 2.1 where the quantity μ_δ must be defined as:

$$\mu_\delta = \delta^2/[\delta^2 + K-1]$$

Corollary γ.2

The function $h_k(.) : \overline{\Theta}_{\delta,k} \to \mathbb{R}$ given as:

$$h_k(\theta) = \mathbb{E}_{\theta_0}[\ \log f(\theta,Y(.))\ ]$$

is well defined and continuous $k \in (1,K)$

We are now in position to make use of the identifiability results of chapter β.

Theorem $\gamma.3$

$h_k(\theta) \leq h_n(\theta_0)$ $(\forall\ k, \theta \in \bar{\Theta}_\delta\ )$

$h_k(\theta) = h_n(\theta_0)$ iff $k=n$ and $(A,B) =$ Permutation (A^0, B^0)

proof:

The first equation follows directly from Jensen inequality. From theorem 3.1 of |Baum] it follows that the inequality is strict unless the induced measures P_θ and P_{θ_0} are equivalent. Therefore theorem $\beta.9$ implies the second equation.

From the abstract point of view theorems $\beta.9$ and $\gamma.3$ solve our problem. If the regular FMC $\{Y_t\}$ is actually generated by P_{θ_0} with $\theta_0 \in \bar{\Theta}_\delta$ then the order n of P_{θ_0} is identifiable and:

Corollary $\gamma.4$

n *is the unique index maximizing the finite sequence*

$h_k = \sup_{\theta\ \in \bar{\Theta}_{\delta,k}} h_k(\theta)$ $k \in (1,K)$

proof:

Since $h_k(\ )$ is continuous and $\bar{\Theta}_{\delta,k}$ is compact, $\forall\ k\ \exists\ \theta_k^{\#}$ such that $h_k = h_k(\theta_k^{\#})$. Conclusion follows from theorem $\gamma.3$

At this stage the missing link is: how do we compute the sequence h_k starting from the data, i.e. from a trajectory of $\{Y_t\}$?

To answer this question we first study the connection between the functions $h_k(\ )$ and the data.

Various special cases of the following (elementary) lemma are tacitly assumed in the litterature.

Lemma $\gamma.5$

If the stationary FMC $\{Y_t\}$ *has pdf* $P \in \mathbb{P}$ *and the corresponding Markov matrix* A *is irreducible and aperiodic then* Y_t *is mixing*

proof:

consists in a direct verification of the mixing condition:

$$\lim_{m \to \infty} P[Y_1^t=\varepsilon_1^t,\ Y_{t+m}^{t+m+s-1}=\delta_1^s] = P[Y_1^t=\varepsilon_1^t]P[Y_1^s=\delta_1^s] \qquad (\forall\ \varepsilon_1^t,\ \delta_1^s,\ t,\ s)$$

Let $T(m) = t+m+s-1$. The LHS is:

$$P[Y_1^t= \varepsilon_1^t,\ Y_{t+m}^{T(m)}=\delta_1^s] = \sum_{\gamma_1^{m-1}} P[Y_1^t=\varepsilon_1^t,\ Y_{t+1}^{t+m-1}=\gamma_1^{m-1},\ Y_{t+m}^{T(m)}=\delta_1^s] =$$

$$= \sum \pi M(\varepsilon_1^t)M(\gamma_1^{m-1})M(\delta_1^s)e =$$

$$= \pi M(\varepsilon_1^t)\ [\sum M(\gamma_1^{m-1})]M(\delta_1^s)e$$

$$= \pi M(\varepsilon_1^t)A^{m-1}M(\delta_1^s)e$$

Since A is irreducible aperiodic $\lim A^{m-1} = e\pi$

The conclusion follows.

In our Hypotheses $P_Y(.)$ satisfy the conditions of lemma $\gamma.5$ since $a_{i,j}^0 > 0$. The process $\{Y_t\}$ is therefore a fortiori ergodic. Next lemma shows the connection between $h_k(\ )$ and the data (for $\theta=\theta_0$ it reduces to the classical Shannon-McMillan-Breiman theorem of ergodic theory). The comment preceding lemma $\gamma.1$ applies here too.

Lemma $\gamma.6$

$$h_k(\theta) = \lim_{T\to\infty} \frac{1}{T} \log P[Y_1^T|\theta] \qquad \text{a.e. } P_{\theta_0}$$

proof:

see theorem 3.2 in [Baum] and the observation in the proof of lemma $\gamma.1$. The use of the ergodic theorem (as made in [Baum]) is justified by lemma $\gamma.5$.

To understand the connection between the sequence h_k and the data $\{y_1, y_2, .., y_T\}$ we have to dwell upon the asymptotic behaviour of the maximum of the log-likelihood function.

Define the sequence (wrt T):

$$h_{k,T} = \frac{1}{T} \log P[Y_1^T | \hat{\theta}_{k,T}] = \max_{\theta_k \in \overline{\Theta}_{\delta,k}} \frac{1}{T} \log P[Y_1^T | \theta_k]$$

Notice that since $P[Y_1^T | \theta_k]$ is continous on $\overline{\Theta}_{\delta,k}$ and bounded from below by δ^T (see lemma $\gamma.7$), $\frac{1}{T} \log P[Y_1^T | \theta_k]$ is continuous on $\overline{\Theta}_{\delta,k}$ and therefore we actually have a max.

Lemma $\gamma.7$

$$\delta^T \le P[Y_1^T | \theta] \le (1-\delta)^T \qquad \forall\ \theta \in \overline{\Theta}_\delta$$

proof:

Let $\theta = (k,A,B)$. From proposition $\sigma.3$ we have:

$$P_\theta[Y_1^{T-1}] - P_\theta[Y_1^T] = \pi(AB_{y_1} AB_{y_2} \dots AB_{y_{T-1}})(I-AB_{y_T})e$$

Since $\theta \in \overline{\Theta}_\delta$, $\delta e \le (I-AB_{y_T})e \le (1-\delta)e$

Therefore:

$$\delta P_\theta[Y_1^{T-1}] \le P_\theta[Y_1^{T-1}] - P_\theta[Y_1^T] \le (1-\delta) P_\theta[Y_1^{T-1}]$$

Rearranging terms:

$$\delta P_\theta[Y_1^{T-1}] \le P_\theta[Y_1^T] \le (1-\delta) P_\theta[Y_1^{T-1}]$$

From $P_\theta[Y_1] = \pi AB_{y_1} e$ we have: $\delta \le P_\theta[Y_1^T] \le 1-\delta$

(observe that $\pi A = \pi$ implies $\pi_i > \delta \quad i=1,..,k$)

Conclusion follows by finite induction.

Proposition γ.8

$$h_{k,\infty} = \lim_{T\to\infty} h_{k,T} \qquad \text{exists } \forall\ k\in(1,K)$$

proof:

We will show that $h_{k,T}$ is (wrt T) a decreasing sequence bounded from below.

$$T\ h_{k,T} = \max_{\theta_k} \log P[Y_1^T|\theta_k] = \max_{\theta_k} \log P[Y_T|Y_1^{T-1},\theta_k]\ P[Y_1^{T-1}|\theta_k] =$$

$$= \max \{ \log P[Y_T|Y_1^{T-1},\theta_k] + \log P[Y_1^{T-1}|\theta_k] \} =$$

$$\leq \max \log P[Y_T|Y_1^{T-1},\theta_k] + \max \log P[Y_1^{T-1}|\theta_k] =$$

$$\leq \max \log P[Y_1^{T-1}|\theta_k] =$$

$$= (T-1)\ h_{k,T-1}$$

Therefore $h_{k,T}$ is decreasing.

From lemma γ.7, $\delta^T \leq P[Y_1^T|\theta_k] \leq (1-\delta)^T$, and since $h_{k,T}= \frac{1}{T} \log P[Y_1^T|\hat{\theta}_{k,T}]$ for some $\hat{\theta}_{k,T} \in \overline{\Theta}_{\delta,k}$, we conclude $\log \delta \leq h_{k,T}$.

At present we know that:

$$h_{k,T} \xrightarrow[T\to\infty]{} h_{k,\infty} \qquad \forall\ k \in (1,K)$$

The nicest situation would be to have:

$$h_{k,\infty} = h_k \qquad \forall\ k \in (1,K) \qquad (*)$$

because if this is the case we can apply corollary γ.4 to $h_{k,\infty}$ and, as we will see there is a practical method to compute $h_{k,\infty}$ starting from the data.

Observe that (*) is equivalent to the following:

$$\lim_{T\to\infty} \max_{\theta\in\overline{\Theta}_{\delta,k}} \frac{1}{T} \log P[Y_1^T|\theta] = \max_{\theta\in\overline{\Theta}_{\delta,k}} \lim_{T\to\infty} \frac{1}{T} \log P[Y_1^T|\theta]$$

The interchange of limit operations in the preceding formula is legal under uniform convergence of the sequence of functions:

$h_{k,T}(.) \ : \ \overline{\Theta}_{\delta,k} \to \mathbb{R}$ defined as:

$h_{k,T}(\theta) = \frac{1}{T} \log P[Y_1^T|\theta]$

Proposition γ.9

$h_{k,T}(.) \xrightarrow[T\to\infty]{} h_k(.)$ uniformly on $\overline{\theta}_{\delta,k}$

proof:

We will check that the sequence $h_{k,T}(.)$ satisfies the conditions of U. Dini's criterion for uniform convergence.

The functions are defined on a compact set ($\overline{\theta}_{\delta,k}$)

they are continous,

$h_{k,T}(.)$ converges pointwise to $h_k(.)$ for $T\to\infty$ (lemma γ.6)

$h_K(.)$ is continous (lemma γ.6)

$h_{k,T}(.) \le h_{k,T-1}(.) \quad \forall \ \theta \in \overline{\theta}_{\delta,k}$

last assertion follows from:

$$T \ h_{k,T}(\theta) = \log P_\theta[Y_1^T] = \log P_\theta[Y_T|Y_1^{T-1}]P_\theta[Y_1^{T-1}]$$
$$= \log P_\theta[Y_T|Y_1^{T-1}] + \log P_\theta[Y_1^{T-1}]$$
$$\log P_\theta[Y_1^{T-1}] = (T-1) \ h_{k,T-1}(\theta)$$

Corollary γ.10

$h_{k,\infty} = h_k \qquad \forall \ k \in (1,K)$

A consistency result

The conclusion that we reached in the preceding section is that to determine the order, all we need to do is to compute the finite sequence $h_{k,\infty}$ and choose n as the maximizing index.

In practice we only have available a finite (hopefully large) number T of observations $\{y_1^T\}$ and therefore the best we can do is to compute $h_{k,T}$ for T large. We prove here that this is enough for the correct determination of the order.

Lemma γ.11

There exists T_0 *such that if* $T \geq T_0$ *then*

$$\max_k h_{k,T} = h_{n,T} \qquad \textit{and}$$

$$h_{k,T} < h_{n,T} \qquad \textit{for} \quad k \neq n$$

proof:

Since K is finite the convergence of $h_{k,T}$ to $h_{k,\infty}$ is trivially uniform with respect to k. From corollaries γ.4 and from the proof of γ.8, $h_{k,T}$ is decreasing in T ($\forall k$). It is then easily seen that for $\xi_* = \inf_k (h_{n,\infty} - h_{k,\infty})$ there exists T_0 such that $T \geq T_0$ implies $h_{n,T} - h_{n,\infty} < \frac{\xi_*}{3}$ and $h_{n,T} - h_{k,T} > \frac{2}{3} \xi_*$ ($k \neq n$). Hence

$$h_{n,\infty} < h_{n,T} < h_{n,\infty} + \frac{\xi_*}{3} \quad \text{and} \quad h_{k,T} < h_{n,\infty} - \frac{\xi_*}{3} \qquad (\forall\, k \neq n).$$

The practical computation of $h_{k,T}$ can be efficiently done using the Baum-Eagon algorithm, see [Levinson] for a description of the algorithm and the analysis of its numerical aspects.

REFERENCES

[Baum]
Baum, L., Petrie, T.
Statistical Inference for Probabilistic Functions of Finite State Markov Chains
Ann. Math. Stat., 37, 1966, pp. 1554-1563

[Carlyle]
Carlyle, J. W.
Stochastic Finite State System Theory
in System Theory (Zadeh, L. A., Polak, E. eds.)
chapter 10, McGraw-hill, New York 1969

[Finesso]
Finesso, L., Picci, G.
Linear Statistical Models and Stochastic Realization Theory
Proc. Symp. Analysis and Optimiz. of Systems, Nice, June 1984
Springer-Verlag, Lect. Notes Contr. Inf. Sc., vol 62, pp. 445-470

[Gilbert]
Gilbert, E. J.
On the Identifiability Problem for Functions of Finite Markov Chains
Ann. Math. Stat., 30, 1959, pp. 688-697

[Grenander]
Grenander, U.
Can we look inside an unreliable automaton?
in Festschrift for J. Neyman (David, ed.)
Wiley (1966) pp. 107-123

[Heller]
Heller, A.
Probabilistic Automata and Stochastic Transformations
Math. Sys. Theory, 1, 1967, pp. 197-208

[Kalman]
Kalman, R. E.
Identifiability and Modeling in Econometrics
in Developments in Statistics (Krishnaiah P. R. ed.)
vol 4, Academic Press, New York, 1983, pp. 97-136

[Levinson]
Levinson, S. E., Rabiner, L. R., Sondhi, M. M.
An Introduction to the Application of the Theory of Probabilistic
Functions of a Markov Process to Automatic Speech Recognition
Bell Tech. Jour., 62, 1983, pp. 1035-1074

[Paz]
Paz, A.
Introduction to Probabilistic Automata
Academic Press, New York, 1971

[Petrie]
Petrie, T.
Probabilistic Functions of Finite State Markov Chains
Ann. Math. Stat., 40, 1969, pp. 97-115

[Picci]
Picci, G.
On the Internal Structure of Finite State Stochastic Processes
in Recent Developments in Variable Structure Systems Proc. of a
U.S.-Italy seminar, Taormina, Sicily 1977, Springer-Verlag
Lect. Notes in Econ and Math Sys. vol 162 (1978)

[Picci, van Schuppen]
Picci, G., van Schuppen, J. H.
On the Weak Finite Stochastic Realization Problem