DEPARTMENT OF OPERATIONS RESEARCH
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK

TECHNICAL REPORT NO. 105

March 1970

ASYMPTOTIC EFFICIENCY OF THE MAXIMUM LIKELIHOOD ESTIMATORS FOR THE PARAMETERS OF CERTAIN STOCHASTIC PROCESSES

by

Dominique Jean-Marie Nocturne

Prepared under contracts
DA-31-124-ARO-D-474, U.S. Army Research Office-Durham
Nonr-401(53), Office of Naval Research
and
National Science Foundation Grant GP 7798

BIOGRAPHICAL SKETCH

The author was born on December 25, 1943, in Paris, France.

In October 1962, he was admitted to the Ecole Nationale Supérieure d'Arts et Métiers (Paris), where he majored in Electrical and Mechanical Engineering, and received his degree in June 1966.

Following his graduation, he came to the United States on a fellowship co-sponsored by the French and United States governments. Since September 1967, he has been a Research Assistant at Cornell.

The author is a member of Phi Kappa Phi, the Institute of Management Sciences, the Operations Research Society of America and the Association Française pour la Cybernétique Economique et Technique.

DEDICATION

to all those who contributed to my education

"If you give a man a fish
He will have a single meal
If you teach him how to fish
He will eat all his life."

Kuan-Tzu

ACKNOWLEDGMENTS

The topic of this dissertation was suggested to me by Professor L. Weiss, under whose direction it was written. I am particularly grateful to Professor L. Weiss for his constant readiness to help and to participate in discussions.

I also wish to thank Professors W. Bussman, K. O. Kortanek, T. C. Liu, W. Maxwell, N. U. Prabhu, who at some time or another served on my special committee.

Thanks also go to Professor R. Bechhofer, whose excellent teaching, added to the motivation I had, to work in the field of Statistics.

During my first year at Cornell, I held a scholarship granted by the French government. The remainder of my study at Cornell, and the preparation of this thesis were supported by the Army Research office, the National Science Foundation, and the Office of Naval Research.

Finally I express my appreciation to Mrs. Esther Huff for her patience and diligence in typing this thesis.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION AND SUMMARY

The purpose of this study is to develop a method of estimation leading to asymptotically efficient estimators.

An important theory is that of Maximum Likelihood which goes back to the 1920's and is due to R. A. Fisher. Its program can successfully be applied in the so-called "regular" case when competing estimators are limited to be asymptotically normally distributed.

To overcome this quite restrictive set of conditions, Weiss and Wolfowitz have recently introduced the concept of "Generalized Maximum Likelihood Estimator" and "Maximum Probability Estimator". Under some conditions the Maximum Probability Estimator, properly normalized, enjoys the asymptotic property of having the maximum probability of concentration around the true value of the unknown parameter.

In Chapter II, we show that for a given class of problems (including as a subclass the "regular" case), possessing some strong regularity conditions, the maximum likelihood estimator is asymptotically efficient.

In Chapter III, we apply our previous results to the problem of estimation for processes which can be described as Markov chains. This problem has been studied by a large number of people and it would be difficult to exhibit here a complete list of authors deserving credits. However, one of the most significant works is probably that of P. Billingsley (See [4] and [5]). In those two publications, the author

derives large sample properties (consistency and asymptotic normality) of the maximum likelihood estimator, when the underlying stochastic process is a time homogeneous, ergodic Markov process, falling in one of 3 classes:

(1) discrete time, discrete state space (Markov chain)

(2) discrete time, continuous state space

(3) continuous time, discrete state space.

In the subsequent study we essentially consider two additional problems:

(1) that of asymptotic efficiency

(2) that of estimation for non-homogeneous Markov chains.

This last question has previously been studied by a few people (see for example Anderson and Goodman [1], Gold [14]). In particular these authors make statistical inference for a class of finite, non-homogeneous Markov chains based on a large number of observations taken at times $t = 0, 1, \ldots\ldots, T$. Consequently this approach fits in the "regular" case which we are not interested in.

Our approach allows only one observation at a time $t = 0, 1, \ldots, T$ and asymptotic results are derived by letting T go to infinity.

In the first part of Chapter IV, we deal with parametric statistical estimation problems with respect to some discrete time - continuous space stochastic processes. Such questions are of fundamental interest in the econometric field and many researchers have devoted their time to this area. See for example the works by:

(1) Mann and Wald [22]

(2) The Cowles Commission for Research in Economics - and in particular:

- monograph 10: Statistical Inference in Dynamic Economic Models. Edited by T. C. Koopmans (1950) John Wiley

- monograph 14: Studies in Econometrics Methods. Edited by W. C. Hood and T. C. Koopmans (1953) John Wiley.

(3) T. W. Anderson and H. Rubin

- Estimation of the parameters of a single equation in a complete system of stochastic equations. Annals of Math. Statistics, Vol. 20 (1949) pp 46-63.

- The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. Annals of Math. Statistics, Vol. 21 (1950) pp 570-582.

In his book on econometrics Tintner [29] suggested looking at forms of functional relationships more complicated than that of stochastic difference equations. In particular he proposed to describe economic systems in terms of stochastic differential equations.

In the second part of Chapter IV we shall look at the estimation problem, which arises in connection with this suggestion, and we shall approach the estimation problem from a uniform point of view, using our basic theorem 2.7 of Chapter II.

Finally in Chapter V we consider the case of continuous time, completely discontinuous Markov processes and Markov renewal processes.

P. Billinsgley has considered in his monograph [4] the problem of analysing statistically a Markov process $\{X(t)\ ,\ 0 \leq t \infty\}$ in which the time parameter is continuous. He focusses attention on processes of the completely discontinuous type and showed that under

certain conditions the maximum likelihood estimator of some unknown parameter (possibly a vector) is consistent and asymptotically normally distributed.

We shall show that in the case of a finite state, time continuous Markov process of the jump type, and under the same conditions given by Billingsley, the maximum likelihood estimator of some unknown parameter is asymptotically efficient.

One can generalize the previous results by looking at a Markov renewal process. In their paper, (see [23]) Moore and Pyke study the large sample properties of a non-parametric estimator of the transition distributions of a Markov renewal process with finitely many states.

In the second part of this chapter, we shall study the problem of parametric estimation for such a process and in particular we shall show that under suitable conditions the maximum likelihood estimator of some unknown parameter is asymptotically efficient.

CHAPTER II

ASYMPTOTIC EQUIVALENCE OF THE MAXIMUM LIKELIHOOD ESTIMATOR AND OF THE MAXIMUM PROBABILITY ESTIMATOR

2.1 Notation

We shall use the following notation, mainly borrowed from [32].

For each positive integer n let:

- $X(n)$ be the finite vector of observations
- $K_n(x|\theta)$ be the density (L measure) of $X(n)$, when θ (an m-vector) is the value of the (unknown to the statistician) parameter. θ lies in Θ the parameter space.
- $\hat{d}(X(n))$ be the maximum likelihood estimator (MLE) of θ
- $\overset{\gamma}{d}(X(n))$ be the maximum probability estimator (MPE) of θ with respect to a measurable region R in m-space.

2.2 Consistency of the Maximum Probability Estimator.

2.2.1 One parameter case

A MPE will be a value of d that maximizes

$$\int_{d-r(n)}^{d+r(n)} K_n(x|\theta)d\theta = Z_n(x,d) \text{ , say where } \{r(n)\} \to 0 \text{ as } n \to \infty .$$

Theorem 2.1

Assume:

(1) The likelihood function can be written in the form:

$$K_n(x|\theta) = \prod_{i=1}^{n} f(x_{i-1}, x_i|\theta) \cdot h_n(x_0, \cdots, x_n)$$

where

- $K_n(x|\theta) \varepsilon L_\theta$ for almost every x
- x_{i-1}, x_i are finite vectors of observations
- $h_n(x_0, \cdots, x_n)$ is independent of θ .

The above format offers the advantage of including:

(i) the independent, identically distributed case i.e.:

$$K_n(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

(ii) the case where $\{x_i\}$ follows a Markov process i.e.:

$$K_n(x|\theta) = \prod_{i=1}^{n} f(x_{i-1}, x_i|\theta)$$

(iii) However it is still more general and will allow us to study other statistical problems, such as those met in econometrics, where often it is assumed that conditionally on a set of exogenous variables, the endogenous variables follow a Markov process.

(2) For almost all ξ and η $g(\xi,\eta|\theta) = \text{Log } f(\xi,\eta|\theta)$ has continuous derivatives throughout θ, up to the third order.

(3) Let $G(\xi,\eta) = \sup_{\theta \epsilon N} \left| \frac{\partial^3}{\partial \theta^3} g(\xi,\eta|\theta) \right|$

where N is some neighborhood of θ_0 - the true value of θ.

Then we assume:

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} g(x_{i-1},x_i|\theta_0) = 0 \quad \text{(in probability)}$$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} g(x_{i-1},x_i|\theta_0) = -k^2(\theta_0) < 0 \quad \text{(in probability)}$$

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} G(x_{i-1},x_i) = M \quad \text{(in probability)}$$

Then $\tilde{d}(x_0,x_1,\ldots,x_n)$ is a consistent estimator of θ_0.

Proof:

(5) $\frac{\partial Z_n}{\partial d}(x,d) = K_n(x|d + r(n)) - K_n(x|d - r(n))$ (for almost every x)

solving for $\tilde{d}_n$ the MPE, we let

(10) $\frac{\partial Z_n}{\partial d}(x,\tilde{d}) = 0 \iff K_n(x|\tilde{d} + r(n)) = K_n(x|\tilde{d} - r(n))$

taking Log on both sides, we get

$$(15)\quad \sum_{i=1}^{n} \text{Log } f(x_{i-1},x_i|\hat{\theta} + r(n)) = \sum_{i=1}^{n} \text{Log } f(x_{i-1},x_i|\hat{\theta} - r(n))$$

by assumption (2) we can expand Log $f(\cdot)$ in a Taylor's series in some neighborhood N around θ_0 .

$$(20)\text{Log } f(x_{i-1},x_i|\hat{\theta} \pm r(n)) = \text{Log } f(x_{i-1},x_i|\theta_0)$$

$$+ (\hat{\theta} \pm r(n) - \theta_0) \cdot \frac{\partial}{\partial\theta} \text{Log } f(x_{i-1},x_i|\theta_0)$$

$$+ \frac{(\hat{\theta} \pm r(n) - \theta_0)^2}{2} \cdot \frac{\partial^2}{\partial\theta^2} \text{Log } f(x_{i-1},x_i|\theta_0)$$

$$+ \frac{\alpha(\hat{\theta} \pm r(n) - \theta_0)^3}{3!} \cdot G(x_{i-1},x_i) \qquad |\alpha| \leq 1 \ .$$

Collecting terms, (15) becomes, after multiplication by $\frac{1}{2\cdot n\cdot r(n)}$

$$(25)\quad \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \text{Log } f(x_{i-1},x_i|\theta_0) + (\hat{\theta} - \theta_0)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \text{Log } f(x_{i-1},x_i|\theta_0)$$

$$+ \frac{\alpha}{2}[(\hat{\theta} - \theta_0)^2 + \frac{r^2(n)}{3}] \cdot \frac{1}{n}\sum_{i=1}^{n} G(x_{i-1},x_i) = 0$$

$$(30)\quad \text{let}\quad B_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \text{Log } f(x_{i-1},x_i|\theta_0)$$

$$B_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \text{Log } f(x_{i-1},x_i|\theta_0)$$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n} G(x_{i-1},x_i)$$

by assumption (3) we have :

(35)
$$\lim_{n\to\infty} B_0 = 0 \qquad \text{(in probability)}$$

$$\lim_{n\to\infty} B_1 = -k^2(\theta_0) \qquad \text{(in probability)}$$

$$\lim_{n\to\infty} B_2 = M \qquad \text{(in probability)}$$

Let δ and ε be given arbitrarily small positive numbers, then for $n > n_0(\varepsilon,\delta,\{r(n)\})$

$$-P(|B_0| \geq \delta^2) < \frac{1}{3}\varepsilon$$

$$-P(B_1 \geq -\frac{1}{2}k^2) < \frac{1}{3}\varepsilon$$

$$-P(|B_2| \geq 2M) < \frac{1}{3}\varepsilon$$

$$-r^2(n) < 3\delta^2$$

Let $S = \{x \mid |B_0| < \delta^2,\ B_1 < -\frac{1}{2}k^2,\ |B_2| < 2M\}$

then following Cramér's argument (see [6] pp 502-503)

$$P(S) > 1 - \varepsilon \quad \text{as soon as} \quad n > n_0$$

For $\vartheta = \theta_0 \pm \delta$ the expression in (25) assumes the values

(40)
$$B_0 \pm B_1\delta + \frac{1}{2}\alpha B_2 \cdot 2\delta^2$$

$$x \in S \rightarrow \left|B_0 + \frac{1}{2}\alpha B_2 \cdot 2\delta^2\right| < \delta^2(1 + 2M)$$

If $\delta < \frac{1}{2}k^2/(1 + 2M)$ the sign of (40) depends only on the second term. So that

$$\frac{\partial Z_n}{\partial d}(x,\theta_0+\delta) < 0 \quad \text{and} \quad \frac{\partial Z_n}{\partial d}(x,\theta_0-\delta) > 0$$

which imply by assumption (2), that with probability greater than $1-\varepsilon$, $|\overset{\gamma}{d}_n - \theta_0| < \delta$ as soon as $n > n_0(\varepsilon,\delta,\{r(n)\})$.

Remark 1: The same argument goes through, with minor changes in the algebra, when taking the MPE w.r.t. a non-symmetric region i.e.

$$Z_n(x,d) = \int_{d-r_1(n)}^{d+r_2(n)} K_n(x|\theta)\, d\theta$$

where $r_i(n) = [k(n)]^{-1} \cdot r_i \qquad i = 1, 2$

There remains to show that $\overset{\gamma}{d}_n$ is a local maximum. For this purpose, we shall study the function $\frac{\partial Z_n}{\partial d}(x,d)$ in a neighborhood N' of $\overset{\gamma}{d}_n$.

$$(41) \qquad \frac{\partial}{\partial d}\left[\frac{\frac{\partial Z_n}{\partial d}(x,d)}{K_n(x|d-r(n))}\right] = \frac{\frac{\partial^2 Z_n}{\partial d^2}(x,d) \cdot K_n(x|d-r(n)) - \frac{\partial Z_n}{\partial d}(x,d) \cdot \frac{\partial K_n}{\partial d}(x|d-r(n))}{[K_n(x|d-r(n))]^2}$$

$$\left\{\frac{\partial}{\partial d}\left[\frac{\frac{\partial Z_n}{\partial d}(x,d)}{K_n(x|d-r(n))}\right]\right\}_{d=\overset{\gamma}{d}_n} = \frac{\partial^2 Z_n}{\partial d^2}(x,\overset{\gamma}{d}_n)/K_n(x|\overset{\gamma}{d}_n - r(n))$$

It is seen from (41) that $\frac{\partial Z_n}{\partial d}(x,d)$ is decreasing over N' iff so are:

$$\frac{\frac{\partial Z_n}{\partial d}(x,d)}{K_n(x|d-r(n))}\ ,\ 1+\frac{\frac{\partial Z_n}{\partial d}(x,d)}{K_n(x|d-r(n))}\ ,\ \text{and Log}\left(1+\frac{\frac{\partial Z_n}{\partial d}(x,d)}{K_n(x|d-r(n))}\right) \equiv Y_n(x|d)$$

$$Y_n(x|d) = \text{Log}\ \frac{K_n(x|d+r(n))}{K_n(x|d-r(n))}$$

By (25) we have

(42)

$$\frac{1}{2n\cdot r(n)}\cdot\frac{\partial Y_n}{\partial d}(x|d) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\text{Log}\ f(x_{i-1},x_i|\theta_0) + \alpha(d-\theta_0)\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}G(x_{i-1},x_i)$$

which shows that with probability going to one

$$\lim_{n\to\infty}\ \frac{1}{2n\cdot r(n)}\cdot\frac{\partial Y_n}{\partial d}(x,\tilde{d}) = -k^2(\theta_0)$$

It follows that with probability going to one, there is at most one solution of (10) with $|\tilde{d}_n - \theta_0| < \delta$ and any such solution is a local maximum of $Z_n(x,d)$.

2.2.2 Multi-parameter case

A MPE will be a value of d that maximizes

$$\int_{d_1-r_1(n)}^{d_1+r_1(n)} \cdots \int_{d_m-r_m(n)}^{d_m+r_m(n)} K_n(x|\theta)\,d\theta = Z_n(x,d)\ ,\ \text{say}$$

where $\{r_i(n)\} \to 0$ as $n \to \infty$ $\quad i = 1,2,\cdots,m$

and $d = (d_1,\cdots,d_m)$

$\theta = (\theta_1,\cdots,\theta_m)$

Theorem 2.2

Assume

(1) The likelihood function can be written in the form:

$$K_n(x|\theta) = \prod_{i=1}^{n} f(x_{i-1},x_i|\theta) \cdot h_n(x_0,\cdots,x_n)$$

where

- $K_n(x|\theta) \in L_\theta$ for almost every x
- x_{i-1},x_i are finite vectors of observations
- $h_n(x_0,\ldots,x_n)$ is independent of θ .

(2) For almost all ξ and η $g(\xi,\eta|\theta) = \text{Log } f(\xi,\eta|\theta)$ has continuous partial derivatives throughout Θ , up to the third order

(3) Let $$G(\xi,\eta) = \sup_{\theta \in N} \left| \frac{\partial^3 g(\xi,\eta|\theta)}{\partial\theta_u \partial\theta_v \partial\theta_w} \right|$$

where N is some neighborhood of θ_0 , the true value of θ .

Then we assume:

$$- \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial\theta_u} g(x_{i-1},x_i|\theta_0) = 0 \qquad \text{(in probability)} \quad u = 1, 2,\cdots,m$$

$$- \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_u\partial\theta_v} g(x_{i-1},x_i|\theta_0) = \sigma_{uv} \qquad \text{(in probability)} \quad u,v = 1, 2,\cdots,m$$

where $\{\sigma_{uv}\}$ is a positive definite matrix

$$- \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} G(x_{i-1},x_i) = M \qquad \text{(in probability)}$$

Then $\hat{d}(x_0,x_1,\ldots,x_n) = (\hat{d}_1,\hat{d}_2,\ldots,\hat{d}_m)$ is a consistent estimator of $\theta_0 = (\theta_{01},\theta_{02},\ldots,\theta_{0m})$.

Proof

$$(45)\quad \frac{\partial Z_n}{\partial d_1}(x,d) = \int_{d_2-r_2(n)}^{d_2+r_2(n)} \cdots \int_{d_m-r_m(n)}^{d_m+r_m(n)} \{K_n(x|d_1 + r_1(n), \theta_2,\ldots,\theta_m) - K_n(x|d_1 - r_1(n), \theta_2,\ldots,\theta_m)\}\, d\theta .$$

solving for the MPE we let:

$$(50)\qquad \frac{\partial Z_n}{\partial d_i}(x,\hat{d}) = 0 \qquad i = 1,\ldots,m$$

let $\quad S_n = \{\theta: \hat{d}_i - r_i(n) \le \theta_i \le \hat{d}_i + r_i(n) \;;\; i = 2,\ldots m\}$

we now show that (50) implies that there exists a point $(\bar{\theta}_2,\cdots,\bar{\theta}_m) \in S$ such that

(55) $K_n(x|\tilde{d}_1 + r_1(n),\bar{\theta}_2,\ldots,\bar{\theta}_m) = K_n(x|\tilde{d}_1 - r_1(n),\bar{\theta}_2,\cdots,\bar{\theta}_m)$

First we claim that there exist 2 points $M_1,M_2 \in S$ such that:

$$K_n(x|\tilde{d}_1 + r_1(n),M_1) > K_n(x|\tilde{d}_1 - r_1(n),M_1)$$

and $$K_n(x|\tilde{d}_1 + r_1(n),M_2) < K_n(x|\tilde{d}_1 - r_1(n),M_2) \quad *$$

This is clear from the fact that otherwise, the value of the integral in (45) could certainly not be zero when the region over which we compute it, is S.

Let the coordinates of M_1 and M_2 be respectively $(\theta_{21},\cdots\theta_{m1})$ and $(\theta_{22},\cdots,\theta_{m2})$.

Let

$$\Delta K_n(x|\tilde{d}_1,t) = K_n[x|\tilde{d}_1 + r_1(n),\theta_{21} + t(\theta_{22} - \theta_{21}),\ldots,\theta_{m1} + t(\theta_{m2} - \theta_{m1})]$$

$$- K_n[x|\tilde{d}_1 - r_1(n),\theta_{21} + t(\theta_{22} - \theta_{21}),\cdots,\theta_{m1} + t(\theta_{m2} - \theta_{m1})]$$

where $t \in [0,1]$

clearly $\Delta K_n(x|\tilde{d}_1,0) > 0$, $\Delta K_n(x|\tilde{d}_1,1) < 0$

For almost every $x, K_n(x|\theta)$ is a continuous function of θ, and $\Delta K_n(x|\tilde{d}_1,t)$ is a continuous function of t. All conditions of the fixed-point theorem hold and the result claimed above follows.

---

* Obviously the case where $K_n(x|\tilde{d}_1 + r_1(n),M) = K_n(x|\tilde{d}_1 - r_1(n),M)$ for all $M \in S$ offers no difficulty.

The proof of Theorem 2.2 consists showing that $(\hat{d}_1, \bar{\theta}_2, \cdots, \bar{\theta}_m)$ converges in probability to $(\theta_{01}, \cdots, \theta_{0m})$.

Since $|\bar{\theta}_i - \hat{d}_i| \leq 2r_i(n)$ $(i = 2, \cdots, m)$, the result just claimed implies that as $n \to \infty$, $\hat{d}(X(n)) \to \theta_0$ in probability.

Expanding Log $f(\cdot)$ in a Taylor's series in some neighborhood of θ_0 , we get:

$$
\begin{aligned}
(60)\quad \text{Log } f(x_{i-1}, x_i | \hat{d}_1 \pm r_1(n), \bar{\theta}_2, \cdots, \bar{\theta}_m) &= \text{Log } f(x_{i-1}, x_i | \theta_0) \\
&+ (\hat{d}_1 \pm r_1(n) - \theta_{01}) \frac{\partial}{\partial \theta_1} \text{ Log } f(x_{i-1}, x_i | \theta_0) \\
&+ \sum_{j=2}^{m} (\bar{\theta}_j - \theta_{0j}) \frac{\partial}{\partial \theta_j} \text{ Log } f(x_{i-1}, x_i | \theta_0) \\
&+ \frac{(\hat{d}_1 \pm r_1(n) - \theta_{01})^2}{2} \cdot \frac{\partial^2}{\partial \theta_1^2} \text{Log } f(x_{i-1}, x_i | \theta_0) \\
&+ \sum_{j=2}^{m} \frac{(\bar{\theta}_j - \theta_{0j})^2}{2} \frac{\partial^2}{\partial \theta_j^2} \text{ Log } f(x_{i-1}, x_i | \theta_0) \\
&+ \sum_{j=2}^{m} (\hat{d}_1 \pm r_1(n) - \theta_{01})(\bar{\theta}_j - \theta_{0j}) \frac{\partial^2}{\partial \theta_1 \partial \theta_j} \text{ Log } f(x_{i-1}, x_i | \theta_0) \\
&+ \frac{\alpha}{3!} \{(\hat{d}_1 \pm r_1(n) - \theta_{01}) + \sum_{j=2}^{m} (\bar{\theta}_j - \theta_{0j})\}^3 G(x_{i-1}, x_i) \quad |\alpha| \leq 1 \ .
\end{aligned}
$$

Collecting terms (55) becomes:

(65)

$$n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_1}\,\text{Log}\,f(x_{i-1},x_i|\theta_0)+(\hat{\gamma}_1-\theta_{01})n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_1^2}\text{Log}\,f(x_{i-1},x_i|\theta_0)$$

$$+\sum_{j=2}^{m}(\bar{\theta}_j-\theta_{0j})n^{-1}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_1\partial\theta_j}\,\text{Log}\,f(x_{i-1},x_i|\theta_0)$$

$$+\frac{\alpha}{3!}\{[(\hat{\gamma}_1+r_1(n)-\theta_{01})+\sum_{j=2}^{m}(\bar{\theta}_j-\theta_{0j})]^2$$

$$+[(\hat{\gamma}_1+r_1(n)-\theta_{01})+\sum_{j=2}^{m}(\bar{\theta}_j-\theta_{0j})]\cdot[(\hat{\gamma}_1-r_1(n)-\theta_{01})+\sum_{j=2}^{m}(\bar{\theta}_j-\theta_{0j})]$$

$$+[(\hat{\gamma}_1-r_1(n)-\theta_{01})+\sum_{j=2}^{m}(\bar{\theta}_j-\theta_{0j})]^2\}\cdot n^{-1}\sum_{i=1}^{n}G(x_{i-1},x_i)=0$$

Let $Q_1(\hat{\gamma}_1,\bar{\theta}_2,\cdots,\bar{\theta}_m)$ be the left hand side of (65), then the fact that $|\bar{\theta}_j-\hat{\gamma}_j|\leq 2r_j(n)$ and assumption (3) imply that for n sufficiently large

$P[|Q_i(\hat{\gamma}_1,\hat{\gamma}_2,\cdots,\hat{\gamma}_m)|<\varepsilon \quad i=1,\cdots,m]>1-\delta$ for any ε,δ positive numbers. It can then be proved in the same way as in Theorem 2.1 of [4] pp 10-13, that the MPE is a consistent estimator.

Remark 2: As states in Remark 1, the same result holds if the region over which we are integrating is not symmetric.

2.3 Asymptotic Distribution of the Maximum Probability Estimator and its Equivalence with that of the Maximum Likelihood Estimator.

2.3.1 One parameter case

Theorem 2.3

Assume the same conditions given in Theorem 2.1, plus:

(4) $\frac{\partial}{\partial\theta}\, g(\xi,n|\theta)$ has moments of order $2+\delta$ for some positive number δ.

(5) For any $\theta \in \Theta$, there exists a neighborhood N of θ such that for all x_{i-1}

$$E\left\{\sup_{\theta\in N}\left|\frac{\partial f(x_{i-1},x_i|\theta)}{\partial\theta}\right|\,\Big|\Big|\, x_{i-1}\right\} < \infty$$

(6) Let $F_0,F_1,\cdots$ be a nondecreasing sequence of Borel fields such that

$$E\left[\frac{\partial}{\partial\theta}\, g(x_{i-1},x_i|\theta)\,||F_{i-1}\right] = 0 \qquad \text{w.p.1} \quad i = 1,\cdots,n$$

then we suppose that:

$$-\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E\left[\left(\frac{\partial}{\partial\theta}\, g(x_{i-1},x_i|\theta_0)\right)^2 \,||\, F_{i-1}\right] = k^2(\theta_0) > 0 \qquad \text{(in probability)}$$

and

$$- \lim_{n\to\infty} n^{-1-\delta/2} \sum_{i=1}^{n} E\left[\left|\frac{\partial}{\partial\theta} g(x_{i-1},x_i|\theta)\right|^{2+\delta} \middle| F_{i-1}\right] = 0 \quad \text{(in probability)}$$

Then $(\sqrt{n}(\hat{d}(x_0,\cdots,x_n) - \theta_0) \xrightarrow{L} N(0,\frac{1}{k^2})$ as $n \to \infty$.

Proof Let $\hat{d}_n$ be the solution of (10). From (25) we obtain:

(70)

$$\sqrt{n}(\hat{d}_n-\theta_0) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} g(x_{i-1},x_i|\theta_0) + \frac{\alpha}{6} \frac{r^2(n)}{\sqrt{n}} \sum_{i=1}^{n} G(x_{i-1},x_i)}{-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} g(x_{i-1},x_i|\theta_0) - \frac{1}{n}\frac{\alpha}{2}(\hat{d}_n-\theta_0) \sum_{i=1}^{n} G(x_{i-1},x_i)}$$

it follows from Theorem 2.1 that the denominator of the R. H. S. of (70) converges in probability to $+ k^2$.

Since we use on the L. H. S. of (70) the normalizing factor $\sqrt{n}$, this implies that $r(n)$ is of order $1/\sqrt{n}$ and consequently that the second term in the numerator goes to zero in probability.

Conditions (2) and (5) make possible to show that the partial sums

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} g(x_{i-1},x_i|\theta)$$

form a martingale - so that we can now apply Billingsley's Theorem 9.1 (see [4] pp 52-61).

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} g(x_{i-1},x_i|\theta_0) \xrightarrow{L} N(0,k^2) \quad \text{as} \quad n \to \infty .$$

From this, it follows that, as $n \to \infty$ $\sqrt{n}(\hat{d}_n - \theta_0) \xrightarrow{L} N(0, \frac{1}{k^2})$

Corollary 2.3.1

Under the same conditions required for Theorem 2.3, we have, denoting by $\hat{d}(x_0,x_1,\cdots,x_n)$ a M. L. E. of θ_0 .

(1) $\sqrt{n}(\hat{d}(x_0,\cdots,x_n) - \theta_0) \xrightarrow{L} N(0,\frac{1}{k^2})$ as $n \to \infty$

(2) $\sqrt{n}(\hat{d}(x_0,\cdots,x_n) - \overset{\vee}{d}(x_0,\cdots,x_n)) \to 0$ in probability as $n \to \infty$

Proof: (1) This is a straightforward extension of Billingsley's results [4] since the likelihood equation $\frac{\partial K_n}{\partial \theta}(x,\hat{d}) = 0$ does not involve $h_n(x_0,\cdots,x_n)$

(2) from (70) we have:

(70)

$$\sqrt{n}(\overset{\vee}{d}_n-\theta_0) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}g(x_{i-1},x_i|\theta_0) + \frac{\alpha_1}{\delta}\frac{r^2(n)}{\sqrt{n}}\sum_{i=1}^{n}G(x_{k-1},x_i)}{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}g(x_{i-1},x_i|\theta_0) - \frac{1}{n}\frac{\alpha_1}{2}(\overset{\vee}{d}_n-\theta_0)\sum_{i=1}^{n}G(x_{i-1},x_i)}$$

if $\hat{d}_n$ is the maximum likelihood estimator of θ_0 we have the similar expression (see [4] page 11).

$$\sqrt{n}(\hat{d}_n-\theta_0) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}g(x_{i-1},x_i|\theta_0)}{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}g(x_{i-1},x_i|\theta_0) - \frac{\alpha_2}{2}(\hat{d}_n-\theta_0)\sum_{i=1}^{n}G(x_{i-1},x_i)}$$

$|\alpha_1| < 1, |\alpha_2| < 1$

from assumption (3) of Theorem 2.1, (70) and (72) it follows that $\sqrt{n}(\hat{d}_n - \tilde{d}_n) \to 0$ in probability, as $n \to \infty$.

2.3.2 Multi-parameter case, $\theta = (\theta_1,\cdots,\theta_m)$

It is a straightforward generalization of the one dimensional case and we shall merely state the results.

Theorem 2.4

Assume the same conditions given in Theorem 2.2, plus:

(4) $\frac{\partial}{\partial\theta_u} g(\xi,\eta|\theta)$ has moments of order $2 + \delta$ $(\delta > 0)$ $u = 1,2,\cdots,m$

(5) for any $\theta \in \Theta$, there exists a neighborhood N of θ such that for all x_{i-1}

$$E\left\{\sup_{\theta\in N}\left|\frac{\partial f(x_{i-1},x_i|\theta)}{\partial\theta_u}\right|\Big|\Big|\; x_{i-1}\right\} < \infty \qquad u = 1,\cdots,m$$

(6) Let $F_0,F_1,\ldots$ be a non decreasing sequence of Borel fields such that

$$E\left[\frac{\partial}{\partial\theta_u} g(x_{k-1},x_k|\theta)\,||\, F_{k-1}\right] = 0 \quad \text{w.p.1} \qquad \begin{matrix} k = 1,\cdots,n \\ u = 1,\ldots,m \end{matrix}$$

then we suppose that:

$$-\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E\left[\frac{\partial}{\partial\theta_u} g(x_{k-1},x_k|\theta) \cdot \frac{\partial}{\partial\theta_v} g(x_{k-1},x_k|\theta)\,||F_{k-1}\right] = \sigma_{uv}(\theta)$$

(in probability) $u,v = 1,\cdots,m$

and

$$-\lim_{n\to\infty} n^{-1-\delta/2} \sum_{k=1}^{n} E\left[\left|\frac{\partial}{\partial\theta_u} g(x_{k-1},x_k|\theta)\right|^{2+\delta}\,\Big|\,F_{k-1}\right] = 0$$

(in probability) $u = 1,\cdots,m$

Then $\sqrt{n}(\tilde{d}(x_0,\cdots,x_n) - \theta_0) \xrightarrow{L} N(0,\sigma^{-1}(\theta_0))$ as $n \to \infty$.

Corollary 2.4.1

Under the same conditions required for Theorem 2.4, we have:

(1) $\sqrt{n}(\hat{d}(x_0,\cdots,x_n) - \theta_0) \xrightarrow{L} N(0,\sigma^{-1}(\theta_0))$ as $n \to \infty$

(2) $\sqrt{n}(\hat{d}(x_0,\cdots,x_n) - \tilde{d}(x_0,\cdots,x_n)) \to 0$ in probability as $n \to \infty$.

2.4 Asymptotic Efficiency of the Maximum Likelihood Estimator for a Certain Class of Problems

First, we shall recall a very useful result due to P. Billingsley (see [4] pp 52-61).

Theorem 2.5

Let $u_1, u_2, \cdots$ be random variables with moments of order $2 + \delta$, $\delta > 0$ and let $F_0, F_1, \cdots$ be a nondecreasing sequence of Borel Fields such that

(1) $E[u_n || F_{n-1}] = 0$ w.p.1 $n = 1,2,\cdots$

(2) $\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} E[u_k^2 || F_{k-1}] = \beta^2 \geq 0$ w.p.1

and $\lim_{n\to\infty} n^{-1-\delta/2} \sum_{k=1}^{n} E[|u_k|^{2+\delta} || F_{k-1}] = 0$ w.p.1

Then $n^{-1/2} \sum_{k=1}^{n} u_k \xrightarrow{L} N(0,\beta^2)$ as $n \to \infty$.

Remark 3: The result just stated still holds if the limits in condition (2) hold in probability.

Corollary 2.5.1

Assume the u's depend on some parameter $\theta = (\theta_1, \cdots, \theta_m)$ $\theta \in \Theta$, and that for all $\theta_0 \in \Theta$, there exists a positive number $r(\theta_0)$ such that the limits in condition (2) of Theorem 2.5 hold uniformly for all θ such that $|\theta - \theta_0| \leq r$.

Then $n^{-1/2} \sum_{k=1}^{n} u_k \xrightarrow{L} N(0,\beta^2(\theta))$ holds uniformly as $n \to \infty$ for $|\theta - \theta_0| \leq r$.

Proof In order to establish the result just claimed, we merely need to follow Billingsley's proof of Theorem 2.5 and make sure that the relevant limiting statements hold uniformly for $|\theta - \theta_0| \leq r$.

These steps do not involve any new idea, so that we shall only sketch where they occur.

1. Lemma 9.2 of [4] pp 53-55 does not involve any limiting argument and therefore hold for all $\theta \in \Theta$.

2. In order that the conclusion of Lemma 9.3 of [4] pp 55-58 holds uniformly for $|\theta - \theta_0| \leq r$ we assume the following:

let $\sigma_n^2 = E[u_n^2 | F_{n-1}]$, $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$,

for $t > 0$ let $m_t = \min \{n : s_n^2 \geq t\}$,

and $\gamma_n^{2+\delta} = E[|u_n|^{2+\delta} | F_{n-1}]$

then we suppose: (a) $\lim_{n\to\infty} \sum_{k=1}^{n} \sigma_k^2(\theta) = \infty$

(b) $\lim_{n\to\infty} n^{-1-\delta/2} \sum_{k=1}^{n} \gamma_k^{2+\delta} = 0$

both holding in probability and uniformly for $|\theta - \theta_0| \leq r$

(c) $\sup_t m_t(\theta)/t$ is stochastically bounded* for $|\theta - \theta_0| \leq r$

(d) $\lim_{n\to\infty} s_n^{-2-\delta} \sum_{k=1}^{n} \gamma_k^{2+\delta} \leq 1$ in probability and uniformly for $|\theta - \theta_0| \leq r$.

3. Finally the proof of Theorem 9.1 itself (see [4] pp 58-61) assuming (a) $\lim_{n\to\infty} s_n^2/n = \beta^2(\theta)$

(b) $\lim_{n\to\infty} n^{-1-\delta/2} \sum_{k=1}^{n} \gamma_k^{2+\delta} = 0$

both statements true in probability and uniformly for $|\theta - \theta_0| \leq r$.

Then it can be shown that:

- $\lim_{t\to\infty} m_t(\theta)/t = \beta^{-2}(\theta)$
- $\lim_{n\to\infty} s_n^{-2-\delta} \sum_{k=1}^{n} \gamma_k^{2+\delta} < 1$

both limits true in probability and uniformly for $|\theta - \theta_0| \leq r$ which in turn imply the conclusion of the corollary.

Next, we recall the basic result given in [32] by Weiss and Wolfowitz:

---

* For definition see [12] page 247.

Theorem 2.6

Let $\{Z_n\}$ be a maximum probability estimator, which satisfies the following conditions for some sequence

$$\mathbf{k}(n) = (k_1(n),\cdots,k_m(n)) \quad (n = 1,2,\cdots)$$

where $k_i(n) \to \infty$ as $n \to \infty$ $(i = 1,..,m)$

and for any $h > 0$.

As $n \to \infty$ we have:

(1) Lim $P[k(n)(Z_n - \theta) \in R|\theta] = \beta$, say, uniformly for all $\theta \in H$ where

$$H = \{\theta \,|\, |k(n)(\theta - \theta_0)| \leq h\}^* \ , \ \theta_0 \in \Theta$$

(2) as $n \to \infty$ and $M \to \infty$ we have

Lim $P[|k(n)(Z_n - \theta)| < M|\theta] = 1$ uniformly for all θ in some neighborhood of θ_0 .

Let $\{T_n\}$ be any estimator such that, as $n \to \infty$,

$$\text{Lim } \{P[k(n)(T_n - \theta) \in R|\theta] - P[k(n)(T_n - \theta_0) \in R|\theta_0]\} = 0$$

uniformly for all $\theta \in H$.

Then $\beta \geq \overline{\lim}\, P[k(n)(T_n - \theta_0) \in R|\theta_0]$

We now turn to an important result which will frequently be used in the sequel.

---

* If $v = (v_1,\cdots,v_m)$ then $|v|$ stands for $\max_i |v_i|$.

Theorem 2.7

Assume the same conditions (1) - (6) given in Theorem 2.4, but now all limits are reached uniformly for $|\theta - \theta_0| \leq r$ for some positive number r .

(7) $\sigma_{uv}(\theta)$ is a continuous function of $\theta, \theta \in \Theta$ $u,v = 1,\cdots,m$

Then The maximum likelihood estimator of θ_0 is asymptotically efficient.

Proof

1 Conditions (1) - (6) , with the additional restriction of uniform convergence imply that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0,\sigma^{-1}(\theta)) \quad \text{uniformly for} \quad |\theta - \theta_0| \leq r \quad \text{as} \quad n \to \infty \ .$$

The proof of the statement just made, is not at all involved. For sake of conciseness, we shall give the argument for the one-parameter case. Recall that we have by (70)

$$(70) \qquad \sqrt{n}(\tilde{d}_n - \theta) = \frac{\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\partial}{\partial\theta} g(x_{k-1},x_k|\theta) + \frac{\alpha}{6} \frac{r^2(n)}{\sqrt{n}} \sum_{k=1}^{n} G(x_{k-1},x_k)}{-\frac{1}{n} \sum_{k=1}^{n} \frac{\partial^2}{\partial\theta^2} g(x_{k-1},x_k|\theta) - \frac{1}{n}\frac{\alpha}{2}(\tilde{d}_n - \theta) \sum_{k=1}^{n} G(x_{k-1},x_k)}$$

. by Corollary 2.5.1 we have

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\partial}{\partial\theta} g(x_{k-1}, x_k | \theta) \xrightarrow{L} N(0, \sigma^2(\theta)) \quad \text{as} \quad n \to \infty \quad \text{uniformly for}$$

$$|\theta - \theta_0| \leq r$$

· the second term in the numerator does not depend on θ

· the first term in the denominator converges uniformly (by assumption) to $\sigma^2(\theta)$ for $|\theta - \theta_0| \leq r$

· finally, the second term in the denominator converges uniformly to zero, since all assumptions needed for consistency hold uniformly for $|\theta - \theta_0| \leq r$.

By the same argument it can be shown that, as $n \to \infty$

$$\sqrt{n}(\hat{d}_n - \theta) \xrightarrow{L} N(0, \sigma^{-1}(\theta)) \quad \text{uniformly for} \quad |\theta - \theta_0| \leq r \ .$$

2. Let $f(z|\theta) = (2\pi)^{-\frac{1}{2}m} ||\sigma^{-1}(\theta)||^{-1/2} \exp\left\{-\frac{1}{2} z'[\sigma(\theta)]z\right\}$

where z is an m-vector

and $\sigma(\theta) \equiv (\sigma_{uv})$ $\quad u,v = 1, \cdots, m$.

Let R_1 be a compact set in m-space such that

$$\int_{R_1} f(z|\theta_0)dz > 1 - \frac{\varepsilon}{17} \qquad \varepsilon > 0 \text{ arbitrarily small}$$

Let $S(\theta) = \max_{x \in R_1} |f(x|\theta) - f(x|\theta_0)|$

Since the correlation matrix is assumed to be a continuous function of θ , so is $S(\theta)$.

Let $v = \int_{R_1} dz$

we choose a number r', $(0 < r' \leq r)$ small enough so that

$$S(\theta) < \frac{\varepsilon}{17v} \quad \text{for} \quad |\theta - \theta_0| \leq r'$$

Let R_3 be a set in m-space

- suppose $R_3 \subseteq R_1$

(75) then $\int_{R_3} |f(z|\theta) - f(z|\theta_0)| dz \leq \frac{\varepsilon}{17v} \cdot v' \leq \frac{\varepsilon}{17}$

- suppose $R_3 \cap R_1 = \phi$

$$\int_{R_1} f(z|\theta) dz > 1 - \frac{2\varepsilon}{17}$$

(80) $\rightarrow \int_{R_3} |f(z|\theta) - f(z|\theta_0)| dz < \frac{3\varepsilon}{17}$

finally (75) and (80) together imply that for all R_3

$$\int_{R_3} |f(z|\theta) - f(z|\theta_0)| dz < \frac{4\varepsilon}{17} \quad \text{for} \quad |\theta - \theta_0| \leq r'$$

Now, by part 1 of the proof, we know that for all θ such that $|\theta - \theta_0| \leq r' \leq r$ there exists n_0 such that $n > n_0$ implies

(85) $$\left| P[k(n)(Z_n - \theta) \in R | \theta] - \int_R f(z|\theta) dz \right| < \frac{\varepsilon}{17}$$

$$(90)\quad \left|P[k(n)(Z_n - \theta) \in R|\theta] - P[k(n)(Z_n - \theta_0) \in R|\theta_0]\right| =$$

$$\left|P[k(n)(Z_n - \theta) \in R|\theta] - \int_R f(z|\theta)dz\right.$$

$$-P[k(n)(Z_n - \theta_0) \in R|\theta_0] + \int_R f(z|\theta_0)dz$$

$$\left. + \int_R f(z|\theta)dz - \int_R f(z|\theta_0)dz\right| \leq \frac{\varepsilon}{17} + \frac{\varepsilon}{17} + \frac{4\varepsilon}{17}$$

let $n_1 = \{\min n: k(n) > \frac{h}{r'}\} < \infty$

then for $n \geq \max(n_0, n_1)$, condition (1) of Theorem 2.6 is satisfied.

Similarly it is seen that condition (2) is also satisfied.

Since we have seen that the MPE and MLE of θ are uniformly asymptotically equivalent for $|\theta - \theta_0| \leq r$. This concludes the proof.

2.5 The Impact of the Invariance Property of the Method of Maximum Likelihood Estimation on the Efficiency of a Class of Estimators.

Let θ be an m-vector of unknown parameters and let $g(\cdot)$ be an invertible mapping from m-space to m-space. The invariance property is that if $\hat{\theta}_n$ is a MLE of θ, then $g(\hat{\theta}_n)$ is a MLE of $g(\theta)$. Now if $g(\cdot)$ is a mapping from m-space to m'-space, $(m' < m)$, Zehna [34] and Berk [3] both propose to use $g(\hat{\theta}_n)$ as a MLE of $g(\theta)$. Zehna proposes to use $g(\hat{\theta}_n)$ since, if with $g(\theta)$ one associates the largest of the likelihoods of those θ' such that $g(\theta') = g(\theta)$, this "induced

likelihood function" is maximized at $g(\hat{\theta}_n)$. ($g(\hat{\theta}_n)$ could actually be a minimum (see [9] pp 70-71)) Berk proposes to use $g(\hat{\theta}_n)$ since, if one simply adjoins to $g(\theta)$ another function $h(\theta)$ so that the mapping

$$\theta \longrightarrow [g(\theta), h\ \theta)]$$

is 1 - 1 , then $[g(\hat{\theta}_n), h(\hat{\theta}_n)]$ is the MLE of $[g(\theta),h(\theta)]$.
The addition of $h(\theta)$ is only aimed at preserving the status of $g(\hat{\theta}_n)$ as a MLE.

Corollary 2.7.1

Consider the mapping

$$(\theta_1,\cdots,\theta_m) \longrightarrow \{\psi_1(\theta_1,\cdots,\theta_m),\cdots,\psi_{m'}(\theta_1,\cdots,\theta_m)\}$$

from m-space to m'-space ($m' \leq m$)

(1) If $m' = m$ assume the mapping is 1 - 1 .

If $m' < m$ assume there exist functions

$\psi_{m'+1}(\theta_1,\cdots,\theta_m),\cdots,\psi_m(\theta_1,\cdots,\theta_m)$ such that the mapping

$\theta = (\theta_1,\cdots,\theta_m) <-> \psi = [\psi_1(\theta_1,\cdots,\theta_m),\cdots,\psi_m(\theta_1,\cdots,\theta_m)]$ is 1 - 1 .

(2) Moreover we assume that:

· the conditions given in Theorem 2.7 hold and that

· $\frac{\partial\theta_i}{\partial\psi_j}$, $\frac{\partial^2\theta_i}{\partial\psi_j\partial\psi_k}$, $\frac{\partial^3\theta_i}{\partial\psi_j\partial\psi_k\partial\psi_l}$ exist and are continuous throughout

Θ $i,j,k,l = 1,\cdots,m$

Then $\{\psi(\hat{\theta}_n)\}$ is an asymptotically efficient estimator of $\psi(\theta)$.

Proof

Let $\overline{K_n}(x|\psi) = \prod_{i=1}^{n} \overline{f}(x_{i-1},x_i|\psi) \cdot h_n(x_0,\cdots,x_n)$

and $\overline{g}(x_{i-1},x_i|\psi) = \text{Log } \overline{f}(x_{i-1},x_i|\psi)$

It is clear that:

$$\frac{\partial \overline{g}}{\partial \psi_j} , \frac{\partial^2 \overline{g}}{\partial \psi_j \partial \psi_k} , \frac{\partial^3 \overline{g}}{\partial \psi_j \partial \psi_k \partial \psi_l}$$

are linear functions of $\frac{\partial g}{\partial \theta_i}$ $i = 1,\cdots,m$ and partial derivatives of higher order. For example:

$$\frac{\partial \overline{g}}{\partial \psi_i} = \sum_{j=1}^{m} \frac{\partial g}{\partial \theta_j} \cdot \frac{\partial \theta_j}{\partial \psi_i}$$

(where the partial derivatives such as $\frac{\partial \theta_j}{\partial \psi_i}$ are not random variables).

We see that $\psi(\hat{\theta}_n)$ is a maximum likelihood estimator in its own rights, and that all assumptions of Theorem 2.7 are fulfilled and accordingly enjoys the property given in the concluding statement of Theorem 2.7.

Finally it seems reasonable to consider that in the case $m' < m$, $\{\psi_1(\hat{\theta}_n),\cdots,\psi_{m'}(\hat{\theta}_n)\}$ is a vector of asymptotically efficient estimators of $\{\psi_1(\theta),\cdots,\psi_{m'}(\theta)\}$.

CHAPTER III

APPLICATIONS TO THE ESTIMATION PROBLEM FOR MARKOV CHAINS

3.1 Estimation of parameters for strongly ergodic, irreducible, non-homogeneous, finite Markov chains.

Consider a non-homogeneous Markov chain with N states $(N < \infty)$ and with successive transition matrices $P_1, P_2, \cdots$.

Let $H_n = \prod_{i=1}^{n} P_i$, $P_i = (P_{jk}^{(i)})$

where $P_{jk}^{(i)}$ may depend on $\theta = (\theta_1, \cdots, \theta_m)$ a vector of unknown parameters, we are willing to estimate.

Assume the chain is ergodic in the strong sense (see [16], [17], [20]).

let $S = \lim_{n\to\infty} H_n$ then $h_{ij}^{(n)} \to \pi_j$ as $n \to \infty$ independently of i.

Since the $h_{ij}^{(n)}$'s are finite in number (for each n), there exists n_0 such that $n > n_0 \to |H_n - S| < (\xi)$ where (ξ) is an $N \times N$ matrix of ε's > 0 arbitrarily small.

Given any initial distribution $a = (a_1, \cdots, a_N)$, the absolute probability vector is

$$a^{(n)} = a \cdot H_n$$

or for the i^{th} component

$$a_i^{(n)} = \sum_{j=1}^{N} a_j \cdot h_{ji}^{(n)}$$

$$|a_i^{(n)} - \pi_i| = |\sum_{j=1}^{N} a_j h_{ji}^{(n)} - \pi_i| = |\sum_{j=1}^{N} a_j h_{ji}^{(n)} - \sum_{j=1}^{N} a_j \pi_i| \leq$$

$$\sum_{j=1}^{N} a_j |h_{ji}^{(n)} - \pi_i| \leq \varepsilon \quad \text{as soon as} \quad n > n_0 \ .$$

Lemma 3.1

Let P_n be the n^{th} transition matrix of an $N \times N$ non-homogeneous, strongly ergodic finite Markov chain.

Let $H_n = \prod_{i=1}^{n} P_i$, $\pi = \lim_{n\to\infty} a \cdot H_n$ for all initial distributions a .

Let $Y_t = f(X_t, X_{t+1})$ be a bounded function of the random variables X_t, X_{t+1} (i.e.: $|Y_t| < M$ for some fixed M) .

Then $\mathrm{Cov}\,(Y_t, Y_{t+n}) \to 0$ as $n \to \infty$.

Proof for all $n > n_0(\varepsilon)$ $|a_i^{(n)} - \pi_i| \leq \varepsilon$

$$E[Y_t Y_{t+n+1}] = E[E\{Y_t Y_{t+n+1} | | X_t, X_{t+1}\}]$$

$$= E[Y_t \{E[Y_{t+n+1} | | X_t, X_{t+1}]\}]$$

but $|E[Y_{t+n+1}|X_{t+1}] - E[Y_{t+n+1}]| \leq 2 \cdot \varepsilon \cdot M$

$$E[Y_t Y_{t+n+1}] = E[Y_t\{E[Y_{t+n+1}] + \varepsilon'\}] \qquad |\varepsilon'| \leq 2 \cdot \varepsilon \cdot M$$

$$= E[Y_t] \cdot E[Y_{t+n+1}] + \varepsilon' E[Y_t]$$

finally $|E[Y_t Y_{t+n+1}] - E[Y_t] \cdot E[Y_{t+n+1}]| \leq 2 \cdot \varepsilon \cdot M \cdot M$

which can be made as small as desired.

Remark 4: It is clear that if $f(\cdot)$ is a function only of X_t, then the same conclusion holds.

Corollary 3.1

Assume the same conditions given in Lemma 3.1.

Then the weak law of large numbers apply and we obtain

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{Y_i - E[Y_i]}{n} = 0 \qquad \text{(in probability)}$$

Proof We choose n_0 such that $|\text{cov}(Y_t, Y_{t+n})| < \varepsilon$ for $n \geq n_0(\varepsilon)$.

$$\text{Var}\left[\sum_{i=i}^{n} Y_i\right] = \sum_{i=1}^{n} \text{Var}[Y_i] + 2 \sum_{1\leq i<j\leq n} \sum \text{Cov}(Y_i, Y_j)$$

$$\text{Var}\left[n^{-1}\sum_{i=1}^{n} Y_i\right] \leq \frac{nM^2+2n\cdot n_0(\varepsilon)\cdot M^2+n^2\varepsilon}{n^2}$$

and clearly $\lim_{n\to\infty} \text{Var}\left[n^{-1}\sum_{i=1}^{n} Y_i\right] = 0$

the result follows from Tchebycheff's inequality (see [21]).

Theorem 3.1

Using the same notations as before, we assume:

(1) $S(\theta) = \lim_{n\to\infty} H_n(\theta)$ holds uniformly for $|\theta - \theta_0| \leq r$ for some positive number $r,\cdots$ and that $\pi(\theta)$, the limiting probability vector is a continuous vector function of $\theta, \theta \in \Theta$,

(2) for all i, j, k, $P_{jk}^{(i)}$ has continuous partial derivatives w.r.t. θ, up to the third order.

Letting $g_{jk}^{(i)} = \text{Log } P_{jk}^{(i)}(\theta)$ we assume that for all (j,k) pairs $(j, k = 1,2,\cdots,N)$ and for any u,v,w $(u,v,w = 1,\cdots,m)$, the sequences

$$\left\{\frac{\partial g_{jk}^{(i)}}{\partial\theta_u}\right\}, \left\{\frac{\partial^2 g_{jk}^{(i)}}{\partial\theta_u\partial\theta_v}\right\}, \left\{\frac{\partial^3 g_{jk}^{(i)}}{\partial\theta_u\partial\theta_v\partial\theta_w}\right\} \quad i = 1,2,\cdots$$

are uniformly bounded in some neighborhood of θ_0,

(3) let $G_{jk}^{(i)} = \sup_{\theta \in N_\varepsilon(\theta_0)} \left| \frac{\partial^3 g(i)}{\partial\theta_u \partial\theta_v \partial\theta_w} \right|$

where $N_\varepsilon(\theta_0)$ is some neighborhood of θ_0 , then we assume that:

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E\left[\frac{\partial^2 g(i)}{\partial\theta_u \partial\theta_v} (x_{i-1}, x_i | \theta) \,||\, x_{i-1} = k \right] \equiv - \lambda_{uv}^{(k)}(\theta)$$

and $$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E[G^{(i)}(x_{i-1}, x_i) \,||\, x_{i-1} = k] \equiv M_k$$

exist, for $k = 1, \cdots, N$. Moreover, the first limit holds uniformly for $|\theta - \theta_0| \leq r$ and is continuous in θ ,

(4) for any $\theta \in \Theta$ there exists a neighborhood N_1 of θ such that for all i,j,k

$$E\left[\sup_{\theta' \in N_1} \left| \frac{\partial P_{jk(\theta')}^{(i)}}{\partial\theta_u} \right| \,\Big|\Big|\, x_{i-1} \right] < \infty \quad u = 1, \ldots, m$$

$$E\left[\sup_{\theta' \in N_1} \left| \frac{\partial^2 P_{jk(\theta')}^{(i)}}{\partial\theta_u \partial\theta_v} \right| \,\Big|\Big|\, x_{i-1} \right] < \infty \quad u,v = 1, \cdots, m$$

(5) let $\sigma_{uv}(\theta) = (\pi_1, \cdots, \pi_N) \cdot (\lambda_{uv}^{(1)}, \cdots, \lambda_{uv}^{(N)})^T$

then $(\sigma_{uv}(\theta))$ is a positive definite $(N \times N)$ matrix.

Then the maximum likelihood estimator of θ is:

(1) consistent

(2) asymptotically normally distributed, i.e.:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0,\sigma^{-1}(\theta_0)) \quad \text{as} \quad n \to \infty$$

where $\sigma(\theta_0) = (\sigma_{uv}(\theta_0))$

(3) asymptotically efficient .

Proof

We simply verify that all conditions, required for Theorem 2.7 to hold,are fulfilled:

(a) clearly the likelihood function $L(\theta) = \prod_{i=1}^{n} p_{jk}^{(i)}(\theta)$

fits in the general format proposed

$$K_n(x|\theta) = \prod_{i=1}^{n} f(x_{i-1},x_i|\theta) \cdot h_n(x_0,\cdots,x_n)$$

and by assumption (2), $L(\theta) \in L_\theta$.

(b) By assumption (1) and (2) we can apply Corollary 3.1

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial g^{(i)}}{\partial \theta_u}(x_{i-1},x_i|\theta_0) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E\left[\frac{\partial g^{(i)}}{\partial \theta_u}(x_{i-1},x_i|\theta_0)\right]$$

(in probability)

but $E\left[\frac{\partial g^{(i)}}{\partial \theta_u}(x_{i-1},x_i|\theta_0)\right] = 0$ w.p.1 so that the condition

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial g^{(i)}}{\partial \theta_u}(x_{i-1},x_i|\theta_0) = 0 \qquad \text{(in probability)}$$

is satisfied for $\theta = \theta_0$

Now since the limit in assumption (1) holds uniformly for $|\theta - \theta_0| \leq r$, this implies that both Lemma 3.1 and Corollary 3.1 hold uniformly over the same range of θ , and so does the above limit.

(c) repeating the same argument it is now clear that

$$-\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} G^{(i)}(x_{i-1},x_i) = (\pi_1,\cdots,\pi_N)\cdot(M_1,\cdots,M_N)^T$$

(in probability)

$$-\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial^2 g^{(i)}}{\partial\theta_u\partial\theta_v}(x_{i-1},x_i|\theta) = -\sigma_{uv}(\theta) \quad \text{(in probability)}$$

where $\sigma_{uv}(\theta)$ is continuous in θ since all π_i's and $\lambda_{uv}^{(i)}$'s are

$$-\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E\left[\frac{\partial g^{(i)}}{\partial\theta_u}(x_{i-1},x_i|\theta)\cdot\frac{\partial g^{(i)}}{\partial\theta_v}(x_{i-1},x_i|\theta)\,||\,x_{i-1}\right] = \sigma_{uv}(\theta)$$

(in probability) .

which follows from the previous line and assumption (4) which implies

$$E\left[\frac{\partial^2 g^{(i)}}{\partial\theta_u \partial\theta_v}(x_{i-1},x_i|\theta)\,||\,x_{i-1}\right]$$

$$= -E\left[\frac{\partial g^{(i)}}{\partial\theta_u}(x_{i-1},x_i|\theta)\cdot\frac{\partial g^{(i)}}{\partial\theta_v}(x_{i-1},x_i|\theta)\,||\,x_{i-1}\right]$$

$$-\lim_{n\to\infty} n^{-1-\delta/2}\sum_{k=1}^{n} E\left[\left|\frac{\partial g^{(k)}}{\partial\theta_u}(x_{k-1},x_k)\right|^{2+\delta}\,\Big|\,\Big|x_{k-1}\right] = 0 \quad \text{(in probability)}$$

which follows from (2).

All limits holding uniformly for $|\theta - \theta_0| \leq r$, this concludes the proof.

Example:

Let $P_n = \begin{pmatrix} 1 - \cos\theta \cdot \cos^2 n & \cos^2 n \cdot \cos\theta \\ \sin\theta \cdot \cos^2 n & 1 - \sin\theta \cdot \cos^2 n \end{pmatrix}$

where θ is an unknown parameter $(0 < \theta < \frac{\pi}{2})$, which we want to estimate.

We now expand P_n in terms of its eigenvalues (see [2] pp 24-28) which are solutions of

$$||\lambda I - P_n|| = 0$$

where I is the (2×2) identity matrix.

This gives us:

$$\lambda_1 = 1 \ , \ \lambda_2^{(n)} = 1 - \cos^2 n \cdot (\sin\theta + \cos\theta)$$

and clearly $|\lambda_2^{(n)}| < 1$, for all n .

P_n can now be rewritten as:

$$P_n = \begin{pmatrix} \frac{\sin\theta}{\sin\theta+\cos\theta} & \frac{\cos\theta}{\sin\theta+\cos\theta} \\ \frac{\sin\theta}{\sin\theta+\cos\theta} & \frac{\cos\theta}{\sin\theta+\cos\theta} \end{pmatrix} + \lambda_2^{(n)} \begin{pmatrix} \frac{\cos\theta}{\sin\theta+\cos\theta} & \frac{-\cos\theta}{\sin\theta+\cos\theta} \\ \frac{-\sin\theta}{\sin\theta+\cos\theta} & \frac{\sin\theta}{\sin\theta+\cos\theta} \end{pmatrix}$$

or in short

$P_n = Q_1 + \lambda_2^{(n)} Q_2$, where we observe that $Q_1 \cdot Q_1 = Q_1$, $Q_2 \cdot Q_2 = Q_2$, $Q_1 \cdot Q_2 = Q_2 \cdot Q_1 = 0$.

Clearly the observations just made imply:

$$H_n = Q_1 + \prod_{i=1}^{n} \lambda_2^{(i)} \cdot Q_2$$

Let us look at $\lim_{n\to\infty} \prod_{i=1}^{n} \lambda_2^{(i)}$:

$$1 < \sin\theta + \cos\theta \leq \sqrt{2}$$

Given any $\delta > 0$, arbitrarily small we can choose a positive integer N , large enough, and a small positive number ε such that for $n > N$, $(1 - \varepsilon)^n < \delta$. We now pick n_0 , positive integer so that $\cos^2 i > \varepsilon$ at least N times $(i = 1,\cdots,n_0)$, so that for $n > n_0$, we have:

$$\prod_{i=1}^{n} |\lambda_2^{(i)}| = \prod_{i=1}^{n} |1 - \cos^2 i(\sin\theta + \cos\theta)| < \delta \text{ for all } \theta .$$

Therefore the limiting distribution is:

$$(\pi_1(\theta), \pi_2(\theta)) = \left(\frac{\sin\theta}{\sin\theta+\cos\theta} \quad \frac{\cos\theta}{\sin\theta+\cos\theta} \right)$$

and the limit is reached uniformly for all θ .

Next, it can be verified that assumption (2) holds.

Our next point is to show that (3) holds. For that purpose we need the following fact:

If $f(\cos^2 x)$ is a continuous, bounded function of x , $(0 \leq x \leq 2\pi)$ then

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f(\cos^2 i) \quad \text{exists}$$

Proof

As an immediate consequence of the continuity of $f(\cdot)$ on $[0,2\pi]$, we know that this function is uniformly continuous on the same set.

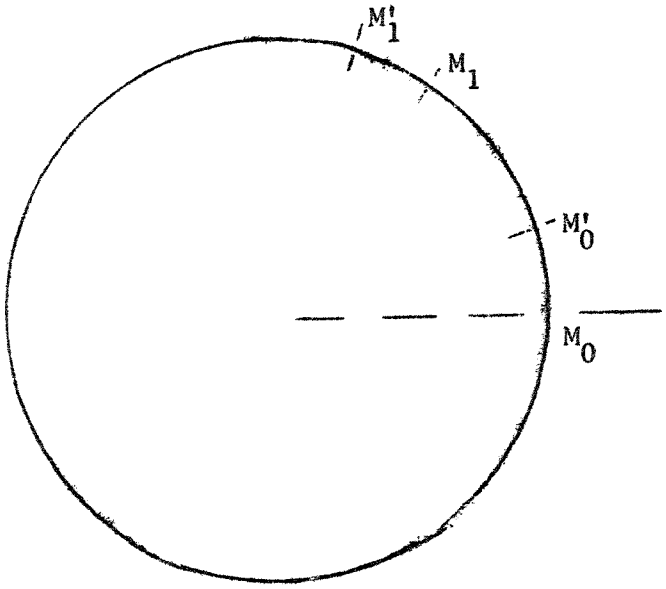
Given $\varepsilon_1 > 0$, arbitratily small we can find $\varepsilon > 0$ suth that

$$|f(\cos^2 x) - f(\cos^2(x+\delta))| < \varepsilon_1$$

for all x on the real line and δ such that $0 \leq \delta \leq \varepsilon$.

We now show that there exist 2 positive integers n_0, k_0 such that

$$0 < n_0 - 2k_0 \cdot \pi < \varepsilon/2 .$$



This is shown by contradiction: let $n = 0,1,\cdots$ and put the points $M_0, M_1, \cdots$ on a circle such that

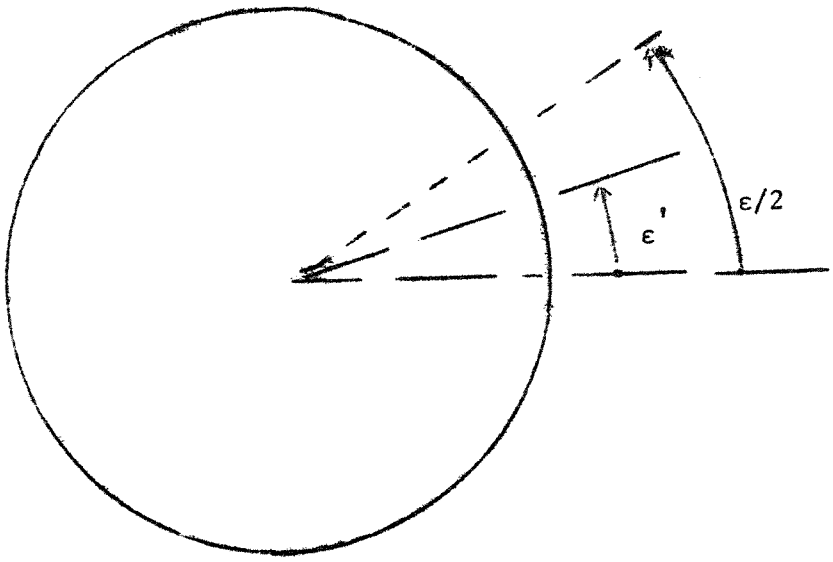
$$\overset{\frown}{M_i M_{i+1}} = 1 \text{ rad } (\text{mod. } 2\pi) \quad ,$$

and $M'_0, M'_1, \cdots$ such that

$$\overset{\frown}{M_i M'_i} = \frac{\varepsilon}{2} \quad (\text{mod. } 2\pi)$$

Let $S_i = \{P \mid 0 \leq M_i P \leq \varepsilon/2 \quad , (\text{mod. } 2\pi)\}$

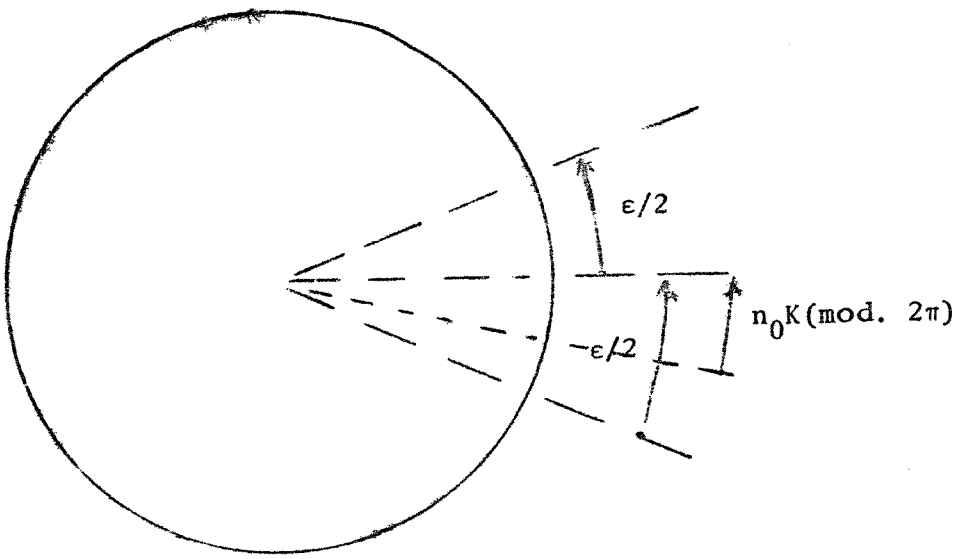
then clearly we have for some i , $M_i \in S_j$, $j < i$, since the circle is finite in length. This implies that $M_{i-j} \in S_0$. Here, we simply have $n_0 = i - j$.



Let an upper bound of $f(\cdot)$ be M

let $\varepsilon' = n_0 - 2k_0 \cdot \pi$, $0 < \varepsilon' < \frac{\varepsilon}{2}$

let $K = \left[\frac{2\pi J}{\varepsilon'}\right]^*$ where J is chosen so that $K > \frac{M}{\varepsilon_2}$, $\varepsilon_2 > 0$ arbitrarily small.



Then we have:

$$2\pi - \varepsilon/2 \leq n_0 K - \left[\frac{n_0 K}{2\pi}\right] \cdot 2\pi \leq 2\pi$$

pick $N \geq 2$ integer such that for the first time, we have

---

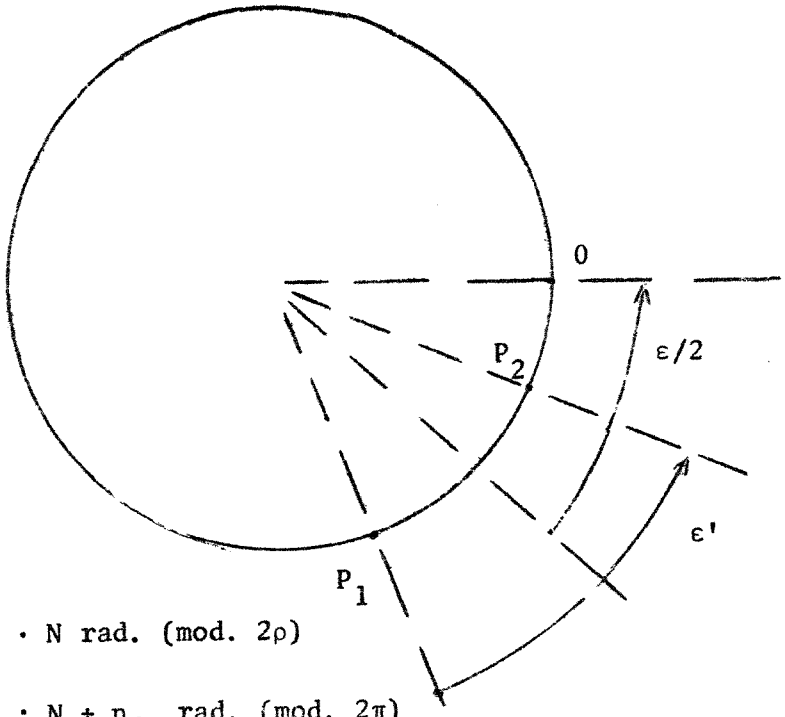
* $[x]$ is the largest integer less or equal to x .

$$2\pi - \varepsilon/2 - \varepsilon' < N \cdot n_0 \cdot K - \left[\frac{Nn_0K}{2\pi}\right] \cdot 2\pi < 2\pi - \varepsilon/2$$

by the inequality $0 < \varepsilon' < \varepsilon/2$ this implies that

$$2\pi - \varepsilon/2 < n_0(KN+1) - \left[\frac{n_0(KN+1)}{2\pi}\right] \cdot 2\pi < 2\pi$$

This, we shall call a "correction of n_0 rad."



$\widetilde{OP_1} = n_0 \cdot K \cdot N$ rad. (mod. 2ρ)

$\widetilde{OP_2} = n_0 \cdot K \cdot N + n_0$ rad. (mod. 2π)

Suppose we have been through Z cycles (a cycle being defined as a rotation of $n_0 \cdot K$ rad.), Y of which requiring a correction of n_0 rad.

let $\sum_{n=1}^{n_0 k} f(\cos^2 n) = n_0 K \mu$

Let μ' be the average for the Z cycles

$$\text{Max } \mu' = \frac{(Z - Y)n_0K(\mu + \varepsilon_1) + Yn_0(K(\mu + \varepsilon_1) + M)}{(Z - Y)n_0K + Yn_0(K + 1)}$$

$$= \frac{Zn_0K(\mu + \varepsilon_1) + Yn_0M}{Zn_0K + Yn_0} \leq \frac{Zn_0K(\mu + \varepsilon_1) + Yn_0M}{Zn_0K}$$

$$= \mu + \varepsilon_1 + \frac{Y}{Z}\frac{M}{K} < \mu + \varepsilon_1 + \varepsilon_2$$

and by a symmetric argument we get

$$|\mu' - \mu| < \varepsilon_1 + \varepsilon_2$$

where $\varepsilon_1 + \varepsilon_2$ can be made as small as desired for any Z positive integer. Also it is clear that for Z large enough, we still have $|\mu' - \mu|$ as small as wanted, when adding a fraction of a cycle. Q.E.D.

If $f(\cos^2 x)$ depends continuously on a parameter θ, then it is clear that the limit of

$$n^{-1} \sum_{i=1}^{n} f(\cos^2 n) = \mu(\theta)$$

holds uniformly for all θ. Also it should be obvious that $\mu(\theta)$ is continuous in θ. Consequently, assumption (3) of Theorem 3.1, is fulfilled, and it is readily seen that so are assumptions (4), (5).

We conclude that the maximum likelihood estimator of θ has the properties of consistency, asymptotic normality, and asymptotic efficiency.

3.2 On Ergodicity of Random Markovian Matrices and Application to Estimation Problems

Let us consider a finite $(N \times N)$ Markov chain, in which P, the transition matrix is a random matrix, i.e.:

$$P_{ij} \text{ is a random variable such that } \sum_{j=1}^{N} P_{ij} = 1$$

Letting $P^{(n)} = (P_{ij}^{(n)})$ be the n^{th} transition matrix.

We assume that:

(1) The $P_{ij}^{(n)}$'s may be either dependent or independent for all $i,j = 1,\cdots,N$

(2) $(P_{ij}^{(n)})$ and $(P_{ij}^{(m)})$ are 2 sets of independent random variables for $n \neq m$

Let $H_n = P^{(1)} \cdot P^{(2)} \cdot \ldots \cdot P^{(n)}$

then the study of H_n, as $n \to \infty$ arises naturally. We show next that for a particular class of random Markovian matrices, a solution is obtained.

Theorem 3.2

Assume:

(1) for any two outcomes $P^{(i)}$ and $P^{(j)}$ we have

$$P^{(i)} \cdot P^{(j)} = P^{(j)} \cdot P^{(i)}$$

(2) P has N distinct characteristic roots with probability 1

(3) P is aperiodic with positive probability.

Then the probability limit of H_n, as $n \to \infty$, exists.

Proof A random Markovian matrix of order N, with N distinct characteristic roots may be written for every outcome as.

$$P^{(i)} = A_0^{(i)} + \lambda_1^{(i)} \cdot A_1^{(i)} + \cdots + \lambda_{N-1}^{(i)} \cdot A_{N-1}^{(i)}$$

· assumptions (1) and (2) imply (see [16] page 74) that we have

$$A_k^{(i)} = A_k^{(j)} = A_k \quad , \text{(say)}$$

$$A_k \cdot A_k = A_k \, , \quad A_k \cdot A_j = 0 \qquad k \neq j \; .$$

- the λ's are of course random variables and $|\lambda_i| \leq 1$ with probability one. Also $\lambda_0 = 1$ w.p.1
- the A's matrices are constant.

· assumption (2) implies that there is only one root $\lambda_0 = 1$ for each outcome.

· Since P is by (3) aperiodic with positive probability, $|\lambda_i|$, $(i = 1, \cdots, N-1)$, is less than one with some probability greater than zero. (See [18] page 101)

So that

$$\prod_{j=1}^{n} \lambda_i^{(j)} \to 0 \text{ , (in probability) as } n \to \infty \qquad i = 1,\cdots,N-1$$

and finally $H_n \to A_0$ (in probability) as $n \to \infty$ Q.E.D.

As a consequence of this, we can show, by the same argument used in the case of non-homogeneous finite Markov chains, that if X_n is the state of the process at stage n, and if we let

$$Y_n = f(X_n) \text{ , } |f(X_n)| < M < \infty$$

then $\text{Cov}\,(Y_t, Y_{t+n}) \to 0$ as $n \to \infty$

$$\sum_{i=1}^{n} \frac{Y_i - E[Y_i]}{n} \to 0 \text{ (in probability) as } n \to \infty$$

Suppose the P_{ij}'s are a set of random variables depending on a finite number of unknown parameters, $\theta = (\theta_1,\cdots,\theta_m)$, that we are willing to estimate. In this respect, we have the following results.

Theorem 3.3

Assume:

(1) $P \lim_{n\to\infty} H_n = H$, (say) exists, uniformly for $|\theta - \theta_0| \leq r$, for

some positive number r, all rows of H being identically $(\pi_1,\cdots,\pi_N)$, where $\pi_i(\theta)$ is a continuous function of θ, $\theta \in \Theta$, $i = 1,\cdots,N$

(2) the Markov chain $\{E(P_{ij}^{(n)})\}$ - where the (i,j)th component is taken to be the expectation of $P_{ij}^{(n)}$ - enjoys the properties (2) to (7) given in Theorem 3.1.

Then the maximum likelihood estimator of θ is

(1) consistent

(2) asymptotically normally distributed when properly normalized

(3) asymptotically efficient.

Proof

At each stage i, we observe the state of the process T_i, but we do not observe the outcome of the P_{ij}'s.

Assume we take n successive observations. To estimate θ we want to maximize:

$$\underset{\{P_{ij}'s\}}{E}\left\{P\left[T_1,T_2,\cdots,T_n \,\middle|\middle|\, (P_{ij}^{(1)}),\cdots,(P_{ij}^{(n)}),\ T_0,\theta\right]\right\}$$

$$= \underset{\{P_{ij}'s\}}{E}\left\{P\left[T_1 \,\middle|\middle|\, (P_{ij}^{(1)}),T_0,\theta\right] \cdot P\left[T_2 \,\middle|\middle|\, (P_{ij}^{(1)}),(P_{ij}^{(2)}),T_0,T_1,\theta\right] \cdots \right.$$

$$\left. \cdot\ P\left[T_n \,\middle|\middle|\, (P_{ij}^{(1)}),(P_{ij}^{(2)}),\cdots,(P_{ij}^{(n)}),T_0,\cdots T_{n-1},\theta\right]\right\}$$

$$\overset{*}{=} E\left\{P\left[T_1 \,||\, (P_{ij}^{(1)}), T_0, \theta\right] \cdot P\left[T_2 \,||\, (P_{ij}^{(2)}), T_1, \theta\right] \cdot \cdots \cdot \right.$$

$$\left. \cdot P\left[T_n \,||\, (P_{ij}^{(n)}), T_{n-1}, \theta\right]\right\}$$

$$\overset{**}{=} \prod_{k=1}^{n} E\left\{P\left[T_k \,||\, (P_{ij}^{(k)}),\ T_{k-1}, \theta\right]\right\} \quad .$$

With this approach we are now back to the conditions of Theorem 3.1, and this concludes the proof.

Example:

Let $P = \begin{pmatrix} 1 - \theta Y & \theta Y \\ Y & 1 - Y \end{pmatrix}$ be the one step transition matrix

$\theta \in (0,1)$, where Y is a random variable with uniform density on $(0,1)$.

We now look for the characteristic roots of P . So that we must solve the determinantal equation

---

* by the Markov property.

** by the independence of $(P_{ij}^{(n)})$, $(P_{ij}^{(m)})$ $m \neq n$ and conditional independence of the T_i's.

$$||P - \lambda I|| = 0 \quad \text{or} \quad \left|\left| \begin{matrix} 1 - \theta Y - \lambda & \theta Y \\ Y & 1-Y-\lambda \end{matrix} \right|\right| = 0$$

we get $\lambda' = 1$

$$\lambda'' = 1 - Y(1 + \theta)$$

Expanding P in terms of its characteristic roots we have

$$P = \begin{pmatrix} \frac{1}{1+\theta} & \frac{\theta}{1+\theta} \\ \frac{1}{1+\theta} & \frac{\theta}{1+\theta} \end{pmatrix} + [1 - Y(1+\theta)] \begin{pmatrix} \frac{\theta}{1+\theta} & \frac{-\theta}{1+\theta} \\ \frac{-1}{1+\theta} & \frac{1}{1+\theta} \end{pmatrix}$$

or in short

$$P = P_1 + \lambda'' P_2 \quad \text{with} \begin{cases} P_i^2 = P_i & i = 1,2 \\ P_i P_j = 0 & i \neq j \end{cases}$$

$$P^{(1)} \cdot P^{(2)} \cdot \ldots \cdot P^{(n)} = P_1 + \left(\prod_{i=1}^{n} \lambda''_i\right) \cdot P_2$$

λ'' is clearly a random variable with $|\lambda''| < 1$. It is easily seen that for $|\theta - \theta_0| \leq r$, for some $r > 0$

$$\lim_{n\to\infty} \prod_{i=1}^{n} \lambda''_i = 0 \quad \text{(in probability), uniformly.}$$

and finally

$P^{(1)} \cdot P^{(2)} \cdot \ldots \cdot P^{(n)} \to P_1$ (in probability), as $n \to \infty$, i.e.:

$$H = \begin{pmatrix} \frac{1}{1+\theta} & \frac{\theta}{1+\theta} \\ \frac{1}{1+\theta} & \frac{\theta}{1+\theta} \end{pmatrix}$$

when observing n successive transitions, one would only record the N_{ij}'s which constitute a set of sufficient statistics.

The MLE of θ is $\dfrac{2N_{12}}{N_{11} + N_{12}}$.

3.3 Estimation of Parameters for a Finite, Homogeneous Markov Chain

3.3.1 Consider the Markov chain with transition matrix

$$P = \begin{pmatrix} P_{11}, P_{12}, \ldots\ldots, P_{1N} \\ P_{21}, \cdots \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ P_{N1}, \ldots\ldots\ldots, P_{NN} \end{pmatrix}$$

with the restrictions $\sum_{j=1}^{N} P_{ij} = 1$, $\quad (i = 1, \cdots, N)$

$$P_{ij} > 0 \ , \quad (i,j = 1, \cdots, N)$$

then P is irreducible, and we know (see [24], chapter 2), that the chain has a unique stationary distribution, coinciding with the limiting distribution $\Pi = (\Pi_1,\cdots,\Pi_N)$, where Π is the solution of:

$$(200)\qquad \begin{cases} \Pi = \Pi P \\ \sum_{i=1}^{N} \Pi_i = 1 \end{cases}$$

The vector of unknown parameters we are trying to estimate is

$$\theta = (P_{11},P_{12},\cdots,P_{1,N-1},P_{21},\cdots,P_{2,N-1},\cdots,P_{N,N-1})$$

Assume θ_0 is the true value of the unknown parameter, then it is clear that

$$\lim_{n\to\infty} P_{ij}^{(n)}(\theta_0) = \Pi_j(\theta_0) \quad .$$

To show that the convergence is uniform for all i,j over some neighborhood of θ_0 we follow an argument given by Doob (see [8] page 173).

> For $\theta = \theta_0$, there is an integer $\nu(\theta_0)$ and a set $J(\theta_0)$ of $N_1(\theta_0)$, $(1 \le N_1(\theta_0) \le N)$, values of j, such that

$$\min_{\substack{1\le i\le N \\ j\in J}} P_{ij}^{(\nu)} = \delta(\theta_0) > 0 \ , \ (0 < \delta(\theta_0) \le \frac{1}{N})$$

Moreover we have

$$\left| P_{ij}^{(n)}(\theta_0) - \Pi_j(\theta_0) \right| \le (1 - N_1 \cdot \delta)^{(n/\nu)} - 1 \tag{205}$$

Consider now a neighborhood of θ_0 such that $P_{ij}(\theta) > 0$, for all $\theta \in N_{\theta_0} = \{\theta: |\theta - \theta_0| \le r\}$. Then obviously $\nu = 1$, $N_1 = N$ and from (205) we can conclude that the convergence is uniform. Consequently, Lemma 3.1 and Corollary 3.1, given in the section on non-homogeneous Markov chains, apply.

Looking at the assumptions of Theorem 3.1, we see that all of them are readily fulfilled.

In the case under consideration it is clear that a set of sufficient statistics is the matrix (n_{ij}) called by Billingsley, (see [5]), transition count, and that the maximum likelihood estimator of P_{ij} is

$$\hat{P}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{N} n_{ij}}$$

Next, we consider the estimation of the Π's and claim that the MLE of Π_i is

$$\hat{\Pi}_i = \frac{\sum_j n_{ij}}{n} + o_i(n)$$

where $n = \sum_{i=1}^{N} \sum_{j=1}^{N} n_{ij}$

In fact by the invariance property of the MLE the $\hat{\Pi}_i$'s are solution of the following system of equations:

$$(210) \quad (\hat{\Pi}_1, \cdots, \hat{\Pi}_N) = (\hat{\Pi}_1, \cdots, \hat{\Pi}_N) \begin{pmatrix} \hat{P}_{11} & \cdots & \hat{P}_{1N} \\ \vdots & & \vdots \\ \hat{P}_{N1} & \cdots & \hat{P}_{NN} \end{pmatrix}$$

$$(220) \quad \sum_{i=1}^{N} \hat{\Pi}_i = 1$$

$$\text{let} \quad \hat{\hat{\Pi}}_i = \frac{\sum_{j=1}^{N} n_{ij}}{n}$$

then it is easily checked that

$$(225) \quad (\hat{\hat{\Pi}}_1, \cdots, \hat{\hat{\Pi}}_N) = (\hat{\hat{\Pi}}_1, \cdots, \hat{\hat{\Pi}}_N) \begin{pmatrix} \hat{P}_{11} & \cdots & \hat{P}_{1N} \\ \vdots & & \vdots \\ \hat{P}_{N1} & \cdots & \hat{P}_{NN} \end{pmatrix} + o(n)$$

$$(230) \qquad \sum_{i=1}^{N} \hat{\hat{\Pi}}_i = 1$$

let $\hat{\Pi}_i - \hat{\hat{\Pi}}_i = x_i$, then from (210), (220), (225),(230) we get:

$$(235) \quad (x_1,\cdots,x_N) = (x_1,\cdots,x_N) \begin{pmatrix} \hat{P}_{11} & \cdots & \hat{P}_{1N} \\ \vdots & & \vdots \\ \hat{P}_{N1} & \cdots & \hat{P}_{NN} \end{pmatrix} + o(n)$$

$$(240) \qquad \sum_{i=1}^{N} x_i = 0$$

from (235) and (240) it is clear that x_i is of order n^{-1}, $(i = 1,\cdots,N)$

so that $\hat{\Pi}_i - \hat{\hat{\Pi}}_i \to 0$ as $n \to \infty$, in probability

$$\sqrt{n}(\hat{\Pi}_i - \hat{\hat{\Pi}}_i) \to 0 \quad \text{as} \quad n \to \infty \text{ , in probability.}$$

This leads us to the following conclusion:

Theorem 3.4

Assume P to be the transition matrix of a finite irreducible Markov chain, where the P_{ij}'s, $(P_{ij} > 0 \; ; \; i,j = 1,\cdots,N)$ are all unknown.

Then (1)* $\hat{P}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{N} n_{ij}}$ is a consistent estimator of P_{ij}

(2)* $\sqrt{n}(\hat{P}_{ij} - P_{ij})$ is asymptotically normally distributed

(3) $\hat{P}_{ij}$ is an asymptotically efficient estimator of P_{ij} .

(4) $\hat{\hat{\Pi}}_i = \frac{\sum_{j=1}^{N} n_{ij}}{n}$ is a consistent estimator of Π_i , the stationary transition probability.

- $\sqrt{n}(\hat{\hat{\Pi}}_i - \Pi_i)$ is asymptotically normally distributed.

- $\hat{\hat{\Pi}}_i$ is an asymptotically efficient estimator of Π_i**.

3.3.2 Consider a finite, irreducible Markov chain where P_{ij} is now assumed to be a function of $\theta = (\theta_1, \cdots, \theta_m)$, a finite vector of unknown parameters. Any P_{ij} may be known, or function of one unknown parameter (example: $P_{ij} = \theta_{ij}$) or else function of several unknown parameters (example: $P_{ij} = \cos^2(\prod_{i=1}^{m} \theta_i)$) .

---

* Results stated by Billingsley in [5] .

** It should be pointed out, here, that the mapping $(P_{ij}) \to (\Pi_i)$ is not 1 - 1 . However, it can be verified that the following mapping fulfills the mild conditions of Corollary 2.7.1

$$(P_{ij})_{\substack{i=1,\cdots,N \\ j=1,\cdots,N-1}} \leftrightarrow \begin{cases} \Pi_i & i = 1,\cdots,N-1 \\ P_{ij} & \begin{cases} i = 2,\cdots,N \\ j = 1,\cdots,N-1 \end{cases} \end{cases}$$

We next show that under mild conditions, the MLE of θ enjoys some desirable large sample properties.

Theorem 3.5

Let $D = \{(i,j): P_{ij}(\theta) > 0$, independently of $\theta \in \Theta\}$,

then we assume that:

(1) each $P_{ij}(\theta)$ has continuous partial derivatives of third order throughout Θ

(2) moreover the $(d \times m)$ matrix

$$\left\{\frac{\partial P_{ij}}{\partial \theta_u}(\theta)\right\} \qquad (i,j) \in D \; ; \quad u = 1,\cdots,m$$

(d being the number of elements in D) , has rank m throughout Θ . For each $\theta \in \Theta$, there is only one ergodic set.

Then the maximum likelihood estimator of $\theta = (\theta_1,\cdots,\theta_m)$, has the following characteristics:

(1) $\hat{\theta}_n$ is consistent

(2) $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normally distributed

(3) $\hat{\theta}_n$ is asymptotically efficient.

The proof will be easy, since it is enough to make sure that the assumptions (1) - (5) of Theorem 3.1 are satisfied.

- by the ergodicity assumption on P , we know that

$$P_{ij}^{(n)}(\theta_0) \to \Pi_j(\theta_0) \quad \text{as} \quad n \to \infty \ .$$

If we follow once more Doob's argument, previously summarized, we see that since the P_{ij}'s are finite in number, we can find a positive number r such that for $|\theta - \theta_0| \leq r$, $N_1(\theta_0)$, $\nu(\theta_0)$, $\delta(\theta_0)$ will work independently of θ . Therefore (205) holds for all θ satisfying $|\theta - \theta_0| \leq r$, and the convergence is uniform for this set of values of θ .

- condition (1), here, is equivalent to condition (2) of Theorem 3.1, since we are now considering an homogeneous Markov chain and that the $P_{ij}(\theta)$'s are finite in number.
- for the same reason, conditions (3) and (4) of Theorem 3.1 are automatically fulfilled as condition (1) of Theorem 3.4, holds.
- finally, we see, following Billingsley's argument (see [4] pp 23-24) that condition (2) of Theorem 3.5 implies that $\{\sigma_{uv}(\theta)\}$ is a positive definite matrix. Q.E.D.

Remark 5: Instead of deriving Theorem 3.5 as a special case of Theorem 3.1, we could have simply followed Billingsley's proof, adding to it the uniformity argument. However, we believe that this approach offers some more generality.

If we are interested estimating the steady state probabilities Π_i's, $(i = 1,\cdots,N)$, we proceed as follows:

(1) suppose $m \geq N - 1$ and that in the case of strict inequality, there exist functions of θ

$$\nu_N(\theta),\cdots,\nu_m(\theta)$$

such that the mapping

$$(\theta_1,\cdots,\theta_m) \leftrightarrow (\Pi_1,\cdots,\Pi_{N-1},\nu_N,\cdots,\nu_m)$$

be $1 - 1$. Then under the mild conditions of Corollary 2.7.1, the MLE of $(\Pi_1,\cdots,\Pi_{N-1})$, $(\hat{\Pi}_1,\cdots,\hat{\Pi}_{N-1})$, which is obtained by solving

$$\hat{\Pi} = \hat{\Pi}\hat{P} \ , \ \sum_{i=1}^{N} \hat{\Pi}_i = 1$$

enjoys the properties of

- consistency
- asymptotic normality, when properly normalized
- asymptotic efficiency

(2) Suppose now $m < N - 1$, then we can estimate simultaneously any m out of the $(N - 1)$ Π_i's . Any such subset enjoys the same properties just given above. The estimators of the remaining Π_i's are immediately determined.

3.4 Estimation of Parameters for Denumerable, Irreducible, Persistent Non-null Markov Chains

Let us consider a denumerable, ergodic Markov chain, whose P_{ij}'s depend on a finite vector of unknown parameters $\theta = (\theta_1, \cdots, \theta_m)$. Billingsley, in his monograph (see [4]), has proved that under a certain set of conditions, the maximum likelihood estimator of $\theta = (\theta_1, \cdots \theta_m)$ is consistent and asymptotically normally distributed, when properly normalized. We shall try here to explain the difficulties met when questioning whether the MLE of θ is asymptotically efficient or not.

It is clearly seen that here, our likelihood function fits into the general format of Theorem 2.7. Also under some regularity conditions

$$g_{ij} = \text{Log } P_{ij}(\theta)$$

will have derivatives as required. However the remaining assumptions of Theorem 2.7 are not easily checked and this is where a more careful examination is needed. For example, we have to see whether

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\partial g(x_{i-1}, x_i|\theta)}{\partial \theta_u}$$

holds uniformly for $|\theta - \theta_0| \leq r$, for some positive r .

We recall that here, we are dealing with infinite dimensional vectors and consequently the results of Lemma 3.1 and Corollary 3.1 do not hold any longer. We also recall that in the finite, homogeneous case, we

used a result on the geometric ergodicity of a Markov chain. We shall next indicate some results in the same direction.

(1) D. G. Kendall, (see [19]),has proved that if a Markov chain is irreducible and aperiodic, and if for some state i , there exists finite numbers M_{ii}, ρ_{ii} such that $M_{ii} \geq 0$, $0 \leq \rho_{ii} < 1$ and the matrix elements $P_{ii}^{(n)}$ satisfy the inequalities

$$(245) \qquad \left| P_{ii}^{(n)} - \Pi_{ii} \right| \leq M_{ii}\rho_{ii}^{n}$$

then for each transition $i - j$, there exist finite numbers M_{ij}, ρ_{ij} ($M_{ij} \geq 0$, $0 \leq \rho_{ij} < 1$) and

$$(250) \qquad \left| P_{ij}^{(n)} - \Pi_{ij} \right| \leq M_{ij}\rho_{ij}^{n}$$

the chain is then said to be geometrically ergodic.

(2) Later on, Vere-Jones (see [30]) investigated the possibility of asserting inequalities such as (245), (250) in which the convergence parameters ρ_{ij} may be replaced by a single parameter independent of i and j .

A way to study $\{P_{ij}^{(n)}\}$ $n = 1,2,\cdots$, is to consider generating functions such as

$$P_{ij}(s) = \sum_{n=0}^{\infty} P_{ij}^{(n)} s^{n} \quad - \quad \text{for all} \quad i, j$$

This is by no means easy, and in fact, does not help getting at the M_{ij}'s, which we know little about. In particular, we would like to find whether $\{M_{ij}\}$ is a bounded double sequence or not. The former case would allow us to proceed forward, as in Lemma 3.1 and its Corrolary.

3.5 A Word on Multiple Markov Chains

Up to now, we have considered statistical estimation for the case of first order Markov chain. Here, we briefly indicate how this can be generalized to multiple Markov chains. Let $\{x_n\}$ be a t^{th} order Markov chain with transition probabilities

$$P_{a_1,\cdots,a_t:a_{t+1}} = P[x_n = a_{t+1} \,||\, x_{n-t} = a_1,\cdots,x_{n-1} = a_t]$$

Problems involving multiple Markov chain are easily reduced to problems about simple ones by a simple device (see [8] page 89 and 185). Here we quote Billingsley (see [5] page 29):

"Consider the process $\{y_m;\ m = 1,2,\cdots\}$ where $y_m = (x_m, x_{m+1},\cdots,x_{m+t-1})$, then $\{y_m\}$ is a first order Markov chain. If x_n can only take a finite number of possible values (say N), then the state space of the chain defined by $\{y_m\}$ consists of N^t different tuples. The transition probabilities being

$$P_{(a_1,\cdots,a_t)(b_1,\cdots,b_t)} = \begin{cases} P_{a_1,\cdots,a_t:b_t} & \text{if } b_i = a_{i+1};\ i = 1,\cdots,t-1 \\ 0 \quad \text{otherwise} & \text{"} \end{cases}$$

Example: Assume $\{x_n\}$ is a 2^{nd} order Markov chain, with 2 possible states. The transition probabilities can be gathered in the following tableau:

	(1)	(2)
(11)	P_{11}	$1 - P_{11}$
(12)	P_{12}	$1 - P_{12}$
(21)	P_{21}	$1 - P_{21}$
(22)	P_{22}	$1 - P_{22}$

We now use the device given by Billingsley and get:

$$P = \begin{array}{c|cccc} & (11) & (12) & (21) & (22) \\ \hline (11) & P_{11} & 1 - P_{11} & 0 & 0 \\ (12) & 0 & 0 & P_{12} & 1 - P_{12} \\ (21) & P_{21} & 1 - P_{21} & 0 & 0 \\ (22) & 0 & 0 & P_{22} & 1 - P_{21} \end{array}$$

If the P_{ij}'s are unknown, the question of their estimation arises naturally and in general it should be clear how to proceed from here, using our previous results.

For the above example, we simply call upon Theorem 3.4.

CHAPTER IV

APPLICATION TO THE ESTIMATION PROBLEM IN THE FIELD OF ECONOMETRICS

4.1 Estimation of unknown parameters in a single equation

4.1.1 Linear regression with stochastic regressors.

Let us consider the following mathematical model

(255) $$Y_t = \beta^T \cdot X_t + \varepsilon_t \qquad t = 0,1,\cdots$$

where (1) $X_t = (X_{t1},\cdots, X_{ts})^T$ is a stationary stochastic process.

(2) $\beta^T = (\beta_1,\cdots,\beta_s)$ is a vector of unknown parameters.

(3) $\{\varepsilon_t\}$ is a sequence of independent, identically distributed random variables with

$$E[\varepsilon_t] = 0 \ , \ E[\varepsilon_t^2] = \sigma^2 \quad \text{(unknown)}.$$

(4) the two sets of random variables $\{X_t\}$ and $\{\varepsilon_t\}$ are independent.

Then it is known that under mild conditions on the $\{X_t\}$ process, the least-squares estimators of β and σ^2 are consistent. (See [15] chapter 6). If in addition, the distribution of ε_t is known, then in some cases, we can derive optimal asymptotic properties of the maximum likelihood estimator.

Theorem 4.1

Consider the model given in (255); in addition to conditions (1) - (4) we assume that:

(5) ε_t is normally distributed $N(0,\sigma^2)$.

(6) there exist numbers c,ρ,δ, $(c \geq 0\ ,\ 0 \leq \rho < 1\ ,\delta > 0)$ such that

$$\left|\mathrm{Cov}\ (X^u_{tj},\ X^v_{t+i,k})\right| \leq c \cdot \rho^i$$

$$j,k = 1,\cdots,s$$

$$i = 0,1,\cdots$$

$$u,v = 1,2,3,4,4+\delta$$

Then the maximum likelihood estimator is

(1) consistent

(2) asymptotically normally distributed

(3) asymptotically efficient.

The proof is easy and we shall only sketch the main ideas. At each $t = 1,2,\cdots,n$ we observe $X_t = (X_{t1},\cdots,X_{ts})^T$ and Y_t . Let $\theta = (\beta_1,\cdots,\beta_s,\sigma^2) \equiv (\theta_1,\cdots,\theta_{s+1})$ the likelihood function is:

(260) $P_\theta[X_1 = x_1] \cdot P_\theta[Y_1 = y_1 | X_1 = x_1] \cdot \cdots\cdots$

$$\cdot\ P_\theta[X_n = x_n | X_i = x_i,\ Y_i = y_i;\ i = 1,\cdots,\ n-1]$$

$$\cdot\ P_\theta[Y_n = y_n |\ X_i = x_i,\ \ i = 1,\cdots,n;\ \ Y_j = y_j,\ j = 1,\cdots,n-1]$$

$$= P[X_1 = x_1] \cdot P_\theta[Y_1 = y_1 | X_1 = x_1] \cdot \cdots \cdot$$

$$\cdot P[X_n = x_n | X_i = x_i, \quad i = 1,\cdots,n-1] \cdot P_\theta[Y_n = y_n | X_n = x_n]$$

$$= P[X_1 = x_1] \cdot \prod_{i=2}^{n} P[X_i = x_i | X_j = x_j, \; j = 1,\cdots,i-1]$$

$$\cdot \prod_{i=1}^{n} P_\theta[Y_i = y_i | X_i = x_i] \qquad \text{where } P_\theta[Y_i = y_i | X_i = x_i]$$

$$= P_\theta[\varepsilon_i = y_i - \beta^T \cdot x_i] = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i - \beta^T \cdot x_i}{\sigma}\right)^2\right]$$

Therefore the Log likelihood function Z_n is given by

$$(265) \quad Z_n = \text{Log } P[X_1 = x_1] + \sum_{i=2}^{n} \text{Log } P[X_i = x_i | X_j = x_j \quad j = 1,\cdots,i-1]$$

$$+ \sum_{i=1}^{n} \text{Log } P_\theta[Y_i = y_i | X_i = x_i]$$

$$(270) \quad \frac{\partial Z_n}{\partial \theta_k} = \sum_{i=1}^{n} \frac{\partial \text{Log } P_\theta[Y_i = y_i | X_i = x_i]}{\partial \theta_k} \qquad k = 1,\cdots, s+1$$

Such expressions as $\dfrac{\partial \text{Log } P_\theta[Y_i = y_i | X_i = x_i]}{\partial \theta_k}$ and derivatives of higher order, will involve polynomial in y_i, x_i at most of degree 2 , so that by our assumption (6), we can see that for example

$$(275) \quad \text{Cov}\left(\frac{\partial \text{Log } P_\theta[Y_i = y_i | X_i = x_i]}{\partial \theta_k}, \frac{\partial \text{Log } P_\theta[Y_{i+t} | X_{i+t} = x_{i+t}]}{\partial \theta_{k'}}\right)$$

is less than $c'\rho^t$. $c' \geq 0$

Moreover we assumed that the $\{X_t\}$ process is stationary and by the assumption (3), we see that the $\{Y_t\}$ process is itself stationary.

We readily see that the weak law of large number is applicable and it is a routine matter to verify that the set of conditions of Theorem 2.7 holds. Q.E.D.

4.1.2 Autoregressive process

Let us consider a stochastic process where the following type of relationship holds.

$$(300) \quad Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t$$

Assume that the disturbance ε_t satisfies another autoregressive process of the q^{th} order, i.e.:

$$(310) \quad \varepsilon_t = \gamma_1 \varepsilon_{t-1} + \cdots\cdots + \gamma_q \varepsilon_{t-q} + u_t$$

where $\{u_t\}$ is a sequence of independent, identically distributed random variables with

(315) $\quad E[u_t] = 0 \quad , \quad E[u_t^2] = \sigma^2$

Then it is seen that (300) can be transformed by means of successive substitutions into a third autoregressive process where the disturbance is temporally independent. By (300) we have:

(320) $\quad \varepsilon_{t-i} = Y_{t-i} - \alpha_0 - \alpha_1 Y_{t-i-1} - \cdots\cdots - \alpha_p Y_{t-i-p}$

substituting (320) into (310) for $i = 1,\ldots,q$, then into (300) we can express Y_t as:

(330) $\quad Y_t = \delta_0 + \delta_1 Y_{t-1} + \cdots + \delta_{p+q} Y_{t-p-q} + u_t$

Assume the distribution of u_t depends on r unknown parameters $(\lambda_1,\cdots\cdots\lambda_r)$. We, then, have a total of $(p + q + r + 1)$ unknown parameters that we wish to estimate.

Let $\theta = (\theta_1,\cdots,\theta_{p+q+r+1}) \equiv (\delta_0,\cdots,\delta_{p+q},\lambda_1,\cdots,\lambda_r)$

Assume we are given $Y_1,\cdots,Y_{p+q}$, then we observe $Y_{p+q+1},\cdots,Y_{p+q+n}$.

The likelihood function is:

$$(340)\quad P_\theta[Y_{p+q+1} = y_{p+q+1} | Y_i = y_i \ , \ i = 1,\cdots, p+q]$$

$$\cdot\ P_\theta[Y_{p+q+2} = y_{p+q+2} | Y_i = y_i \ , \ i = 2,\cdots, p+q+1] \cdot \cdots\cdot$$

$$\cdots \cdot P_\theta[Y_{p+q+n} = y_{p+q+n} | Y_i = y_i \ , \ i = n,\cdots, p+q+n-1]$$

which can also be written as - using (330)

$$(350)\quad P_\theta[u_{p+q+1} = y_{p+q+1} - \delta_0 - \delta_{p+q} y_1] \cdot \cdots\cdot$$

$$\cdots \cdot P_\theta[u_{p+q+n} = y_{p+q+n} - \delta_0 - \cdots\cdot - \delta_{p+q} y_n]$$

We now focus attention on the conditions of Theorem 2.7. Clearly, we are dealing here with a multiple-Markov process and accordingly (340) fits into the general format proposed in Chapter II.

We now turn to the problem of the existence of various limits as our number of observations increases. Assume that all roots of equation (360) in ρ, are less than 1, in absolute value.

$$(360)\quad \rho^{p+q} - \delta_1 \rho^{p+q-1} - \cdots\cdot - \delta_{p+q} = 0$$

Also, we suppose that the $\{Y_t\}$ process has been going on, for a long time. Then, following Mann and Wald's argument (see [22]) we have:

$$(370)\quad \mathrm{Cov}(Y_t,Y_{t+i}) = \sigma^2 \sum_{j=1}^{p+q} \sum_{k=1}^{p+q} \nu_j \nu_k \rho_j^{i+1} \rho_k/(1-\rho_j\rho_k)$$

where ν_j ; $j = 1,\cdots,p+q$, are constants depending on $\delta_1,\cdots,\delta_{p+q}$.

ρ_j; $j = 1,\cdots,p+q$, are the roots of equation (360)

Let θ_0 be the true value of the parameter $(\theta_0 \in \Theta)$. Then whatever $\varepsilon_1 > 0$, arbitrarily small, we can find N so that $i \geq N$ implies $\mathrm{Cov}(Y_t,Y_{t+i}) < \frac{\varepsilon_1}{17}$, when $\theta = \theta_0$.

Also it is clear that $\mathrm{Cov}\,(Y_t,Y_{t+i})$ is a continuous function of $\rho_j (j = 1,\cdots, p+q)$, and σ^2 . Looking at the left hand side of (360), we have a polynomial in ρ of degree $p+q$. As such, it is a continuous function of ρ . For all ε_1 , we can find a small positive number ε_2 , so that if we perturb any δ_i by an amount less than ε_2 in absolute value, any root ρ will change by less than $\varepsilon_1/17$. This leads us to the conclusion that there exists a positive number r , such that for all $|\theta - \theta_0| \leq r$

$$|\mathrm{Cov}\,(Y_t,Y_{t+i})| \leq \varepsilon \text{ as soon as } i \geq N$$

Once more, if we follow Mann and Wald's argument it can be proved that for any positive, fixed integers m,n

$$\text{Cov}\,(Y_t^m, Y_{t+i}^n) \to 0 \quad \text{uniformly for} \quad |\theta - \theta_0| \leq r(m,n)$$

$$r(m,n) > 0 \quad \text{as} \quad i \to \infty$$

Another question of interest concerns the distribution of Y_t as t gets large. Assume $\{u_t\}$ is a sequence of independent, identically distributed random variables such that

$$E[u_t] = 0 \ , \ E[u_t^k] = \mu_k < \infty$$

Mann and Wald argue that Y_t can be expressed as*

$$(380) \quad Y_t = \phi_0(t) + \phi_1(t)\varepsilon_1 + \cdots + \phi_t(t)\varepsilon_t$$

$$(390) \quad \text{where} \quad \phi_0(t) = \sum_{i=1}^{p+q} \mu_i \rho_i^t + \frac{\delta_0}{1 - \sum_{j=1}^{p+q} \delta_j}$$

$$\phi_i(t) = \sum_{k=1}^{p+q} \lambda_k \rho_k^{t-i+1} \qquad 0 < i < t$$

$$\phi_t(t) = 1$$

μ_i, λ_i are constants.

---

* here, we assume $\rho_i \neq \rho_j$ for $i \neq j$, however the result carries over if equation (360) has some multiple roots.

(400) Then $E[Y_t] = \phi_0(t)$

$$\lim_{t\to\infty} E[Y_t] = \frac{\delta_0}{1 - \sum_{j=1}^{p+q} \delta_j}$$

$$(410)\quad E[Y_t - \phi_0(t)]^k = \mu_k \sum_{i=1}^{t} [\phi_i(t)]^k$$

$$= \mu_k \sum_{i=1}^{t} \sum_{j=1}^{p+q} \lambda_j \rho_j^{t-i+1}$$

$$= \mu_k \sum_{j=1}^{p+q} \lambda_j \rho_j \frac{1 - \rho_j^t}{1 - \rho_j}$$

and finally $\lim_{t\to\infty} E\left[(Y_t - E[Y_t])^k\right] = \mu_k \sum_{j=1}^{p+q} \lambda_j \rho_j/(1-\rho_j)$

Looking once more at (380) we see that Y_t is a sum of independent random variables and as such, we know that under mild conditions the limiting distribution of Y_t will be normal with

$$\text{mean} \quad \frac{\delta_0}{1 - \sum_{j=1}^{p+q} \delta_j}$$

and variance $\sigma^2 \sum_{j=1}^{p+q} \lambda_j \rho_j/(1-\rho_j)$

(see for example [33] pp 257-258, [13] pp 202-203, and [6] pp 215-218)

It should now be sufficiently clear, how one would proceed to check whether the assumptions of Theorem 2.7 hold, given the distribution function of the disturbance $\{\varepsilon_t\}$.

Here we could slightly specialize those conditions, but without really getting them simpler. Consequently, we shall skip that step and instead concentrate on the case where u_t is normally distributed $N(0,\sigma^2)$, which nevertheless is a very important particular case.

Theorem 4.2

Let $\{Y_t\}$ be a stochastic process satisfying

$$Y_t = \delta_0 + \delta_1 Y_{t-1} + \cdots\cdots + \delta_{p+q} Y_{t-p-q} + u_t \qquad t = 0,1,\cdots$$

where $\{u_t\}$ is a sequence of independent, normally distributed random variables $N(0,\sigma^2)$.

Assume that all roots of (360) are less than one in absolute value. Then the maximum likelihood estimator of

$$\theta = (\delta_0, \delta_1, \cdots, \delta_{p+q}, \sigma^2)$$

is: (1) consistent*

(2) asymptotically normally distributed*

(3) asymptotically efficient

Proof: The log-likelihood function is:

$$(420)\quad Z_n = \sum_{j=1}^{n} - \frac{1}{2}\frac{(y_{p+q+j} - \delta_0 - \cdots - \delta_{p+q}y_j)^2}{\sigma^2} - n \text{ Log } \sigma\sqrt{2\pi}$$

then using our standard notation

$$g(x_{i-1},x_i|\theta) = \text{Log } f(x_{i-1},x_i|\theta)$$

where $x_i = (y_{p+q+i},y_{p+q+i-1},\cdots,y_i)$

$$\frac{\partial g(x_{i-1},x_i|\theta)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{(y_{p+q+i} - \delta_0 - \cdots - \delta_{p+q}y_i)^2}{\sigma^3}$$

$$\frac{\partial^2 g(x_{i-1},x_i|\theta)}{\partial\sigma^2}\ ,\ \frac{\partial g(x_{i-1},x_i|\theta)}{\partial\delta_0}$$ and other derivatives of interest

are easily computed. It is particularly interesting to notice that the above expressions are at most quadratic functions of the observations

* results given by Mann and Wald in [22] .

and accordingly the covariance of any two such expressions goes to zero as the time log increases and the weak law of large numbers is applicable.

From (380) it is seen that for any t , Y_t is normally distributed and the reader should be convinced that the limiting of Y_t is

$$N(0,\sigma^2 \sum_{j=1}^{p+q} \lambda_j \rho_j/(1-\rho_j)) \quad \text{and that the convergence is uniform}$$

for $|\theta - \theta_0| \leq r'$, $r' > 0$.

Appealing to Theorem 2.7 it is clear that the whole set of assumptions is fulfilled and this concludes the proof.

4.1.3 We now turn to a more general situation where $\{Y_t\}$ is given by

$$(430) \quad Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \beta_1^T \cdot X_t + \varepsilon_t$$

where $\varepsilon_t = \gamma_1 \varepsilon_{t-1} + \cdots + \gamma_q \varepsilon_{t-q} + u_t$.

We shall assume (1) $\{u_t\}$ to be a sequence of independent identically distributed random variables with zero mean.

(2) the $\{X_t\}$ process to be independent of the sequence $\{u_t\}$ and to be looked at, as a set of exogenous random variables.

$$\beta_1^T = (\beta_{11}, \cdots, \beta_{1s})$$

By successive substitutions (430) can be written as:

$$Y_t = \delta_0 + \delta_1 Y_{t-1} + \cdots + \delta_{p+q} Y_{t-p-q} + \beta_1^T \cdot X_t + \beta_2^T \cdot X_{t-1} + \cdots$$

$$\cdots + \beta_{q+1}^T \cdot X_{t-q} + u_t$$

with $\beta_i^T = -\gamma_{i-1} \beta_1^T \qquad i = 2, \cdots, q+1$

Assume the distribution of u_t depends on r unknown parameters $(\lambda_1, \cdots, \lambda_r)$. Then, in the total we have $p + q + (q+1)\, s + r + 1$ unknown parameters. Denote them by:

$$\theta = (\delta_0, \cdots, \delta_{p+q}, \beta_{11}, \cdots, \beta_{1s}, \beta_{21}, \cdots\cdots \beta_{2s}, \cdots \beta_{q+1,s}, \lambda_1, \cdots, \lambda_r)$$

Suppose we are given $X_0, X_1, \cdots, X_q, Y_{1-p}, \cdots, Y_q$, we now observe (X_{q+1}, Y_{q+1}) , $(X_{q+2}, Y_{q+2}), \cdots,$ (X_{q+n}, Y_{q+n}) .

The likelihood function is:

$$(450) \quad P_\theta[X_{q+1} = x_{q+1} | X_0 = x_0, \cdots, X_q = x_q]$$

$$\cdot P_\theta(Y_{q+1} = y_{q+1} | X_i = x_i,\ i = 0, \cdots,\ 1 + q;\ Y_j = y_j\ ,\ j = 1 - p, \cdots, q]$$

$$\cdots \cdot P_\theta[X_{q+n} = x_{q+n} | X_i = x_i,\ i = 0, \cdots, q+n-1]$$

$$\cdot P_\theta[Y_{q+n} = y_{q+n} | X_i = x_i,\ i = 0, \cdots, q+n;\ Y_j = y_j, \quad = 1-p, \cdots, q+n-1]$$

We now point out the fact that

$$P_\theta[X_{q+i}|X_0 = x_0,\cdots,X_{q+i-1} = x_{q+i-1}]$$

does not depend on θ. Consequently the derivative of the log-likelihood function with respect to some unknown parameter will not involve such expression. Dropping the subscript θ of $P_\theta[X_j|\cdots]$ we recognize that the likelihood function fits into our general format.

We now turn to study the conditions under which we can guarantee that the limits in Theorem 2.7 exist. Let us look at the $\{Y_t\}$ process at $t \to \infty$. In most econometrics publications, people assume that $\{Y_t\}$ is a stationary stochastic process and do not analyze under which conditions this is true. They have in fact a good reason for doing so, since this problem is very difficult to answer.

Let us state what are the assumptions which are most of the time justifiable on operational ground in the econometric field.

(460) ||

(1) $\{X_t\}$ is a second-order stationary process such that

$$|\mathrm{Cov}(X_t,X_{t+i})| \leq c \cdot \rho^i \qquad c \geq 0, \quad 0 \leq \rho < 1$$

(2) all roots of the polynomial below are less than one in absolute value

$$\rho^{p+q} - \delta_1\rho^{p+q-1} - \cdots - \delta_{p+q} = 0$$

(470) Let $v_t = \beta_1^T \cdot X_t + \beta_2^T \cdot X_{t-1} + \cdots + \beta_{q+1}^T \cdot X_{t-q} + u_t$

then (440) can be rewritten as

(480) $Y_t = \delta_0 + \delta_1 Y_{t-1} + \cdots + \delta_{p+q} Y_{t-p-q} + v_t$

where obviously $\{v_t\}$ is not a sequence of independent random variables.

Using Mann and Wald's argument Y_t can be expressed as:

(490) $Y_t = \phi_0(t) + \phi_1(t)v_1 + \cdots + \phi_t(t)v_t$

Although we know that the dependence of Y_t and Y_{t+i} is going to wear off rapidly as i increases, no sufficiently general result allows us to draw a conclusion on the limiting distribution of Y_t. For some literature on the theory of limiting distribution for dependent random variables we refer the reader to a paper by R. J. Serfling. (See [28]) If in addition to (460) we assume that $E[X_t^k]$ and $E[X_t^k X_{t+i}^\ell]$ are known for all positive k,ℓ,i then one can compute all moments of Y_t. However the existence of

$$\lim_{t\to\infty} E[Y_t^k]$$

for all positive integers k will not in general be sufficient to insure that the limiting distribution of Y_t as t goes to infinity, exists.

see for example [6] page 176, [13] pp 73-74, [12] page 224 and footnote.

Finally we reach the conclusion that stationarity (or asymptotic stationarity) of the $\{Y_t\}$ process must be part of our assumptions.

Next, we are concerned with the degree of time dependence of such random variables as

$$\frac{\partial g(x_{i-1}, x_i \mid \theta)}{\partial \theta_u} \quad \text{and} \quad \frac{\partial g(x_{i-1}, x_i \mid \theta)}{\partial \theta_v}$$

We have discussed a similar question in the previous paragraph, where no exogenous variable appeared in the Y_t process. We now state a result similar to that of Theorem 4.2.

Theorem 4.3

Let $\{Y_t\}$ be a stochastic process $(t = 0,1,\cdots)$ satisfying

$$Y_t = \delta_0 + \delta_1 Y_{t-1} + \cdots + \delta_{p+q} Y_{t-p-q}$$

$$+ \beta_1^T \cdot X_t + \cdots + \beta_{q+1}^T \cdot X_{t-q} + u_t$$

where

(1) $\{u_t\}$ is a sequence of independent, normally distributed random variable $N(0,\sigma^2)$

(2) $\{X_t\}$ is a stationary stochastic process independent of $\{u_t\}$, and there exist numbers c,ρ,δ $(0 \leq c,\ 0 \leq \rho < 1\ ,\delta > 0)$, such that

$$|\mathrm{Cov}(X^k_{t,m}, X^j_{t+i,n})| \leq c \cdot \rho^i \qquad k,j = 1,2,3,4,4+\delta$$
$$i = 0,1,\cdots$$

and for every m^{th} and n^{th} component

(3) $\{Y_t\}$ is stationary*.

Then the maximum likelihood estimator is:

(1) consistent

(2) asymptotically normally distributed

(3) asymptotically efficient

The proof is easy and follows that of Theorem 4.2.

4.2 Estimation of Unknown Parameters in a Complete System of Linear Equations

4.2.1 System of linear autoregressive stochastic difference equations

Let us consider the following mathematical model of a multivariate stochastic process

$$Y_t = (Y_{t1}, Y_{t2}, \cdots, Y_{tM})^T$$

(500) $$A_1 Y_t + A_2 Y_{t-1} + \cdots + A_{p+1} Y_{t-p} + A_0 = \varepsilon_t \qquad t = 0,1,\cdots$$

where: Y_t, A_0, ε_t are M vectors; $A_1, A_2, \cdots, A_{p+1}$ are $M \times M$ matrices.

---

* It can be seen that this last assumption implies that the second part of (460) is true.

Assume that the disturbance vector ε_t satisfies another autoregressive process:

$$(510) \quad \varepsilon_t = \Gamma_1\varepsilon_{t-1} + \Gamma_2\varepsilon_{t-2} + \cdots + \Gamma_q\varepsilon_{t-q} + u_t$$

where $\Gamma_1,\Gamma_2,\cdots,\Gamma_q$ are $M \times M$ matrices and $\{u_t\}$ is a sequence of independent, identically distributed multivariate random variables with

$$E[u_t] = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \quad E[u_{ti}u_{tj}] = \sigma_{ij} \qquad i,j = 1,\cdots,M$$

or $$E[u_t u_t^T] = \sum_u$$

By successive substitutions (500) can be rewritten as:

$$(520) \quad \Delta_1 Y_t + \Delta_2 Y_{t-1} + \cdots + \Delta_{p+q+1} Y_{t-p-q} + \Delta_0 = u_t$$

$\Delta_1,\cdots,\Delta_{p+q+1}$ are $M \times M$ matrices; Δ_0 is a M vector.

Assume the distribution of u_t depends on r unknown parameters $(\lambda_1,\cdots,\lambda_r)$. Moreover, we suppose that all diagonal elements of Δ_1 are equal to one (this is not a restriction). Each $\Delta_i (i = 0,\cdots,p+q+1)$, contains a number of unknown parameters that we wish to estimate jointly with $(\lambda_1,\cdots,\lambda_r)$.

Furthermore we shall assume that:

(525)
(1) Δ_1 is non-singular

(2) the system in (520) is identified (see [15] chapter 7)

then premultiplying (520) by Δ_1^{-1} we get

$$(530) \quad Y_t = \Pi_0 + \Pi_1 Y_{t-1} + \cdots + \Pi_{p+q} Y_{t-p-q} + v_t$$

where: Π_0 is an M vector

$\Pi_1, \cdots, \Pi_{p+q}$ are $M \times M$ matrices

$\{v_t\}$ is a sequence of independent, identically distributed multivariate random variables such that

$$(540) \quad v_t = \Delta_1^{-1} u_t \qquad E[v_t] = 0$$

$$E[v_t v_t^T] = \Delta_1^{-1} \cdot E[u_t \cdot u_t^T] \cdot (\Delta_1^{-1})^T$$

$$= \Delta_1^{-1} \sum_u (\Delta_1^{-1})^T \equiv \Omega_v$$

The hypothesis of identification allows us to estimate indifferently the set of unknown parameters of the structural form (520) or that of the reduced form (530). We shall concentrate on the last one.

In the next Theorem, we shall assume that u_t is a multivariate normal random variable. Although we lose here some generality, this last assumption is generally accepted in parametric estimation in the econometric field.

Theorem 4.4

Let Y_t be a multivariate stochastic process satisfying (500), where the disturbances are generated by another autoregressive process (510).

We assume that:

(1) $\{u_t\}$ is a sequence of independent, multivariate normal random variables with

$$E[u_t] = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \quad E[u_t u_t^T] = \sum_u$$

(2) the conditions in (525) hold

(550) (3) let $\delta_{ij}(\rho) = \sum_{k=1}^{p+q+1} \delta_{ijk}\rho^{-k}$

where δ_{ijk} is the $(i,j)^{th}$ component of Δ_k

we, then, assume that all roots of the equation

(555) $$||\delta_{ij}(\rho)|| = 0$$

are smaller than 1 in absolute value

Then the maximum likelihood estimator of θ, the vector of unknown parameters of the reduced form (or equivalently - the MLE of ψ, the vector of unknown parameters of the structural form) is:

(1) consistent

(2) asymptotically normally distributed

(3) asymptotically efficient

Proof; given $Y_1, Y_2, \cdots, Y_{p+q}$ we observe $Y_{p+q+1}, \cdots, Y_{p+q+n}$, the likelihood function is:

$$(560) \quad \prod_{t=p+q+1}^{p+q+n} P_\theta[Y_t = y_t \mid Y_{t-1} = y_{t-1}, \cdots, Y_{t-p-q} = y_{t-p-q}]$$

$$= \prod_{t=p+q+1}^{p+q+n} \frac{1}{(\sqrt{2\pi})^p} \cdot \frac{1}{\sqrt{\det \Omega_v}} \exp \left\{ -\frac{1}{2}(y_t - \Pi_0 - \Pi_1 \cdot y_{t-1} - \cdots - \Pi_{p+q} \cdot y_{t-p-q}) \right.$$

$$\left. \cdot \Omega_v^{-1} \cdot (y_t - \Pi_0 - \cdots - \Pi_{p+q} \cdot y_{t-p-q})^T \right\}$$

The Y_t process falls clearly in the class of Markov processes. As in the single equation case we show next that the limiting distribution of Y_t exists.

Let Y_{ti} be the i^{th} component of Y_t $\quad i = 1, \cdots, M$

u_{tj} be the j^{th} component of u_t $\quad j = 1, \cdots, M$

then following once more Mann and Wald's argument Y_{ti} can be expressed as:

$$(570) \quad Y_{ti} = \phi_i(t) + \sum_{\tau=1}^{t} \sum_{j=1}^{M} \phi_{ij\tau}(t) u_{\tau j}$$

$$(575) \quad \text{with (1)} \quad \phi_i(t) = \lambda_1 A_{1i} \rho_1^t + \cdots + \lambda_\nu A_{\nu i} \rho_\nu^t + c_i \ *$$

where the ρ_i are the solution of equation (555), ν being the total number of roots. The A_{ji}'s and λ_j's are constants.

$$(580) \qquad (2) \quad \phi_{ij\tau}(t) = \lambda_{ij} A_{1i} \rho_1^{t-\tau+1} + \cdots + \lambda_{\nu j} A_{\nu i} \rho_\nu^{t-\tau+1}$$

the λ_{ij}'s are constants.

By a result due to Cramer and Wold (see [7]) we know that for any t Y_t is a multivariate normal random variable. From (570) we have $E[Y_{ti}] = \phi_i(t)$

and $\lim_{t\to\infty} E[Y_{ti}] = c_i$ follows from (575).

It should also be clear that due to condition (555)

$$\lim_{t\to\infty} \mathrm{Cov}(Y_{ti}, Y_{tj})$$

exists and is finite.

---

* for sake of simplicity we assume that all roots of equation (555) are distinct. The results carry over in the general case.

Next, we indicate why the dependence of Y_{ti} and $Y_{t+k,m}$ goes to zero as the lag k increases.

$$(590)\quad \mathrm{Cov}(Y_{ti},Y_{t+k,m}) = \mathrm{Cov}\left(\sum_{\tau=1}^{t}\sum_{j=1}^{M}\phi_{ij\tau}(t)u_{\tau j}, \sum_{\tau=1}^{t+k}\sum_{\ell=1}^{M}\phi_{m\ell\tau}(t+k)u_{\tau\ell}\right)$$

using the fact that u_t is independent of $u_{t+j}(j\neq 0)$, we get

$$\mathrm{Cov}(Y_{ti},Y_{t+k,m}) = \mathrm{Cov}\left(\sum_{\tau=1}^{t}\sum_{j=1}^{M}\phi_{ij\tau}(t)u_{\tau j}, \sum_{\tau=1}^{t}\sum_{\ell=1}^{M}\phi_{m\ell\tau}(t+k)u_{\tau\ell}\right)$$

$$= \sum_{\tau=1}^{t}\sum_{j=1}^{M}\sum_{\ell=1}^{M}\phi_{ij\tau}(t)\cdot\phi_{m\ell\tau}(t+k)\sigma_{j\ell}$$

Without going into further details, using (580), it should be felt that $\mathrm{Cov}(Y_{ti},Y_{t+k,m})$ will approach zero geometrically fast as k increases. A similar result holds for $\mathrm{Cov}(Y^{e}_{ti},Y^{f}_{t+k,m})$ for fixed positive integers e and f. Furthermore, using an analogous argument to that of the single equation case, we could show that the above limits are reached uniformly for $|\theta - \theta_0| \leq r$, for some positive r. θ_0 being the true value of the vector of unknown parameters. The remaining of the proof is believed to be a routine matter.

4.2.2. We now generalize the results of section 4.2.1 to the case where exogenous variables are included.

Consider the following model of a multivariate stochastic process $Y_t = (Y_{t1},\cdots,Y_{tM})$

$$(600)\quad A_1 \cdot Y_t + A_2 \cdot Y_{t-1} + \cdots + A_{p+1} \cdot Y_{t-p} + A_0 + B_1 \cdot X_t = \varepsilon_t$$

where - X_t is a K-variate stochastic process, to be considered as representing a set of exogenous variables $X_t = (X_{t1},\cdots,X_{tk})^T$

- $A_1, A_2, \cdots, A_{p+1}$ are $M \times M$ matrices
- A_0, ε_t are M vectors
- B_1 is an $M \times k$ matrix.

Assume the disturbance vector is generated by the autoregressive process given in (510). Then by successive substitutions (600) can be rewritten as:

$$(610)\quad \Delta_1 \cdot Y_t + \cdots + \Delta_{p+q+1} \cdot Y_{t-p-q} + \Delta_0 + \Delta_{p+q+2} \cdot X_t + \cdots$$

$$+ \Delta_{p+2q+2} \cdot X_{t-q} = u_t$$

Moreover if we suppose that

(615) | (1) Δ_1 is non-singular
(2) the system in (610) is identified

then (610) is equivalent to

$$(620)\quad Y_t = \Pi_0 + \Pi_1 \cdot Y_{t-1} + \cdots + \Pi_{p+q} \cdot Y_{t-p-q} + \Pi_{p+q+1} \cdot X_t + \cdots + \Pi_{p+2q+1} \cdot X_{t-q} + v_t$$

Theorem 4.5

Let Y_t be a multivariate stationary process generated by (600) and (510) where:

(1) $\{u_t\}$ is a sequence of independent, normally distributed multivariate random variables $N(0, \sum_u)$

(2) $\{X_t\}$ is a stationary stochastic process independent of $\{u_t\}$, and there exist numbers c,ρ,δ $(c \geq 0,\ 0 \leq \rho < 1, \delta > 0)$, such that

$$\left|\mathrm{Cov}(X^k_{t,m}, X^j_{t+i,n})\right| \leq c \cdot \rho^i \qquad k,j = 1,2,3,4+\delta;\ i = 0,1,2,\cdots$$

and for every m^{th} and n^{th} component

(3) condition (615) holds.

Then the maximum likelihood estimator of θ, the vector of unknown parameters of the reduced form (620) (or equivalently ψ, the vector of unknown parameters of the structural form) is:

(1) consistent

(2) asymptotically normally distributed

(3) asymptotically efficient

The proof follows the same arguments presented in earlier theorems and will be omitted.

4.3 A Word on the Problem of Non-stationarity

All along this chapter we have seen that a crucial assumption (although not a necessary one) was the stationarity of the stochastic process under consideration. For example, in section 4.1.1, we considered the following model:

$$Y_t = \beta \cdot X_t + \varepsilon_t , \qquad t = 0,1,\cdots$$

assuming (1) $\{\varepsilon_t\}$ to be a sequence of independent, normally distributed, $N(0,\sigma^2)$, random variables

(2) $\{X_t\}$ to be a stationary stochastic process, independent of $\{\varepsilon_t\}$

then we showed that, under some other mild conditions on the $\{X_t\}$ process, the maximum likelihood estimator of (β,σ^2) has desirable asymptotic properties.

Consider now the modified model

$$Y_t = \beta \cdot X_t + \varepsilon_t \qquad t = 0,1,\cdots$$

where (1) $\{\varepsilon_t\}$ is a sequence of independent, identically distributed $N(0,\sigma^2)$ random variables

(2) X_t can be written as

$X_t = \hat{X}_t + \alpha \cdot t$ where $\hat{X}_t$ is a stationary stochastic process and α is an unknown constant.

$\{X_t\}$ and $\{\varepsilon_t\}$ are independent sequences. At each $t = 0,1,\cdots$ we observe (X_t,Y_t), and we wish to estimate (α,β,σ^2), assuming the distribution of $\hat{X}_t$ unknown.

It is clear that we are now dealing with a non-parametric statistical problem and although there exists methods of estimation for that particular situation, this type of question does not fit into our framework.

4.4 Linear Stochastic Differential Equation of the First Order

Let $X(t)$ be a one dimensional stochastic process with

$$E[X(t)]^2 < \infty \quad ,$$

$\dot{X}(t)$ its derivative (in the mean square) which we assume to exist.*

Assume $X(t)$ satisfies the following relationship:

(700) $\dot{X}(t) + aX(t) = Z(t) + b$

* for definition of differentiability of a regular process, see [24] chapter 1.

where: (1) a and b are constants

(2) $Z(t)$ is a stationary Wiener process such that

$$\text{Var}\,[Z(t) - Z(s)] = \sigma^2 |t - s|$$

for any t,s .

Based upon observations of $X(t)$ taken at time $t = 0,1,2,\cdots$ we are now considering the problem of estimating the unknown parameter $\theta = (a,b,\sigma^2)$. It is known (see [2] pp 156-164, [10]) that (700) has the unique solution:

$$(710)\quad X(t_{i+1}) = x_i e^{-a(t_{i+1}-t_i)} + \int_{t_i}^{t_{i+1}} e^{-a(t_{i+1}-v)}\, dY(v) + \frac{b}{a}$$

where - $Y(t)$ is the (mean square) differential of the $Z(t)$ process

- $X(t_i) = x_i$ is the initial condition.

Since the observations are equally spaced in time, we have

$$(720)\quad X(t_{i+1}) = x_i e^{-a} + \int_0^1 e^{-az}\, dY(z) + \frac{b}{a}$$

Finally we see that conditionally on $X(t_i) = x_i$, the distribution of $X(t_{i+1})$ is found to be normal with

mean: $\quad x_i e^{-a} + \frac{b}{a}$

variance: $\quad \sigma^2 \int_0^1 e^{-2az}\, dz = \sigma^2\, \frac{1-e^{-2a}}{2a}$

Moreover, if we assume $a > 0$, then from (710) it follows that $X(t)$ is asymptotically stationary, and its limiting distribution is $N(\frac{b}{a}, \frac{\sigma^2}{2a})$.

From (710) we see that the discrete process $\{X(t_i)\}$ $i = 0,1,\cdots$ is a Markov process since the only piece of information used at t_{i+1} is the observed value of $X(t_i)$.

The likelihood function is:

$$(730)\quad L_n = \prod_{i=1}^{n} P_\theta[X_i = x_i | X_{i-1} = x_{i-1}]$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\sqrt{\frac{2a}{1-e^{-2a}}}\ \exp\left\{ -\frac{1}{2}\left(\frac{x_i - x_{i-1}e^{-a} - b/a}{\sigma}\right)^2 \cdot \frac{2a}{1-e^{-2a}} \right\}$$

We see that the Log-likelihood function Z_n is a polynomial in the x_i's at most of degree 2 . So are $\frac{\partial Z_n}{\partial a}$, $\frac{\partial Z_n}{\partial b}$, $\cdots$ and other derivatives of interest.

From (710) we compute $Cov(X(t),X(u))$, $(u < t)$

$$(740)\quad Cov(X(t),X(u)) = e^{-a(t-u)} \cdot Var(X(u))$$

Assuming the process has reached its equilibrium we get

$$(750)\quad Cov(X(t),X(u)) = \frac{\sigma^2}{2a} e^{-a(t-u)}$$ which goes uniformly to zero, for $|\theta - \theta_0| \leq r$, as $(t-u)$ increases, and for some positive number r .

Obviously this result holds for $Cov(X^e(t),X^f(u))$, where e,f, are positive integers no greater than some constant.

We finally end up with the following result:

Theorem 4.6

Consider a regular stochastic process $X(t)$ satisfying

$$\dot{X}(t) + aX(t) = Z(t) + b \quad .$$

Where: (1) $a > 0$ and b are constants

(2) $Z(t)$ is a Wiener process with diffusion coefficient σ^2 .

Then based on observations of the $X(t)$ process taken at time $t = 0,1,\cdots$ the maximum likelihood estimator of θ is:

(1) consistent

(2) asymptotically normally distributed

(3) asymptotically efficient

Remark 6:

(1) The conclusion does not change if the observations are not equally spaced in time. Under the additional restriction that

$$Cov(X^e(t),X^f(u)) \to 0$$

geometrically fast as the number of observation in the interval (t,u) increases and where $e,f, = 1,2,3,4 + \delta$ (for some positive number δ) , uniformly for $|\theta - \theta_0| \leq r$. A sufficient condition

which will guarantee that the above restriction holds is: $t_{i-1} - t_i \geq \Delta$, where Δ is a fixed positive number.

(2) If instead of observing $X(t)$, $t = 0,1,\cdots,$ we observe the sequence $W(t) = X(t) + \varepsilon(t)$, where the $\varepsilon(t)$ are independent, normally distributed random variables $N(0,s^2)$, the asymptotic properties of $\theta' = (a,b,\sigma^2,s^2)$ are the same as those given above for θ .

4.5 Linear Stochastic Differential Equation of Order $r > 1$

As in the previous case, we consider $X(t)$, a regular stochastic process such that

$$\frac{dX(t)}{dt} , \frac{d^2X(t)}{dt^2} , \ldots, \frac{d^rX(t)}{dt^r} \qquad \text{exist (in the mean square)}$$

Assume $X(t)$ satisfies the equation:

$$(760) \quad \frac{d^rX(t)}{dt} + a_1 \frac{d^{r-1}X(t)}{dt^{r-1}} + \cdots + a_{r-1} \frac{dX(t)}{dt} + a_rX(t) + a_{r+1} = Z(t)$$

where $Z(t)$ is defined as before.

Let $\lambda_1,\cdots,\lambda_r$ be the roots of

$$(765) \qquad \lambda^r + a_1\lambda^{r-1} + \ldots + a_r = 0$$

Then it is known (see the same references as in the 1^{st} order case) that (760) has the unique solution:

$$(770)\quad X(t) = A_1(u)e^{\lambda_1(t-u)} + \cdots + A_r(u)e^{\lambda_1(t-u)} + \int_u^t g(t-v)\,dY(v)$$

where (1) $Y(t)$ is the (mean square) differential of $Z(t)$

(2) $A_1(u),\cdots,A_r(u)$ are determined from the values of $X(t)$,

$$\frac{dX(t)}{dr}\ ,\ \cdots\ ,\frac{d^{r-1}X(t)}{dt^{r-1}} \quad \text{at} \quad t = u$$

(3) $g(t)$ is the solution of

$$(775)\quad \frac{d^r g(t)}{dt^r} + a_1 \frac{d^{r-1}g(t)}{dt^{r-1}} + \cdots + a_r g(t) = 0$$

with the r initial conditions:

$$\left.\frac{d^{r-1}g(t)}{dt^{r-1}}\right|_{t=0} + a_1 \left.\frac{d^{r-2}g(t)}{dt^{r-2}}\right|_{t=0} + \cdots + a_{r-1}g(0) = 1$$

$$\left.\frac{d^{r-2}g(t)}{dt^{r-2}}\right|_{t=0} + a_1\left.\frac{d^{r-3}g(t)}{dt^{r-3}}\right|_{t=0} + \cdots + a_{r-2}\,g(0) = 0$$

$$\vdots$$

$$g(0) = 0$$

If in addition, we assume the real part of λ_i to be negative, then it is easy to see that $X(t)$ is asymptotically stationary.

Suppose that we observe the vector

$$V(t_i) = \left(X(t_i), \left.\frac{dX(t)}{dt}\right|_{t=t_i}, \cdots, \left.\frac{d^{r-1}X(t)}{dt^{r-1}}\right|_{t=t_i} \right)$$

at $t_0 < t_1 < t_2 < \cdots < t_n$, with $t_i - t_{i-1} \geq \Delta > 0$. From (770), (775) we see that $V(t_i)$, $i = 0,1,\cdots,n$, is a Markov process. More precisely, conditionally on $V(t_{i-1}) = v_{i-1}, V(t_i)$ is distributed as a multivariate normal random variable.

One would not have any trouble writing in details the likelihood function and proving

Theorem 4.7

In the problem outlined above, the maximum likelihood estimator of the unknown parameter $\theta = (a_1,\cdots,a_{r+1},\sigma^2)$ is:

(1) consistent

(2) asymptotically normally distributed

(3) asymptotically efficient

Remark 7: We used the fact that at each t_i , $i = 0,1,\cdots,n$, one observes the random vector $V(t_i)$. Although this might be possible in physics, where $X(t)$ could be the location of an object, affected by some random noise, $\frac{dX(t)}{dt}$ and $\frac{d^2X(t)}{dt^2}$, its velocity and acceleration

respectively, in econometrics one would only expect to record the value of $X(t)$ for $t = t_0, t_1, \cdots, t_n$. Based on $X(t_i) = x_i$, the $A_j(t_i)$'s are now random variables, whose steady state distributions depend on $x_0, x_1, \cdots, x_i$. So that the likelihood function can now be written as:

$$P_\theta[X_1 = x_1 | X_0 = x_0] \cdot P_\theta[X_2 = x_2 | X_i = x_i,\ i = 0,1] \cdot \cdots$$

$$\cdot P_\theta[X_n = x_n | X_i = x_i,\ i = 0,\cdots,n-1]$$

The $X(t)$ process alone is not Markovian and the problem of estimation in this situation does not fit into out framework.

4.6 Generalization to System of Stochastic Differential Equations

Let us now consider a system of stochastic differential equations:

$$(790)\quad A_0 \frac{d^r X(t)}{dt^r} + \cdots + A_r X(t) + B = Z(t)$$

where $X(t) = (X_1(t), \cdots, X_m(t))^T$

$A_0, \cdots A_r$ are $m \times m$ matrices

B is a m-vector

$Z(t)$ is an m-dimensional Wiener process.

Under suitable conditions on the A_i's, $i = 0, \cdots, r$. (790) has a unique and asymptotically stationary solution.

If for $t_0, t_1, \cdots, t_n$ with $t_i - t_{i-1} \geq \Delta > 0$ one observes $X(t)$, $\frac{dX(t)}{dt}$, $\cdots$, $\frac{d^{r+m-2}X(t)}{dt^{r+m-2}}$, then we are in a situation analogous to that of Theorem 4.7 and the maximum likelihood estimator of the vector of unknown parameters will enjoy desirable properties. However, if at each t_i, we only observe $X(t_i)$ then, for the same reasons as those stated in the remark following Theorem 4.7, our approach is inadequate.

CHAPTER V

APPLICATION TO THE ESTIMATION PROBLEM OF CONTINUOUS TIME, COMPLETELY DISCONTINUOUS MARKOV AND MARKOV RENEWAL PROCESSES

5.1 Estimation in a Continuous Time, Jump Type Markov Process

Let $S = 1,2,\cdots,N$ be the state space of the process. We now borrow notation from part II of [4].

Let $P_\theta[t,i,A] = P_\theta[X(s+t) \in A \,|\, X(s) = i]$ for each $\theta \in \Theta$

where $\theta = (\theta_1,\cdots,\theta_r)$ is a vector of unknown parameters; Θ is the parameter space. Also the process is supposed to be time-homogeneous.

Assume:

$$\lim_{t\to 0} P_\theta[t,i,i] = 1 \qquad \text{for all } \theta \in \Theta \tag{800}$$

Condition 1 :

Each sample function is a right continuous step function. The limit (800) holds for all i so that there exist functions $q(i,\theta)$ and $q(i,j,\theta)$ $(i \neq j)$, satisfying (810)

$$\lim_{t\to 0} (1 - P_\theta[t,i,i])/t = q(i,\theta)$$

$$\lim_{t\to 0} P_\theta[t,i,j]/t = q(i,j,\theta) \quad (i \neq j) \tag{810}$$

for all θ and i, $q(i,\theta) > 0$

Let us denote $Z_1, Z_2, \cdots$ the successive states of the system and $\rho_1, \rho_2, \cdots$ the successive sojourn times. So that we have

$$(830) \qquad \begin{array}{ll} X(t) = Z_1 & \text{if } 0 \leq t_1 < \rho_1 \\ \vdots & \vdots \\ X(t) = Z_n & \text{if } \rho_1 + \cdots + \rho_{n-1} \leq t < \rho_1 + \cdots + \rho_n \end{array} \qquad n \geq 1$$

(840) Let $\nu(t) = \max \{k : \rho_1 + \cdots + \rho_k < t\}$

From the general theory of discontinuous Markov processes, it follows that

$$\{(Z_n, \rho_n) \; ; \; n = 1,2,\cdots\}$$

is a Markov process with the following transition probabilities:

$$(850) \qquad P_\theta[Z_{n+1} = z_{n+1},\ r \leq \rho_{n+1} < r + dr \,||\, Z_n = z_n] =$$

$$\frac{q(z_n, z_{n+1}, \theta)}{q(z_n, \theta)} \cdot q(z_{n+1}, \theta) e^{-q(z_{n+1}, \theta) \cdot r} dr$$

$$(860) \quad \text{Let} \quad f(z_n, z_{n+1}, \theta) = \frac{q(z_n, z_{n+1}, \theta)}{q(z_n, \theta)}$$

then the densities of the transition probabilities can be written as:

$$(870)\quad F(z_n,z_{n+1},r,\theta) = f(z_n,z_{n+1},\theta)\cdot q(z_{n+1},\theta)\cdot e^{-q(z_{n+1},\theta)\ r}$$

Following Billingsley, we shall take the function.

$$(880)\quad L_t(\theta) = \sum_{k=1}^{\nu(t)-1} [\text{Log } f(z_k,z_{k+1},\theta) + \text{Log } q(z_{k+1},\theta) - \rho_{k+1}\cdot q(z_{k+1},\theta)]$$

as the Log-likelihood function of θ , given the observation $\{X(\tau)\ ;\ 0 \le \tau < t\}$.

As pointed out in [4], the sample $\{X(\tau)\ ;\ 0 \le \tau < t\}$ from the original process contains a little more information than does the sample $\{Z_k,\rho_k)\ ;\ k = 0,1,\cdots,\nu(t)\}$ from the imbedded process. However, asymptotic results are not affected by ignoring this extra information.

Condition 2:

The set D of (i,j) such that $q(i,j,\theta) > 0$ is independent of θ and the functions $q(i,j,\theta)$ have continuous third order partial derivatives throughout θ . Let d be the number of elements in D . Then the $(d \times r)$ matrix $\left(\frac{q(i,j,\theta)}{\partial\theta_u}\right)\left\{\begin{array}{l}(i,j)\ \epsilon\ D\\ i \le u \le r\end{array}\right.$ has rank r for all $\theta\ \epsilon\ \Theta$.

Condition 3:

The Markov process $\{Z_n\}$, $n = 1,2,\cdots$ is an irreducible Markov chain with state space $1,2,\cdots,N$ and transition matrix

$$(P_{ij}) = \left|\frac{q(i,j,\theta)}{q(i,\theta)}\right|$$

Lemma 5.1

Let $Y_t = f(Z_t, Z_{t+1},\rho_{t+1},\theta)$ be a function of the random variables Z_t,Z_{t+1},ρ_{t+1} and of the unknown parameter θ .

Assume $E[Y_t^2 || Z_t] < \infty$ for all $\theta \in \Theta$ and Z_t .

Then under conditions 1 - 3:

(1) $\mathrm{Cov}\,(Y_t,Y_{t+i}) \to 0$ as $i \to \infty$, uniformly for $|\theta - \theta_0| \leq \varepsilon$ where θ_0 is the true value of θ . ε is some positive number.

(2) $\lim\limits_{n\to\infty} n^{-1} \sum\limits_{i=1}^{n} Y_i = \sum\limits_{j=1}^{N} E_\theta[Y_t||Z_t = j] \cdot \Pi_j = \lambda_\theta$, (say) uniformly for $|\theta - \theta_0| \leq \varepsilon$, where $(\Pi_1,\cdots,\Pi_N)$ is the steady state distribution of Z_t .

(3) $\lim\limits_{t\to\infty} \nu(t)^{-1} \sum\limits_{i=1}^{\nu(t)} f(Z_t,Z_{t+1},\rho_{t+1},\theta) = \lambda_\theta$ (in probability)

and the limit is reached uniformly for $|\theta - \theta_0| \leq \varepsilon$.

Proof

(1) - by Condition (3), it does make sense to consider the statistical equilibrium of the Z_t process

- by Condition (2), the $P_{ij}(\theta)$'s are continuous functions of θ so that appealing once more to Doob's argument (see [8] pp 172-174):

$$P_{ij}^{(n)}(\theta) = P_\theta[Z_{t+n} = j \,||\, Z_t = i] \to \Pi_j \qquad \text{as} \quad n \to \infty$$

Moreover the convergence is geometrically fast and uniform for $|\theta - \theta_0| \leq \varepsilon$ for some positive number ε (the same argument was used in Chapter III)

- now, following the proof ot Lemma 3.1, it is easy to conclude that

$$\lim_{i \to \infty} \text{Cov}(Y_t, Y_{t+i}) = 0$$

uniformly for all $|\theta - \theta_0| \leq \varepsilon$. The only minor change is that here, we do not assume Y_t to be bounded but instead assume Y_t has a finite variance (for all Z_t) and therefore a finite mean.

(2) From the first part of the lemma and the weak law of large numbers, conclusion (2) follows immediately. The uniformity of the convergence being implied by

(i) the uniformity of the convergence in conclusion (1)

(ii) the assumption that $E[Y_t^2 \,||\, Z_t] < \infty$, for all Z_t

(3) Let $\mu_i = E[\rho_t \,||\, Z_t = i]$, then from part 2 of this lemma, we have:

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \rho_k = \sum_{j=1}^{N} \Pi_j \cdot \mu_j = \mu \quad , \quad \text{say} \quad (\mu < \infty) \quad \text{(in probability)}$$

Using the fact (proved by Billingsley in [4]) that

$$\lim_{t\to\infty} \frac{\nu(t)}{t} = \frac{1}{\mu} \quad \text{(in probability)},$$

we see that $\nu(t) \to \infty$ in probability as $t \to \infty$

and the conclusion follows readily from part 2.

Q.E.D.

Letting $G(z_n, z_{n+1}, r, \theta) = \text{Log } F(z_n, z_{n+1}, r, \theta)$, Billingsley showed that under conditions (1) - (3)

$$(890) \quad \operatorname*{Plim}_{t\to\infty} t^{-\frac{1}{2}} \left| \sum_{k=1}^{\nu(t)} \frac{\partial G}{\partial \theta_u}(z_k, z_{k+1}, \rho_{k+1}; \theta) - \sum_{k=1}^{[t/\mu)} \frac{\partial G}{\partial \theta_u}(z_k, z_{k+1}, \rho_{k+1}, \theta) \right| = 0$$

where $[\frac{t}{\mu}]$ is the largest integer less or equal to $\frac{t}{\mu}$. So that (890) allows us to apply the central limit theorem for martingales to a random sum. We are therefore back to the case treated in Chapter II, where we considered the problem of getting the asymptotic distribution of a non-random sum.

Theorem 5.1

Let us consider a completely discontinuous, time continuous, time homogeneous Markov process with finite state space, depending on some unknown parameter θ .

Then under conditions (1) - (2) - (3) the maximum likelihood estimator of θ is:

(1) consistent*

(2) asymptotically normally distributed, when properly normalized*

(3) asymptotically efficient

The proof of the third part of the conclusion is based upon Theorem 2.7 and Lemma 5.1. It is fairly easy to see that the conditions, under which Lemma 5.1 holds, are fulfilled by the expressions involved in Theorem 2.7 $\left(\text{where } \sum_{k=1}^{n} \text{ is now replaced by } \sum_{k=1}^{\nu(t)}\right)$, and therefore that the various limits are reached uniformly for $|\theta - \theta_0| \leq \varepsilon$. The conclusion follows.

Remark 7:

(1) For the same reason that the one given in Chapter III on the geometric ergodicity of denumerable Markov chains, our approach does not allow us to conclude that under some conditions on the process, the

---

* results given by Billingsley in [4].

maximum likelihood estimator of some unknown parameter is asymptotically efficient.

(2) It is particularly interesting to see that in general the maximum likelihood equations will be easily handled and this is the purpose of the following example to illustrate the calculations involved.

Example:

Let us consider the statistical analysis of the model described on pp 416-418 of [11].

We consider a set of m identical automatic machines which are attended by r repairmen $(r < m)$. If a machine breaks down, it is serviced at once unless no repairman is available, in which case it joins a waiting line.

Under reasonable assumptions:

(1) the length of time, during which any machine is in a working state, has an exponential distribution, with mean $\frac{1}{\lambda}$.

(2) the time required for servicing any machine is taken as a random variable with an exponential distribution and mean $\frac{1}{\mu}$.

We say that the system is in state i at time t if i machines are not working. Thus the state space is $X = \{0,1,\ldots,m\}$. Billingsley [4] has shown that the general log-likelihood function can be, in the finite state case, reduced to

$$(900) \qquad L_t(\theta) = \sum_D [t_{ij} \operatorname{Log} q(i,j,\theta) - \gamma_i \cdot q(i,j,\theta)]$$

where: - t_{ij} is the number of direct jumps from i to j in the sample function

- γ_i is the total amount of time the system was in state i, up to time t.

- D is the set of (i,j) such that $q(i,j,\theta) > 0$

in this example we have:

$$\theta = (\lambda,\mu)$$

$$(910)\quad \begin{cases} q(i,i+1,\theta) = (m-i)\lambda & i = 0,1,\cdots,m-1 \\ q(i,i-1,\theta) = i\mu & i = 1,\cdots,r \\ q(i,i-1,\theta) = r\mu & i = r+1,\cdots,m \end{cases}$$

From (900) and (910) we get:

$$(920)\quad L_t(\theta) = \sum_{i=0}^{m-1} [t_{i,i+1}\mathrm{Log}(m-i)\lambda - \gamma_i(m-i)\lambda]$$

$$+ \sum_{i=1}^{r} [t_{i,i-1}\ \mathrm{Log}\ i\mu - \gamma_i \cdot i\mu]$$

$$+ \sum_{i=r+1}^{m} [t_{i,i-1}\mathrm{Log}\ r\mu - r\gamma_i\mu]$$

taking $\frac{\partial L_t}{\partial \lambda}(\theta)$ and $\frac{\partial L_t}{\partial \mu}(\theta)$ and setting them equal to zero we get:

$$\hat{\lambda}_t = \frac{\sum_{i=0}^{m-1} t_{i,i+1}}{\sum_{i=0}^{m-1} \gamma_i (m-i)}$$

(930)

$$\hat{\mu}_t = \frac{\sum_{i=1}^{m} t_{i,i-1}}{\sum_{i=1}^{r} i\gamma_i + r \sum_{i=r+1}^{m} \gamma_i}$$

It is easy to check that for this example conditions (1) - (3) hold and accordingly, that $(\hat{\lambda}_t, \hat{\mu}_t)$ is an asymptotically efficient estimator of (λ, μ) as $t \to \infty$.

5.2 Estimation in a Markov Renewal Process

Let us briefly recall the definition of a Markov renewal process with $m(< \infty)$ states. For more details, we refer the reader to Pike's papers [25] and [26]. One is given:

(1) a matrix of transition distributions (Q_{ij}) where Q_{ij} is a mass function defined on $(-\infty, +\infty)$ satisfying $Q_{ij}(x) = 0$ for $x \leq 0$ and

$$\sum_{j=1}^{m} Q_{ij}(\infty) = 1 \qquad 1 \leq i \leq m$$

(2) an m-tuple of initial propabilities $(p_1, \cdots, p_m)$ which satisfies $p_j \geq 0$ and $\sum_{j=1}^{m} p_j = 1$

Consider any two-dimensional Markov process $\{(J_n, X_n);\ n \geq 0\}$ defined on a probability space that satisfies $X_0 = 0$ (a.s), $P[J_0 = k] = p_k$ and

$$P[J_n = k,\ X_n \leq x \,||\, J_0, J_1, \cdots, J_{n-1}, X_1, \cdots X_{n-1}] = Q_{J_{n-1},k}(x) \quad \text{(a.s)}$$

for all $x \in (-\infty, +\infty)$ and $1 \leq k \leq m$.

Let $N(t) = \sup \{n \geq 0 : \sum_{i=0}^{n} X_i \leq t\}$

$N_j(t)$ be the number of times $J_k = j$ for $1 \leq k \leq N(t)$

$N_{ij}(t)$ be the number of times $J_k = i$ and $J_{k+1} = j$ for $1 \leq k \leq N(t) - 1$

then the stochastic process $\{N_1(t), \cdots, N_m(t);\ t \geq 0\}$ is called a Markov renewal process determined by the initial probabilities and matrix of transition distributions.

Let $P_{ij} = Q_{ij}(\infty)$ and $P = (P_{ij})$ an $m \times m$ matrix

$$H_i(t) = \sum_{j=1}^{m} Q_{ij}(t)$$

Assume the m^2 functions $Q_{ij}(t)$ depend on an unknown parameter $\theta = (\theta_1, \cdots, \theta_r)$. We want to find conditions under which the maximum likelihood estimator of θ has desirable asymptotic properties.

Condition (1) : We assume throughout that the Markov renewal process is irreducible, positive recurrent for all $\theta \in \Theta$. In view of Theorem 5.1 of [25] it is necessary and sufficient that for all $\theta \in \Theta$:

(i) P be a positive recurrent Markov chain

(ii) $\eta_j = \int_0^\infty t dH_j(t) < \infty \quad$ for $1 \leq j \leq m$

Condition (2) : We shall assume that the $Q_{ij}(t|\theta)$ are differentiable w.r.t t for all $\theta \in \Theta$. Although we can derive the same results in the discrete or mixed case, this will make the study easier. The extension to the two other cases is immediate.

$$\text{Let } q_{ij}(t|\theta) = \frac{\partial Q_{ij}}{\partial t}(t|\theta) \qquad (t > 0)$$

then for any i , the set of j and t for which $q_{ij}(t|\theta) > 0$ does not depend on θ . For any i,j,t, the functions $q_{ij}(t|\theta)$ have continuous third order partial derivatives throughout

$$\text{Let } g(i,j,x|\theta) = \text{Log } q_{ij}(x|\theta)$$

Then we suppose that there exists an ε - neighborhood of θ_0 (the true value of the parameter) such that for any u,v,w,i

(1) $E \sup_{\theta \in N_\varepsilon(\theta_0)} \left| \frac{\partial q_{ij}}{\partial \theta_u} (x|\theta) \right| < \infty$

(2) $E \sup_{\theta \in N_\varepsilon(\theta_0)} \left| \frac{\partial^2 q_{ij}}{\partial \theta_u \partial \theta_v} (x|\theta) \right| < \infty$

(3) Let $G(i,j) = \sup_{\theta \in N_\varepsilon(\theta_0)} \left| \frac{\partial^3 g}{\partial \theta_u \partial \theta_v \partial \theta_w} (i,j,x|\theta) \right|$

then $E[G(i,j)]^2 < \infty$

(4) $E\left[\frac{\partial g(i,j,x|\theta)}{\partial \theta_u}\right]^{4+\delta} < \infty$ for all $\theta \in N_\varepsilon(\theta_0)$

and for some $\delta > 0$

(5) $E\left[\frac{\partial^2 g(i,j,x|\theta)}{\partial \theta_u \partial \theta_v}\right]^2 < \infty$ for all $\theta \in N_\varepsilon(\theta_0)$

(6) Let $\sigma_{uv}(\theta) = E\left[\frac{\partial g(i,j,x|\theta)}{\partial \theta_u} \cdot \frac{\partial g(i,j,x|\theta)}{\partial \theta_v}\right]$

where the expectation is taken under the assumption that the distribution of state i has reached its equilibrium.

then $-\sigma_{uv}(\theta)$ is assumed to be continuous in $N_\varepsilon(\theta_0)$

$-\sigma(\theta) = \{\sigma_{uv}(\theta)\}$ is assumed to be non-singular.

Remark 8 : It is clear that the assumptions we need here are far more stringent than those needed in the Markovian case. This is due to the fact that in the former case, the sojourn time in any state has an exponential distribution and that here we are looking for a larger family of distributions.

Assume we observe the process on $[0,t)$, then the likelihood function is:

$$q_{0j_1}(x_1) \cdot q_{j_1 j_2}(x_2) \cdot \cdots \cdot q_{j_{N(t)-1} j_{N(t)}}(x_{N(t)}) \cdot \left\{1 - H_{j_{N(t)}}\left(t - \sum_{i=1}^{N(t)} x_i\right)\right\}$$

As we did in the first part of this chapter, we shall take as Log-likelihood function the following truncated expression $L_t(\theta)$:

$$(950) \quad L_t(\theta) = \sum_{i=1}^{N(t)} q_{i-1,i}(x_i) \quad ,$$

and as noted in the former case, asymptotic results will not be affected.

Theorem 5.2

Let us consider a Markov renewal process with finitely many states, depending on some unknown parameter θ .

Then under Conditions 1 and 2, the maximum likelihood estimator of θ is:

(1) consistent

(2) asymptotically normally distributed, when properly normalized

(3) asymptotically efficient.

Proof: Under condition (1), we know from a result given by Pyke and Schaufele (see [27]) that:

$$N(t) \to \infty \quad \text{w.p.1} \qquad \text{as} \quad t \to \infty$$

this fact and condition (2) enable us to apply lemma 5.1 and typically we conclude for instance that:

$$\lim_{t\to\infty} N(t)^{-1} \sum_{k=1}^{N(t)} \frac{\partial g}{\partial \theta_u}(i_k, j_{k+1}, x_k | \theta) = 0 \qquad \text{in probability}$$

and uniformly for $|\theta - \theta_0| \leq \varepsilon$.

The conclusion follows from Theorem 2.7.

Remark 9: The denumerable state space case offers even more difficulty than in the Markovian case, since there does not seem to be any simple necessary and sufficient condition for the positivity of the process. (See [26] pp 1240-1241)

Example of a Maximum Likelihood Estimation for a Markov Renewal Process

Let us consider a Markov renewal process where the transitions distributions $Q_{ij}(\cdot)$ can be expressed as:

$$Q_{ij}(t) = P_{ij} \cdot H_i(t) \tag{960}$$

where $H_i(t) = P$ [sojourn time in state $i \leq t$]

$$= \int_0^t \frac{e^{-\lambda_i x} \lambda_i^{\alpha_i} x^{\alpha_i - 1}}{\Gamma(\alpha_i)} dx \qquad t \geq 0 \qquad i = 1,\cdots,m$$

$$(0 < P_{ij} < 1 \ , \ \lambda_i > 0 \ , \ \alpha_i > 0 \ ; \ i, j = 1,\cdots,m)$$

Assume that the quantities $P_{ij},\lambda_i,\alpha_i$ $(i = 1,\cdots,m; \ j = 1,\cdots,m-1)$ are not functionally dependent, then we wish to estimate

$$\theta = (P_{11},P_{12},\cdots,P_{1,m-1},P_{2m},\cdots,P_{m,m-1},\lambda_1,\cdots,\lambda_m,\alpha_1,\cdots,\alpha_m)$$

an $(m^2 + 2m - 2)$ dimensional parameter*.

---

* There is no change in the approach if a subset of the (m^2+2m) components is known. The only additional requirement is that (P_{ij}) must be irreducible.

The truncated likelihood function may be written as:

$$(965) \quad \prod_{i=1}^{m} \prod_{j=1}^{m} P_{ij}^{N_{ij}(t)} \prod_{k=1}^{N_i(t)-\delta_i} dH_i(x_{ik})$$

where: x_{ik} is the length of the k^{th} sojourn time in state i

$$\delta_i = \begin{cases} -1 & \text{if we are in state } i \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

By the assumption of the functional independence, the problem of maximizing the expression in (965) can be reduced to two separate maximum likelihood problems:

$$(1) \quad \max_{0< P_{ij} < 1} \prod_{i=1}^{m} \prod_{j=1}^{m} P_{ij}^{N_{ij}(t)}$$

$$(2) \quad \max_{\lambda_i,\alpha_i > 0} \prod_{i=1}^{m} \prod_{k=1}^{N_i(t)-\delta_i} \frac{e^{-\lambda_i x_{ik}} \lambda_i^{\alpha_i} x_{ik}^{\alpha_i-1}}{\Gamma(\alpha_i)}$$

The answer to (1) was given in Chapter III for a fixed total number of observations. Here we simply have:

$$\hat{P}_{ij} = \frac{N_{ij}(t)}{N_i(t)+\delta_i}$$

the answer to (2) is equally easy

$$\text{Let } T_{1i}(t) = \sum_{k=1}^{N_i(t)+\delta_i} X_{ik} \qquad i = 1,\cdots,m$$

$$T_{2i}(t) = \sum_{k=1}^{N_i(t)+\delta_i} \text{Log } X_{ik} \qquad i = 1,\cdots,m$$

Then it is known that $\{T_{1i},T_{2i};\ i = 1,\cdots,m\}$ are sufficient statistics for this problem.

Now it is easy to derive that $(\hat{\alpha}_i,\hat{\lambda}_i)$ is a solution of the system:

$$\begin{cases} \dfrac{\hat{\alpha}_i}{\hat{\lambda}_i} = \dfrac{T_{1i}(t)}{N_i(t)+\delta_i} \\ \text{Log } \hat{\lambda}_i - \dfrac{1}{\Gamma(\hat{\alpha}_i)} \dfrac{d\Gamma(\alpha)}{d\alpha}\Big|_{\alpha=\hat{\alpha}_i} + \dfrac{T_{2i}(t)}{N_i(t)+\delta_i} = 0 \qquad i = 1,\cdots,m \end{cases}$$

Finally it can be verified that conditions (1) and (2) are fulfilled, so that the maximum likelihood estimator of θ is asymptotically efficient.

BIBLIOGRAPHY

[1] Anderson, T. W. and Goodman, L. (1957). "Statistical inference about Markov chains." Annals of Math. Stat. Vol 28, pp 89-110.

[2] Bartlett, M. S. (1966). "An introduction to stochastic processes." Cambridge University press.

[3] Berk, R. H. (1967). "Zehna, Peter W. Invariance of maximum likelihood estimators." Review #1922, Mathematical Reviews, Vol. 33, pp 342-343.

[4] Billingsley, P. (1961). "Statistical inference for Markov processes." Institute of Mathematical Statistics, University of Chicago. Statistical Research Monographs, University of Chicago Press, Chicago.

[5] Billingsley, P. (1961). "Statistical methods in Markov chains." Annals of Math. Stat., Vol 32, pp 12-40.

[6] Cramér, H. (1946). "Mathematical methods of mathematical statistics." Princeton University press.

[7] Cramér, H. and Wold, H. (1936). "Some theorems on distribution functions." J. London Math. Soc., Vol 11, pp 290-295.

[8] Doob, J. L. (1953). "Stochastic processes." John Wiley and Sons, New York.

[9] Dudewicz, E. J. (1969). "Estimation of ordered parameters." Unpublished Ph. D.thesis. Cornell University.

[10] Edwards, D. A. and Moyal, J. E. (1955). "Stochastic differencial equations." Proc. Camb. Phil. Soc., 51, pp 663-677.

[11] Feller, W. (1957). "An introduction to probability theory and its applications." Vol 1, 2nd Ed. John Wiley and Sons, New York.

[12] Feller, W. (1966). "An introduction to probability theory and its applications." Vol 2, John Wiley and Sons, New York.

[13] Fisz, M. (1963). "Probability theory and mathematical statistics." John Wiley and Sons, New York.

[14] Gold, R. Z. (1960). "Inference about Markov chains with non-stationary transition probabilities." Unpublished Ph.D. thesis. Columbia University.

[15] Goldberger, A. (1964). "Econometric theory." John Wiley and Sons, New York.

[16] Hajnal, J. (1956). "The ergodic properties of non-homogeneous finite Markov chains." Proceedings of the Cambridge Philosophical Society, 54, pp 233-246.

[17] Hajnal, J. (1958). "Weak ergodicity in non-homogeneous Markov chains." Proc. Camb. Phil. Soc., 54, pp 233-246.

[18] Karlin, S. (1966) "A first course in stochastic processes." Academic Press.

[19] Kendall, D. G. (1959). "Unitary dilations of Markov transition operators and the corresponding integral representations for transition-probability matrices." Probability and Statistics. Edited by Grenander, pp 139-161.

[20] Kozniewska, I. (1962). "Ergodicité et stationnarité des chaines de Markoff variables à un nombre fini d'états possibles." Colloquium Mathematicum, Vol IX, Fasc. 2, pp 333-346.

[21] Kozniewska, I. (1963). "Sur les lois des grands nombres pour les variables aleatoires dependantes à variances egalement bornées." Coll. Math., Vol X, Fasc. 2, pp 289-304.

[22] Mann, H. B. and Wald, A. (1943). "On the statistical treatment of linear stochastic difference equations." Econometrica, Vol 11, July - Oct. pp 173-220.

[23] Moore, E. H. and Pyke, R. (1964). "Estimation of the transition distributions of a Markov renewal process." Boeing Sc. Res. Labs., Tech. Rep. D1-82-0371.

[24] Prabhu, N. U. (1965). "Stochastic processes." The Macmillan Company, New York.

[25] Pyke, R. (1961). "Markov renewal processes: definitions and preliminary properties." Ann. Math. Stat., 32, pp 1231-1242.

[26] Pyke, R. (1961). "Markov renewal processes with finitely many states." Ann. Math. Stat. 32, pp 1243-1259.

[27] Pyke, R. and Schaufele, R. (1964). "Limit theorems for Markov renewal processes." Ann. Math. Stat., 35, pp 1746-1764.

[28] Serfling, R. J. (1968). "Contributions to central limit theory for dependent variables." Ann. Math. Stat., 39, pp 1158-1175.

[29] Tintner, G. (1952). "Econometrics." John Wiley and Sons, New York.

[30] Vere-Jones, D. (1962). "Geometric ergodicity in denumerable Markov chains." Quarterly Journal of Math. Oxford, series 2, #49-52, pp 7-28.

[31] Weiss, L. and Wolfowitz, J. (1966). "Generalized maximum likelihood estimators." Teoriya Vyeroyatnostey, 11, No. 1, pp 68-93.

[32] Weiss, L. and Wolfowitz, J. (1967). Maximum probability estimators." Annals of the Inst. of Stat. Math., Vol 19, No. 2, pp 193-206.

[33] Wilks, S. (1962). "Mathematical Statistics." John Wiley and Sons, New York.

[34] Zehna, P. W. (1966). "Invariance of maximum likelihood estimators." Ann. of Math. Stat. Vol 37, p 744.

Errata

Page and line	For	Substitute
v, ℓ. 12 from bottom	A work	A word
3, ℓ.3 from bottom	$0 \leq t \quad \infty$	$0 \leq t < \infty$
13, ℓ. 3 from top	$= \sigma_{uv}$	$= - \sigma_{uv}$
19, ℓ. 5 from bottom	$G(x_{k-1,} x_i)$	$G(x_{i-1}, x_i)$
29, ℓ. 9 from top	. This	, this
54, ℓ. 1 from top	$\underset{1 \leq i \leq}{\text{Min}}$	$\underset{1 \leq i \leq N}{\text{Min}}$
68, ℓ. 1 from top	Cov(	|Cov(
70, ℓ. 5 from top	$-\delta_0 - \delta_{p+q} y_1$	$-\delta_0 - \cdots - \delta_{p+q} y_1$
76, ℓ. 4 from top	limiting of	limiting distribution of
78, ℓ. 9 from top	at $t \to \infty$	as $t \to \infty$
81, ℓ. 2 from bottom	$+A_0 =$	$+ A_0 = \epsilon_t$
102, ℓ. 2 from top	the function.	the function
102, ℓ. 2 from bottom	$\left(\frac{q(i,j,\theta)}{\partial \theta_u} \right)$	$\left(\frac{\partial q(i,j,\theta)}{\partial \theta_u} \right)$