IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY

Department of Mathematics

LATENT STRUCTURE MODELS IN TIME SERIES

by

Adelchi Azzalini

Thesis submitted for the degree of

Doctor of Philosophy in the University of London

JANUARY, 1982

ABSTRACT

In the first three chapters, the analysis of a number, say n, of independent time series is considered in a Normal theory context. Chapter 1 deals with the simplest case of n replicates of a first or second order autoregressive stationary process. The emphasis is on asymptotics as n tends to infinity and the length of the series is fixed. The likelihood conditional on the initial values of each of the n series produces estimates of the parameters which are appreciably less efficient than the ones derived from the unconditional likelihood. Chapter 2 and 3 consider more complex situations (allowing for non-stationarity and non-identical distributions) with special attention for Markov processes, using the unconditional likelihood. Chapter 2 deals with systematic effect models, while Chapter 3 deals with latent structure models.

In Chapter 4, a stationary Markov process $\{\theta_t\}$, not necessarily Normal, is observable only indirectly, via a process $\{Y_t\}$. More precisely, the current value Y_t is sampled from a distribution determined by the current value of θ_t, and it is conditionally independent of all past values of $\{\theta_t\}$ and $\{Y_t\}$. A procedure is proposed for approximating the conditional densities of θ_t and θ_{t+1} given $Y_1,\ldots,Y_t$.

A method for approximating the likelihood of a stationary time series is studied in Chapter 5. General results on the asymptotic properties of the associated estimates are given, and numerical comparisons with other estimators are carried out in special cases.

In Appendix A, it is proved that the likelihood of a stationary Normal Markov process with unknown mean, variance and autocorrelation is unimodal. Appendix B studies a Markov process with Beta marginal distribution; its transition law and the autocovariance function are determined.

ACKNOWLEDGMENTS

The research for this thesis has been conducted, and the thesis itself has been compiled, under the supervision of Professor D.R. Cox. It would take several pages to describe adequately how much this thesis owes to his invaluable guidance and to his infinite patience. In an attempt to imitate his well-known conciseness, I shall only say:

"Thank you, Professor Cox".

I would also like to express my gratitude to Professor G. Picci for clarifying points on Kalman filtering and related subjects, to Dr. G. Sawitzki for various remarks on the content of Chapter 4, and to a referee of a paper containing most of Chapter 1 and Appendix A for thorough checking and substantial improvements to the first draft of the paper.

My thanks are also due to the Department of Statistics of the University of Padua, Italy, from which I was on leave, for relieving me from my duties as a research assistant for so long, and to the Consiglio Nazionale delle Ricerche, Italy, for financial support.

TABLE OF CONTENTS

		Page
	List of Tables	6
	List of Figures	7
1.	Replicated observations of low order autoregressive time series	
	1.1 Introduction	8
	1.2 First order autoregressive series	9
	1.3 Second order autoregressive series	15
	1.4 Testing for stationarity	18
	1.5 A design problem	20
2.	Deterministic models for non-identically distributed Markov series	
	2.1 Introduction	23
	2.2 Stationary series with different means	24
	2.3 Non-stationary series with different mean vectors	27
	2.4 Series with different autoregressive parameter	30
	2.5 Non-stationarity caused by the autoregressive parameter	32
	2.6 Testing between two models. An example	34
3.	Stochastic models for non-identically distributed Markov series	
	3.1 Empirical Bayes approach for the mean	37
	3.2 Means with drift	40
	3.3 Empirical Bayes approach for the autoregressive parameter	43
4.	Approximate filtering of parameter driven processes	
	4.1 Parameter driven processes	47
	4.2 Gamma-Poisson process	48
	4.3 Higher order approximation	52
	4.4 Other examples	55
	4.5 Estimating the fixed parameters	57

5. m-order maximum likelihood estimation for stationary stochastic processes

5.1 Introduction 60

5.2 Consistency and asymptotic normality 61

5.3 Efficiency for Normal Markov process plus noise 63

5.4 Efficiency for second order binary Markov chain 67

Appendices

A Proving unimodality of the likelihood of stationary Normal Markov processes 70

B A Markov process with Beta marginal distribution 72

References 74

Subsidiary matter in support of the candidature 77

LIST OF TABLES

Table		Page
1.1	Some values of α_a' and $\lvert\alpha_a''\rvert$.	22
4.1	Relative loss of the filter for the Gamma-Poisson case.	51
4.2	Relative loss of the filter for the Beta-Binomial case.	51
5.1	Asymptotic variance of MLE of ρ, and ARE of $\tilde{\rho}$ and $\hat{\rho}_1$, for AR(1) process plus noise.	66
5.2	Asymptotic variance of $\hat{\theta}_2$, and ARE of $\hat{\theta}_0$ and $\hat{\theta}_1$, for second order binary Markov chain.	69

LIST OF FIGURES

Figure		Page
1.1	ARE of estimators for stationary AR(1) Normal processes based on conditional likelihood plotted versus the autoregressive parameter.	13
1.2	Generalized ARE of estimators for stationary AR(1) Normal processes based on the conditional likelihood plotted versus the autoregressive parameter.	14
1.3	Generalized ARE of stationary AR(2) Normal processes based on ℓ_{c1} likelihood, T = 10.	16
1.4	Generalized ARE of estimators for stationary AR(2) Normal processes based on ℓ_{c2}, T = 10.	17
4.1	Plot of $\omega_t'(\cdot \mid y^{(t)})$ and approximations when $\omega_t(\cdot \mid y^{(t)})$ is Gamma(5, 2.5), $\nu = 2$, $\lambda = 1$, $\rho = 0.8$.	54
5.1	ARE of c_2/c_1 for AR(1) Normal process plus noise, when $\omega = 1/4, \frac{1}{2}, 1, 2$.	65

Chapter 1

REPLICATED OBSERVATIONS OF LOW ORDER AUTOREGRESSIVE TIME SERIES

1.1 Introduction

We shall deal with the analysis of a number of independent short time series. Much time series theory is concerned with asymptotic results holding for a single series, as the length of that series tends to infinity. Here, by contrast, we assume that a moderate or large number of series is available, the length of the component sections being possibly quite short.

Such data can arise in two ways. First, we may have a single series observed in a number of sections, the sections being sufficiently far apart to be treated as independent. Second, suppose that treatments are applied to a number of individuals and a response observed every time unit for each individual, the response for any one individual being assumed to be sampled from some discrete time process.

This kind of problem has received limited attention in the literature. References to known work will be made at appropriate places in the text. We only mention now papers which, although concerned with the same topic, have little relationship with the present thesis. Shumway (1970), Borpujari (1977) and Brillinger (1980) essentially test whether n independent non-stationary series have common mean vector through a frequency domain approach, introducing an "analysis of power" table. Anderson (1978) considers multivariate

series, but he works essentially with the likelihood conditional on the initial values, in a terminology explained later. Farrow (1979) refers mainly to first order autoregressive and first order moving-average models with special emphasis on computational aspects.

The present chapter deals with the simplest, rather restrictive, case of replicated series, namely when n independent realizations of univariate series are sampled from a simple stationary autoregressive process of order one or two. We let n tend to infinity keeping the length of the series fixed. Study of this simple case is valuable in part because it clarifies qualitatively the relation to standard asymptotic theory, that is when n=1 and the length of the series tends to infinity. More complex situations, allowing for non-stationary and non-identical distributions, will be considered in subsequent chapters.

1.2 First order autoregressive series

Let $\{y_{i,t}$ for $t=1,\ldots, T_i$; $i=1,\ldots, n\}$ be a set of Normal random variables (r.v.); $y_{i,t}$ is independent of $y_{j,\tau}$ if $i \neq j$. We suppose here that, for each i, $\{y_{i,t}$ for $t=1,\ldots, T_i\}$ is a section of a first order autoregressive processes and these processes are identically distributed. Assume $T_i \geq 2$ for all i, and define $T=\sum T_i/n$.

First, we consider the case of zero mean. We therefore assume the model

$$y_{i,t} = \alpha\, y_{i,t-1} + \varepsilon_{i,t} \quad , \quad |\alpha|<1 \ , \ t=2,\ldots,T_i; \ i=1,\ldots,n; \qquad (1)$$

where, for each i, the $\varepsilon_{i,t}$ are independent and identically distributed (i.i.d.) $N(0,\sigma^2)$, i.e. $\{\varepsilon_{i,t}\}$ is Gaussian white noise. If the initial values $y_{i,1}$ $(i=1,\ldots,n)$ are regarded as sampled from the stationary distribution $N(0, \sigma^2/(1-\alpha^2))$ the loglikelihood for the unknown parameters (σ^2,α) is

$$\ell(0,\sigma^2,\alpha) = -\tfrac{1}{2} n T \log(2\pi\sigma^2) + \tfrac{1}{2} n \log(1-\alpha^2) - \frac{1}{2\sigma^2}\{ d_{11} - 2\alpha d_{12} + \alpha^2 d_{22}\} , \tag{2}$$

where

$$d_{rs} = \sum_{i=1}^{n} \sum_{t=s}^{T_i-r+1} y_{i,t}\, y_{i,t+r-s} \qquad (1 \le r \le s) ,$$

with the usual convention that a sum is zero if the lower limit exceeds the upper limit.

We consider now $(\hat\sigma^2, \hat\alpha)$, the maximum likelihood estimate (MLE) of (σ^2, α) from (2). When $\hat\alpha$ is known, $\hat\sigma^2$ is easily computed, namely

$$\hat\sigma^2 = (d_{11} - 2\hat\alpha d_{12} + \hat\alpha^2 d_{22})/(n T) ;$$

then $\hat\alpha$ satisfies

$$\hat\alpha^3 d_{22} (T-1)/T - \hat\alpha^2 d_{12} (T-2)/T - \hat\alpha (d_{22} + d_{11}/T) + d_{12} = 0 ,$$

which has one real root for each one of the open intervals $(-\infty, -1)$, $(-1, 1)$, $(1,\infty)$. This is a trivial generalization of a known result for n=1 (White, 1961).

If we assume that $T_i = \text{constant} = T$ and let $n \to \infty$, then strong consistency of $(\hat\sigma^2,\hat\alpha)$ follows using, for instance, theorem 4.1 of Berk (1972); the distribution is asymptotically normal with variance matrix

$$V(\hat\sigma^2,\hat\alpha) \sim \frac{1}{n\{T-(T-2)\alpha^2\}(T-1)} \begin{pmatrix} 2\{T-1-(T-3)\alpha^2\}\sigma^4 & -2\alpha\sigma^2(1-\alpha^2) \\ -2\alpha\sigma^2(1-\alpha^2) & T(1-\alpha^2)^2 \end{pmatrix} . \tag{3}$$

In time series theory it is very common to consider the loglikelihood conditional on the first observation. If we do this for each one of the n time series, the loglikelihood is

$$\ell_c(0,\sigma^2,\alpha) = -\tfrac{1}{2} n (T-1) \log(2\pi\sigma^2) + \\ - \frac{1}{2\sigma^2}\{ d_{11} - \Sigma y_{i,1}^2 - 2\alpha d_{12} + \alpha^2 (d_{22} + \Sigma y_{i,1}^2)\} \tag{4}$$

which holds also for $|\alpha| \ge 1$. The MLE $(\hat\sigma_c^2,\hat\alpha_c)$ of (σ^2,α) is given by

$$\hat\sigma_c^2 = \{d_{11} - \Sigma y_{i,1}^2 - 2\hat\alpha_c d_{12} + \hat\alpha_c^2 (d_{22} + \Sigma y_{i,1}^2)\}/\{n(T-1)\} ,$$

$$\hat\alpha_c = d_{12}/(d_{22} + \Sigma y_{i,1}^2) .$$

Strong consistency and asymptotic normality of $(\hat\sigma_c^2,\hat\alpha_c)$ for series of equal length T can be established as follows. Let us denote by h

the probability density function (p.d.f.) from which the $y_{i,1}$ ($i = 1,\ldots,n$) are sampled, so that, for one individual series, the p.d.f. is

$$f(y_1,\ldots,y_T) = h(y_1)\ \frac{1}{(2\pi\ \sigma^2)^{(T-1)/2}}\ \exp\left[-\frac{1}{2\ \sigma^2}\ \sum_{t=2}^{T}(y_t - \alpha\, y_{t-1})^2\right]\ ,$$

having dropped the suffix i. Clearly, the MLE for (σ^2,α) using density f is still $(\hat{\sigma}_c^2,\hat{\alpha}_c)$ so that we can apply theorems for unconditional maximum likelihood estimation to $(\hat{\sigma}_c^2,\hat{\alpha}_c)$. It is easy to check that conditions for theorems 4.1 and 6.1 of Berk (1972) are satisfied if $E(y_1^2) < \infty$, plus the identifiability condition $E(y_1^2) > 0$ if $T = 2$. In particular, if $E(y_1^2) = \sigma^2/(1-\alpha^2)$, the asymptotic variance matrix is

$$V(\hat{\sigma}_c^2, \hat{\alpha}_c) \sim \frac{1}{n(T-1)} \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & 1-\alpha^2 \end{bmatrix} . \qquad (5)$$

In fact, strong consistency can be established by applying a conditional inference argument to ℓ_c and using a theorem of Andersen (1973, p.51), under the single assumption that $E(y_1^2) > 0$ when $T = 2$.

Consider now the case of unknown mean, i.e. the model

$$y_{i,t} - \mu = \alpha(y_{i,t-1} - \mu) + \varepsilon_{i,t}\ , \qquad t = 2,\ldots,T_i;\ i = 1,\ldots,n; \qquad (6)$$

where μ is an additional unknown parameter, $nT > 2$, $E(y_{i,1}) = \mu$ and everything else is as before. The unconditional loglikelihood is

$$\ell(\mu,\sigma^2,\alpha) = \ell(0,\sigma^2,\alpha) - (2\sigma^2)^{-1}(1-\alpha)\{-2\mu(d_1 - \alpha d_2) + \mu^2 n\left(T - (T-2)\alpha\right)\}\ , \qquad (7)$$

where

$$d_r = \sum_{i=1}^{n}\ \sum_{t=r}^{T_i-r+1} y_{i,t} \qquad (1 \le r)\ .$$

Look for the MLE of (μ,σ^2,α). For any fixed α, the values of μ,σ^2 which maximize (7) are

$$\hat{\mu}(\alpha) = \frac{d_1 - \alpha d_2}{n\{T-(T-2)\alpha\}}\ , \quad \hat{\sigma}^2(\alpha) = \frac{1}{nT}\{d_{11} - 2\alpha d_{12} + \alpha^2 d_{22} - n\hat{\mu}(\alpha)^2\ (1-\alpha) \times \left(T-(T-2)\alpha\right)\}\ , \qquad (8)$$

respectively. Substituting these into $\partial\ell/\partial\alpha = 0$, we find after some

algebra that the MLE $\hat{\alpha}$ of α satisfies

$$C(\hat{\alpha}) \equiv \sum_{j=0}^{5} c_j \hat{\alpha}^j = 0 , \tag{9}$$

where

$$\begin{aligned}
c_5 &= (T-2)(T-1)\{(T-2)d_{22} - b_{22}\}/T , \\
c_4 &= -2(T-2)(T-1)d_{22} + (2T^2-3T+2)b_{22}/T - (T-2)^3 d_{12}/T + \\
&\quad + (T-2)^2 b_{12}/T , \\
c_3 &= -(T-2)^2 d_{11}/T + (3T-4)d_{22} + (T-2)b_{11}/T - b_{22} + \\
&\quad - 2(T^2-2T+2)b_{12}/T + 2(T-2)^2 d_{12} , \\
c_2 &= 2(T-2)d_{11} - (T-2)b_{11}/T - 2(T-2)d_{12} + 2T(T-2)d_{22} - (2T-1)b_{22} , \\
c_1 &= -Td_{11} - T^2 d_{22} + b_{11} + 2Tb_{12} - 2T(T-2)d_{12} + Tb_{22} , \\
c_0 &= T^2 d_{12} - T b_{12} - b_{11}
\end{aligned} \tag{10}$$

and

$$b_{11} = d_1^2/n , \quad b_{12} = d_1 d_2/n , \quad b_{22} = d_2^2/n . \tag{11}$$

In Appendix A, it is proved that (9) has one and only one real root in the open interval $(-1, 1)$, which therefore corresponds to a maximum of ℓ.

Since the information matrix is block diagonal, the asymptotic variance matrix of $(\hat{\sigma}^2, \hat{\alpha})$ is given by (3) and

$$\mathrm{var}(\hat{\mu}) \sim \sigma^2 \{n (1-\alpha) \left(T - (T-2)\alpha\right)\}^{-1} \tag{12}$$

as $n \to \infty$, for series of equal length T.

Consider now the loglikelihood conditional on the starting values, that is

$$\begin{aligned}
\ell_c(\mu,\sigma^2,\alpha) &= \ell_c(0,\sigma^2,\alpha) - (2\sigma^2)^{-1} (1-\alpha) \times \\
&\quad \times \{-2\mu \left(d_1 - \Sigma y_{i,1} - \alpha(d_2 + \Sigma y_{i,1})\right) + n(T-1)\mu^2(1-\alpha)\} .
\end{aligned} \tag{13}$$

The MLE $(\hat{\mu}_c, \hat{\sigma}_c^2, \hat{\alpha}_c)$ of (μ, σ^2, α) can be easily written down in explicit form. The asymptotic variance matrix of $(\hat{\sigma}_c^2, \hat{\alpha}_c)$, when the $y_{i,1}$ $(i=1,\ldots,n)$ are sampled from $N(\mu, \sigma^2/(1-\alpha^2))$, is given by (5) and

$$\mathrm{var}(\hat{\mu}_c) \sim \sigma^2 \{n(T-1) (1-\alpha)^2\}^{-1} . \tag{14}$$

Figure 1.1 and 1.2 give some numerical comparison of $(\hat{\mu}_c, \hat{\sigma}_c^2, \hat{\alpha}_c)$

Figure 1.1: ARE of estimators for stationary AR(1) Normal process based on conditional likelihood plotted versus the autoregressive parameter.

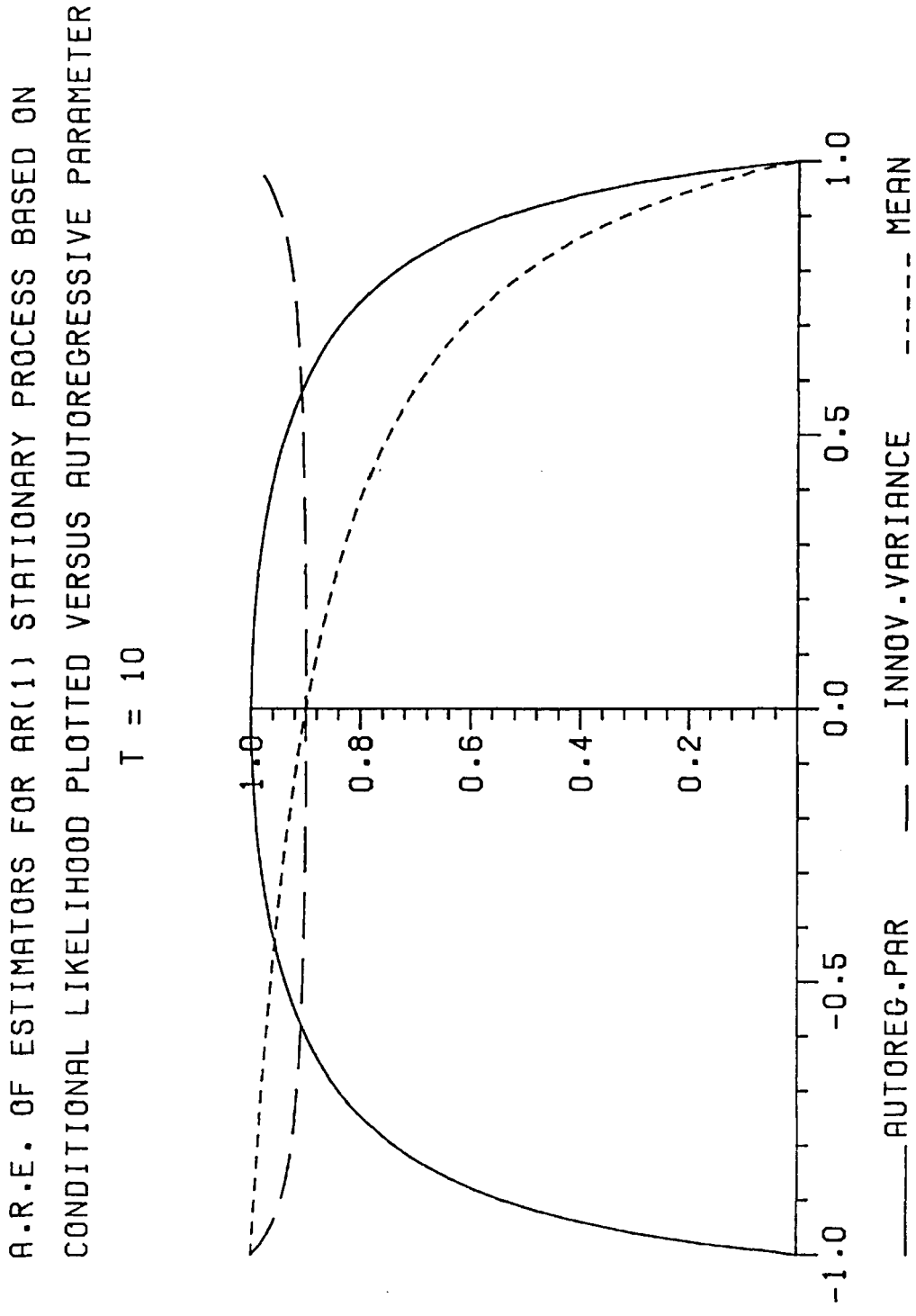
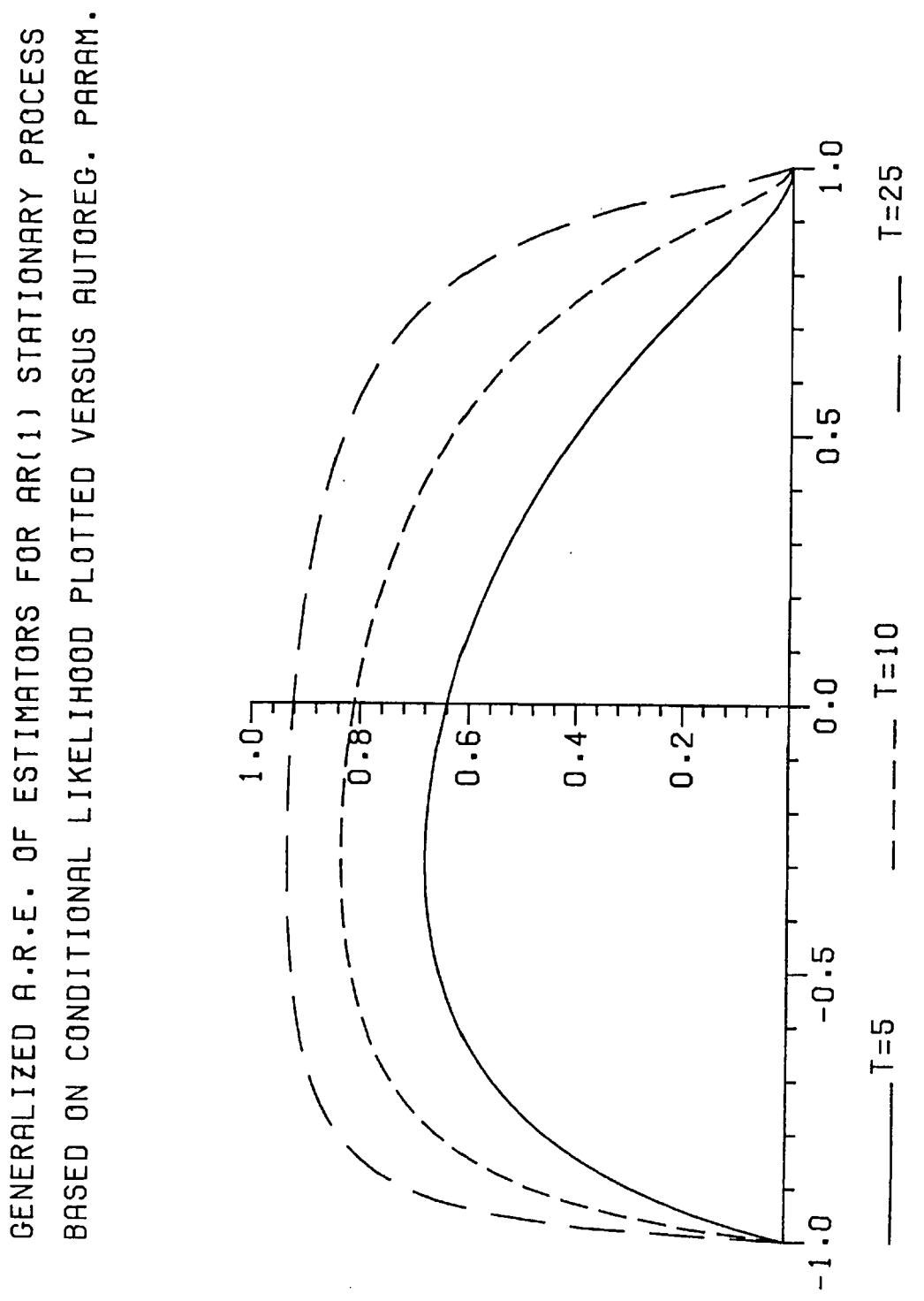


Figure 1.2 Generalized ARE of estimators for stationary AR(1) Normal process based on conditional likelihood plotted versus the autoregressive parameter.



versus $(\hat{\mu},\hat{\sigma}^2,\hat{\alpha})$. In Figure 1.1, the asymptotic relative efficiency (ARE) of each component separately is plotted versus α, for $T = 10$. Figure 1.2 gives the generalized ARE, defined as the ratio of the asymptotic generalized variances when $T = 5, 10, 25$. These figures show that, even for moderate values of T, there is a not negligible loss of efficiency in using ℓ_c instead of ℓ when the latter is appropriate, at least when α is positive, which is the most common case in practice. The fact that no explicit expression for $\hat{\alpha}$ is available is not a serious problem because (9) can be solved efficiently. Therefore ℓ_c should be used only when stationarity cannot be assumed.

1.3 Second order autoregressive series

We now suppose that the data are generated by the model

$$y_{i,t} = \alpha_1 y_{i,t-1} + \alpha_2 y_{i,t-2} + \varepsilon_{i,t}\ ,\ t = 3,\ldots,T_i\ ;\ i = 1,\ldots,n, \qquad (15)$$

where $T_i \geq 3$ $(i = 1,\ldots,n)$, the $\varepsilon_{i,t}$ are i.i.d. $N(0,\sigma^2)$ and $|\alpha_2| \leq 1$, $\alpha_1 + \alpha_2 < 1$, $\alpha_2 - \alpha_1 < 1$. We consider three different types of likelihood for model (15): (i) unconditional likelihood, when the starting values $y_{i,1}$, $y_{i,2}$ $(i = 1,\ldots,n)$ are sampled from the stationary distribution, which is Normal with zero mean vector and variance matrix

$$\sigma^2 \begin{pmatrix} 1-\alpha_2^2 & -\alpha_1(1+\alpha_2) \\ -\alpha_1(1+\alpha_2) & 1-\alpha_2^2 \end{pmatrix}^{-1} ;$$

see, for instance, Box and Jenkins (1970, p.275); (ii) likelihood conditional on $y_{i,1}$ $(i = 1,\ldots,n)$ assuming that $y_{i,2}$ $(i = 1,\ldots,n)$ are sampled from the stationary distribution conditional on the $y_{i,1}$'s; (iii) likelihood conditional on $y_{i,1}$, $y_{i,2}$ $(i = 1,\ldots,n)$. Let us denote by ℓ, ℓ_{c1}, ℓ_{c2} the three loglikelihoods, respectively. Explicit expressions of the asymptotic variance matrices of the estimates are intri-

Figure 1.3: Generalized ARE of estimators for stationary AR(2) Normal processes based on ℓ_{c1} likelihood, T = 10.

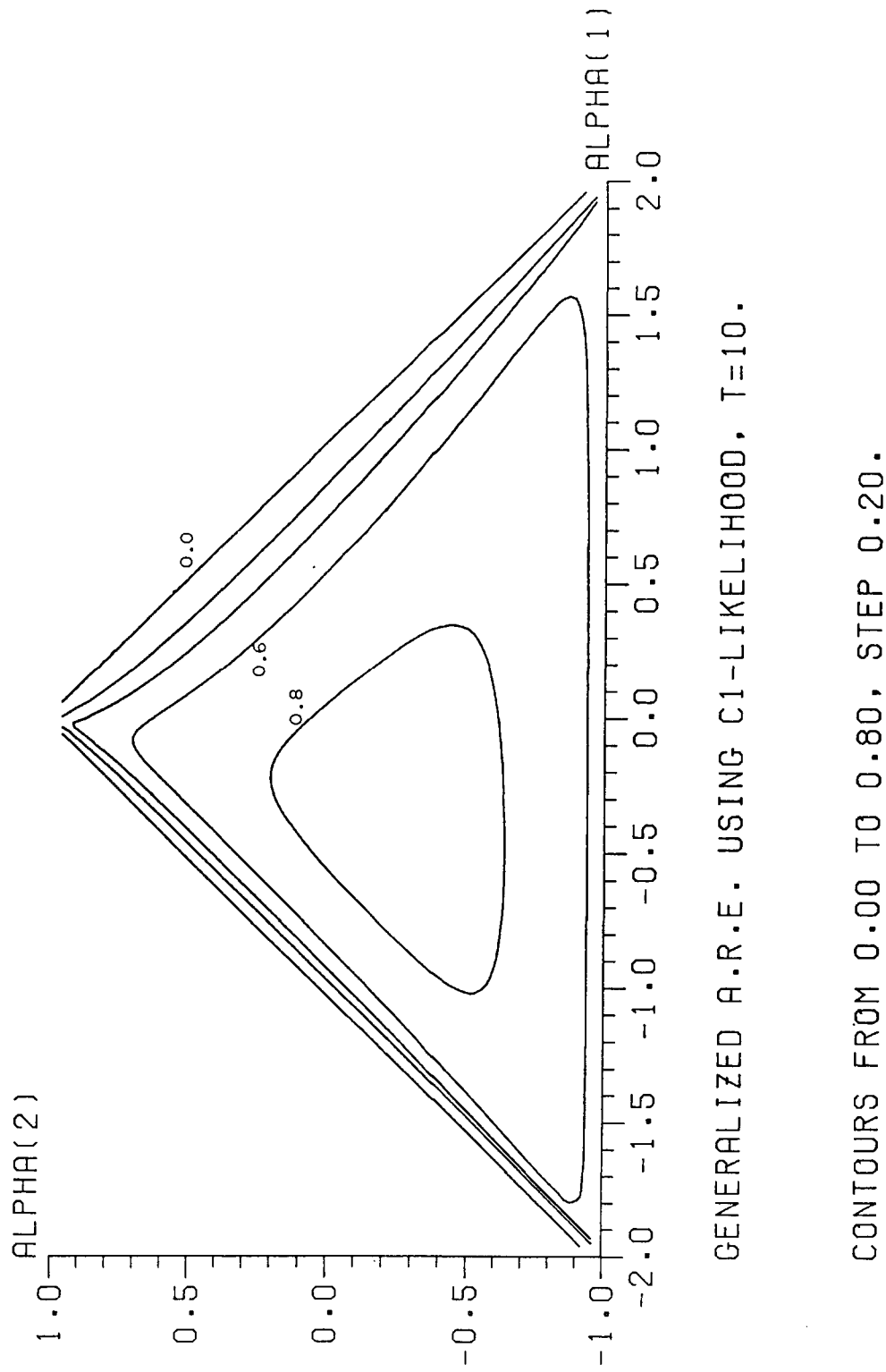
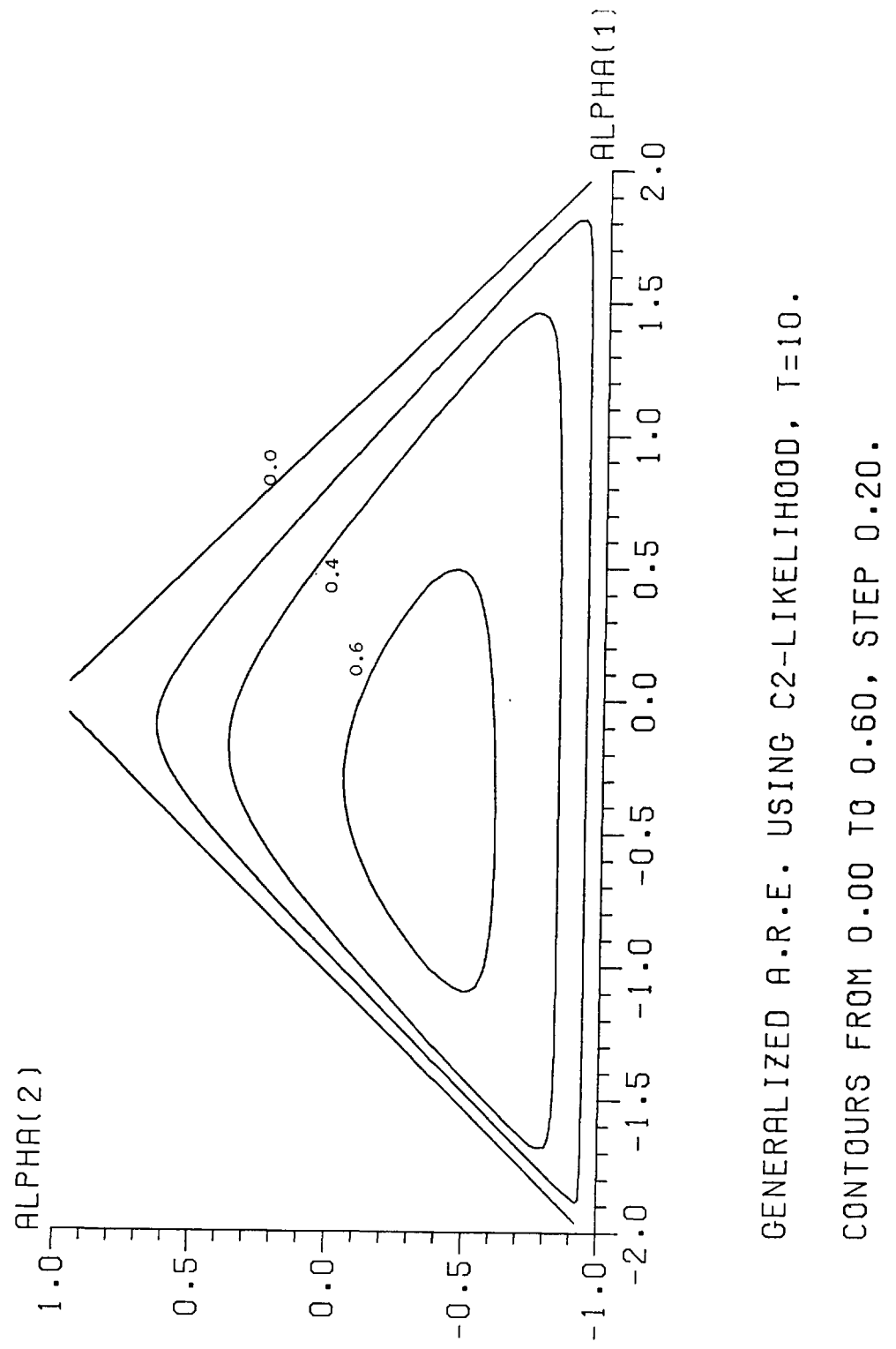


Figure 1.4: Generalized ARE of estimators for stationary AR(2) Normal processes based on ℓ_{c2} likelihood, T = 10.



cate but numerical evaluation is straightforward.

If we allow the mean to be non-zero, say μ, the asymptotic variance matrices of MLE's of $(\sigma^2, \alpha_1, \alpha_2)$ are unchanged, because the information matrix is block diagonal. Asymptotic variances of MLE's of μ are easily computed explicitly.

Figure 1.3 and 1.4 show the generalized ARE of estimates derived from ℓ_{c1} and ℓ_{c2} versus that from ℓ when the mean is unknown, the starting values are sampled from the stationary distribution and $T_i \equiv 10$. The main feature is that ℓ_{c2} has a very low efficiency over a wide area of the parameter space; ℓ_{c1} gives estimates of higher efficiency, but it makes the same assumptions on the parameter space as ℓ without simplifying the actual computation of the MLE. However,it is of interest when $y_{i,1}$ is not sampled from the stationary distribution.

Indeed, the actual computation of the estimates is more complicated than in section 2. If we use ℓ and $\mu = 0$ is assumed, then $(\hat{\alpha}_1, \hat{\alpha}_2)$ is given by the simultaneous solution of a cubic and a quartic. If μ is not known, the values of μ which maximizes ℓ for any given (α_1, α_2) is a simple function of (d_1, d_2, d_3); similarly for σ^2. The parameter space of (α_1, α_2) can be searched starting from the MLE derived from ℓ_{c2}, which is obtained by solving two linear equations in two variables.

1.4 Testing for stationarity

One of the purposes of section 2 and 3 was to evaluate the loss of efficiency using ℓ_c (ℓ_{c1} and ℓ_{c2}, respectively) instead of ℓ when the series are actually stationary. On the other hand, ℓ will be misleading when the series are non stationary. We might want a tool for testing whether the series are stationary or not. In standard asymptotic theory (i.e.

when $n = 1$, $T \to \infty$) there is hardly any need for a formal significance test for the hypothesis $|\alpha| < 1$, because the plot of the data will usually be sufficient; by contrast, departure of the distribution of the initial values from the stationary one usually has little chance of being detected. In the present framework the situation is, to a large extent, reversed: for small T, the plot of the data could easily appear stationary for $|\alpha|$ equal to or just larger than 1, while the distribution of the initial values can be consistently estimated as $n \to \infty$.

Therefore, writing $\mu_1 = E(y_{i,1})$, $\sigma_1^2 = \text{var}(y_{i,1})$, we want to test the null hypothesis

$$H_o = \; |\alpha| < 1, \; \mu_1 = \mu, \; \sigma_1^2 = \sigma^2/(1-\alpha^2) \; ,$$

against $\overline{H}_o$. It appears that the theory of tests does not provide any technique for this system of hypotheses. On the other hand, failure of $|\alpha| < 1$ has an entirely different physical meaning of departure of the initial distribution from the stationary one. So it makes sense to test $|\alpha| < 1$ first, irrespective of μ_1 and σ_1^2. Then, in principle, one could consider the statistic of Davies (1977) for testing $\alpha = 1$ (note that the likelihood does not depend on μ at $\alpha = 1$), but the large number of nuisance parameters makes the approach intractable. Alternatively, one could consider the loglikelihood

$$\ell^*(\mu_1,\sigma_1^2,\mu,\sigma^2,\alpha) = -\tfrac{1}{2} n \log(2\pi\sigma_1^2) - (2\sigma_1^2)^{-1}\textstyle\sum_i (y_{i,1} - \mu)^2 + \ell_c(\mu,\sigma^2,\alpha) \; ,$$

for which the MLE of α is $\hat{\alpha}_c$, and derive a confidence interval for α from a standard asymptotic argument. The only special point is that, although the information matrix is singular at $\alpha = 1$,

$$\lim_{\alpha \to 1} \{\text{asymptotic variance of } \hat{\alpha}_c\} = \frac{1}{n}\left((T-1)\left(\frac{T-2}{2} + \frac{\sigma_1^2}{\sigma^2}\right)\right)^{-1}$$

irrespective of the side from which $\alpha \to 1$.

If $|\alpha| < 1$ has been assumed, one wants to test whether $\mu_1 = \mu$, $\sigma_1^2 = \sigma^2/(1-\alpha^2)$, either simultaneously or separately. Derivation of

testing procedures is routine work and details are not given.

1.5 A design problem

In section 2 and 3, we have taken the number n and the sequence $T_1,\ldots,T_n$ as given. It might be that we can choose them before sampling, under a certain constraint on the total cost of sampling. What is the optimum allocation strategy? There are many possible specifications. Even if we decide to take "optimum" as synonymous with having minimum variance, it is unlikely we are equally interested in all parameters.

Consider the following special version of this problem. For model (6) with initial values taken from the stationary distribution, let a' be the cost of sampling the first observation of each series and b the cost of sampling any observation after the first one. Then, after division by b, the total cost is

$$C = (a + T)\, n \qquad (16)$$

for $a = (a'/b) - 1$, $a > -1$. Given a value $C > 0$, we look for the value of which maximizes the diagonal element of the information matrix, $I_{\theta\theta}$ say, of a parameter θ when the series are of equal length. No *a priori* distribution on the parameter space is introduced, because the goal is to examine the dependence of the optimum sampling strategy on the parameter of interest rather than working out an operating rule.

Substituting (16) into the information matrix we get

$$I_{\mu\mu} = \sigma^{-2}\,(1-\alpha)\,\{(1-\alpha)\,C - n\,(a(1-\alpha) - 2\alpha)\},$$

$$I_{\sigma^2\sigma^2} = \tfrac{1}{2}\sigma^{-4}\,(C - a\,n)\ ,$$

$$I_{\alpha\alpha} = (1-\alpha^2)^{-1}\left\{C - n\left(a + 1 - \frac{2\alpha^2}{1-\alpha^2}\right)\right\}\ ,$$

for $1 \leq n \leq C/(a+2)$. The following points hold: (i) in all cases the optimum sampling strategy is either $n = 1$, $T = C$ or $n = C/(a+2)$, $T = 2$;

(ii) for estimating σ^2, the choice depends only on the sign of a;

(iii) for estimating μ and α, the choice is in favour of $n = C/(a+2)$, $T = 2$, when

$$\frac{\alpha}{1-\alpha} > \frac{a}{2} \qquad \text{or} \qquad \frac{\alpha^2}{1-\alpha^2} > \frac{a+1}{2} \quad , \tag{17}$$

respectively. Let α'_a be the largest value of α such that the first one of (17) is false; analogously $|\alpha''_a|$ for the second inequality. Some numerical values are given in Table 1.1 .

Inspection of Table 1.1 shows that in many practical situations α can be assumed to be greater than α'_a, but $|\alpha| > |\alpha''_a|$ is less common. Supposing that an *a priori* distribution on α has been introduced, conflicting interests in the estimation on μ and α can arise if these parameters are of equal importance for us.

Table 1.1: Some values of α'_a and $|\alpha''_a|$

a	-0.5	0	0.500	1	2
α'_a	-0.333	0	0.200	0.333	0.500
$\lvert\alpha''_a\rvert$	0.447	0.577	0.655	0.707	0.775

Chapter 2

DETERMINISTIC MODELS FOR NON-IDENTICALLY DISTRIBUTED MARKOV SERIES

2.1 Introduction

The situations considered in Chapter 1 are too restricted to be of interest in applications. If we think of each of the n times series as the history (value of the response variable) of an individual over a period of time T, then an allowance must be made for differences between individuals which might influence the response variable. In a clinical experiment, where individuals are given a treatment and then observed every hour for 24 hours, the response variable(s) might be influenced by factors such as sex, age, presence of some disease, and so on.

A second aspect to be taken into account is possible non-stationarity of the series. In the above example, the effect of the treatment is possibly high at the beginning of the observation period and then dicreases. Of course, a further generalization allows non-stationarity to be different from series to series. Similar arguments apply if we think of the n series as fairly spaced sections of a unique realization of one process.

This kind of problem has been studied in the sociological and econometric literature under the name of "panel data". In particular, the issue of April-September 1978 of the Annales de l'INSÉÉ has been entirely dedicated to this topic. The emphasis has been mainly on continuous dependent variables, although there are exceptions, notably

that of Heckmann(1978) who develops a continuous latent structure model for dichotomous observations. For continuous observations, very general models have been considered based on normality assumption or associated second order theory. However, for all papers of which the author is aware either no autocorrelation is supposed for data of the same individual along time, or autoregressive models are considered, but the likelihood conditional on the p initial values is considered, if p is the order of the autoregression. Examples are Jöreskog (1978, p. 360) for a fixed effect model, Marodier and Trognon (1978) for a random effect model; Wansbeek and Kaptein (1978) consider both. Ignoring the contribution to the likelihood from the initial values simplifies dramatically the mathematical treatment of the problem, but has the drawback of producing inefficient estimates if T is not moderately large, as shown in Chapter 1. For this reason, we shall work with the unconditional likelihood as long as possible. Fixed effect models are considered in this chapter; random effect models in the following chapter. Sections 2 and 3 of this chapter deal with modelling of the mean of the process. Section 4 and 5 do the analogous thing for the autoregressive parameter.

2.2 Stationary series with different means

Assume that the n series satisfy the model

$$y_{i,t} - \mu_i = \alpha (y_{i,t} - \mu_i) + \varepsilon_{i,t} , \qquad t=2,\ldots,T_i;\ i=1,\ldots,n, \tag{1}$$

where $|\alpha|<1$, and the $\varepsilon_{i,t}$ are white noise $N(0,\sigma^2)$; $y_{i,1}$ is sampled from the stationary distribution $N(\mu_i,\sigma^2/(1-\alpha^2))$. For the vector of the means $\underline{\mu}=(\mu_1,\ldots,\mu_n)'$, we assume the linear model

$$\underline{\mu} = X\underline{\beta} , \tag{2}$$

where X is a full rank $n \times p$ matrix ($n \geq p$) of covariates and $\underline{\beta}$ is a $p \times 1$ vector of unknown parameters.

If X is the identity matrix of order n, then this model reduces to a one-way classification with autocorrelated errors within each row. Box (1954) has studied the effect of autocorrelation on the significance level of the usual analysis of variance (i.e. ignoring autocorrelation) when testing equality of the means of the rows, showing that even a small autocorrelation has a dramatic effect on the actual significance level of the F-test. Box worked with a first order moving average model but it is usually recognised that, qualitatively, the result holds more generally.

A two-way classification with autocorrelated errors within each row has been considered by Larsen (1969) in a Bayesian framework, the main purpose being computation of the posterior distribution of the parameters given a locally uniform prior density. See also Ljung and Box (1980) who assume a circular autoregressive process within each row and use Jeffreys uninformative prior distribution.

For series of equal length T ($T \geq 2$), the loglikelihood for model (1) is

$$\begin{aligned} \ell(\underline{\beta},\sigma^2,\alpha) = & -\tfrac{1}{2}nT \log (2\pi\sigma^2) + \tfrac{1}{2}n \log (1-\alpha^2) + \\ & - (2\sigma^2)^{-1} \{d_{11} - 2\alpha d_{12} + \alpha^2 d_{22} + (1-\alpha)\times \\ & \times \left((T-(T-2)\alpha)\underline{\beta}'X'X\underline{\beta} - 2\underline{\beta}'X'(\underline{d}_1 - \alpha\underline{d}_2)\right)\} , \end{aligned}$$

where d_{rs} is defined as in Chapter 1 and

$$\underline{d}_r = \left(\sum_{t=r}^{T-r+1} y_{1,t} , \ldots , \sum_{t=r}^{T-r+1} y_{n,t} \right)' .$$

Look for the MLE of $\underline{\beta}$, σ^2, α. For any fixed α, the values of $\underline{\beta}$, σ^2 which maximize the loglikelihood are

$$\hat{\underline{\beta}}(\alpha) = \{T - (T-2)\alpha\}^{-1} (X'X)^{-1} (\underline{d}_1 - \alpha\underline{d}_2) ,$$

$$\begin{aligned} \hat{\sigma}^2(\alpha) = (nT)^{-1} & \{d_{11} - 2\alpha d_{12} + \alpha^2 d_{22} - (1-\alpha)\times \\ & \times (b_{11} - 2\alpha b_{12} + \alpha^2 b_{22})/(T-(T-2)\alpha)\} , \end{aligned}$$

respectively, where

$$b_{rs} = \underline{d}_r' X (X'X)^{-1} X' \underline{d}_s \ . \qquad (3)$$

Substituting $\hat{\underline{\beta}}(\alpha)$, $\hat{\sigma}^2(\alpha)$ into $\partial \ell / \partial \alpha = 0$, we find after some algebra that the MLE $\hat{\alpha}$ of α satisfies

$$\sum_{j=0}^{5} c_j \hat{\alpha}^j = 0 \ , \qquad (4)$$

where the coefficients c_j's are the same of Chapter 1, except that b_{rs} is now given by (3). By similar arguments to those of Appendix A, it can be proved that, with probability 1, equation (4) has one and only one real root in the open interval (-1,1). Note that the whole procedure needs only one matrix inversion.

An alternative estimate of $\underline{\beta}$ is

$$\tilde{\underline{\beta}} = (X'X)^{-1} X' \underline{m} \ ,$$

where $\underline{m} = T^{-1} \underline{d}_1$ is the vector of the sample means; $\tilde{\underline{\beta}}$ does not depend on the autocorrelation. Its variance matrix is

$$V(\tilde{\underline{\beta}}) = \frac{\sigma^2}{T(1-\alpha^2)} \left\{ \frac{1+\alpha}{1-\alpha} - \frac{2\alpha (1-\alpha^T)}{T(1-\alpha)^2} \right\} (X'X)^{-1} \ .$$

We compare $\tilde{\underline{\beta}}$ with the uniformly minimum variance unbiased (UMVU) estimate of $\underline{\beta}$ for fixed α, i.e. $\hat{\underline{\beta}}(\alpha)$, whose variance matrix is

$$V\{\hat{\underline{\beta}}(\alpha)\} = \sigma^2 \{(1-\alpha)(T-(T-2)\alpha) X'X\}^{-1} \ .$$

The ratio of the determinants of these two variances is the same as the ARE of the sample mean versus the UMVU estimate of the mean for one individual series. Chipman *et al.* (1968) have shown that this ARE is very high for $0 < \alpha < 1$, with greatest lower bound 0.87769. Only for α close to -1, there is a strong superiority of $\hat{\underline{\beta}}(\alpha)$ over $\tilde{\underline{\beta}}$, but this situation is fairly rare in practice. All of this suggests that $\tilde{\underline{\beta}} = \hat{\underline{\beta}}(\alpha)$ is not much better than $\tilde{\underline{\beta}}$ for $0 < \alpha < 1$. However, we must not conclude that $\hat{\underline{\beta}}$ is unimportant; for instance, in the one-way classification mentioned earlier in this section, $\tilde{\underline{\beta}}$ reduces to $\underline{m}$ with no estimate of the residual variance, while a likelihood ratio test is possible via $\hat{\underline{\beta}}$.

If the series have different length $T_1,\ldots,T_n$, then

$$\underline{\hat{\beta}}(\alpha) = \{(1-\alpha)X'D_TX + 2\alpha X'X\}^{-1}\, X'(\underline{d}_1 - \alpha\underline{d}_2) \ , \qquad (5)$$

$$\hat{\sigma}^2(\alpha) = (\Sigma T_i)^{-1}\, \{d_{11} - 2\alpha d_{12} + \alpha^2 d_{22} - (1-\alpha)\underline{\hat{\beta}}(\alpha)'X'(\underline{d}_1 - \alpha\underline{d}_2)\} \ ,$$

where $D_T = \mathrm{diag}(T_1,\ldots,T_n)$ and the definition of $\underline{d}_r$ is modified in a obvious way. Unfortunately, the inverse matrix in (5) is difficult to handle and no equation corresponding to (4) could be found. For computing the MLE of α, we evaluate the maximum of the likelihood for any fixed α, that is

$$\begin{aligned} \ell_{max}(\alpha) &= \ell(\underline{\hat{\beta}}(\alpha),\ \hat{\sigma}^2(\alpha),\ \alpha) \\ &= -\tfrac{1}{2}(\Sigma T_i)\log\hat{\sigma}^2(\alpha) + \tfrac{1}{2}n\log(1-\alpha^2) + \text{constant} \ , \end{aligned}$$

and maximize $\ell_{max}(\alpha)$ numerically. This involves a new $p\times p$ matrix inversion for all values of α.

The information matrix is block diagonal. The block for $(\sigma^2,\ \alpha)$ is the same of Chapter 1. The block for $\underline{\beta}$ is

$$\sigma^{-2}\ (1-\alpha)\ \{(1-\alpha)X'D_TX + 2\alpha\, X'X\} \ .$$

2.3 Non-stationary series with different mean vectors

In a more elaborate situation, a time-effect can be combined with an individual-effect, the simplest form of combination being addition. For series of equal length T, consider then

$$y_{i,t} - \mu_{i,t} = \alpha(y_{i,t-1} - \mu_{i,t-1}) + \varepsilon_{i,t} \ ,$$

where $\mu_{i,t}$ is decomposed as

$$\mu_{i,t} = \mu_i + \mu_t^* \ .$$

Suppose that, for $\underline{\mu} = (\mu_1,\ldots,\mu_n)'$, $\underline{\mu}^* = (\mu_1^*,\ldots,\mu_T^*)'$,

$$\underline{\mu} = X\underline{\beta} \ , \qquad \underline{\mu}^* = Z\underline{\zeta} \ ,$$

where X and $\underline{\beta}$ are as in the previous section, Z is a full rank $T\times q$ matrix $(T \geq q)$ of known functions of t at $t = 1,\ldots,T$, and $\underline{\zeta}$ is a $q\times 1$ vector of unknown parameters. Typically, the entries of the columns

of Z are values of "proper" functions of t, such as $\sqrt{t}$, $\cos t$, e^{-t}, and so on. But Z can also be used to allow for stimuli with transient effect; for instance, if we measure the blood glucose level of n individuals every hour, the t-th entry of a column of Z may represent the amount of food eaten in the hour preceding time t. Biological growth curves provide other applications of the model; the "natural" development of an individual is interpreted via $Z\underline{\zeta}$, while differences between individuals such as breed or initial weight are accounted via $X\underline{\beta}$.

In the above model, non-stationarity is allowed for, but its form is restricted to be the same for all n series. In practice, such a restriction needs often to be violated, since different individuals may receive different stimuli during the observation period, or they may receive them at different times. Therefore we generalize further the model, letting Z to depend on the individual. In other words, we write

$$\mu_{i,t} = \mu_i + \mu^*_{i,t} , \qquad \underline{\mu}^*_i = Z_i \underline{\zeta}$$

with $\underline{\mu}^*_i = (\mu^*_{i,1},\ldots,\mu^*_{i,T})'$. There is a danger of overparameterization, which would occur, for instance, if X and all Z_i's had a column of 1's; but these columns of 1's are not the only possible trap. We shall return to this point; for the moment, assume that no overparameterization occurs.

The loglikelihood is

$$\ell(\underline{\beta},\underline{\zeta},\sigma^2,\alpha) = -\tfrac{1}{2}nT \log (2\pi\sigma^2) + \tfrac{1}{2} n \log(1-\alpha^2) + \\ -(2\sigma^2)^{-1} \sum_{i=1}^{n} (\underline{y}_i - W_i\underline{\beta} - Z_i\underline{\zeta})' \Omega^{-1} (\underline{y}_i - W_i\underline{\beta} - Z_i\underline{\zeta}) , \tag{6}$$

where $W_i = \underline{1}\,\underline{x}_i'$ is a $T \times p$ matrix with all rows equal to the i-th row of X, and $\sigma^{-2}\,\Omega^{-1}$ is the inverse of the variance matrix. For a first order autoregressive process,

$$\Omega^{-1} = \begin{pmatrix} 1 & -\alpha & 0 & \cdots & 0 & 0 \\ -\alpha & 1+\alpha^2 & -\alpha & \cdots & 0 & 0 \\ 0 & -\alpha & 1+\alpha^2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\alpha & 1 \end{pmatrix} .$$

For any fixed α, computation of the MLE of the remaining parameters is simple. Define

$$H_{ZW} = \sum_{i=1}^{n} Z_i' \Omega^{-1} W_i ,$$

similarly for H_{WZ}, H_{ZZ}, H_{WW}. Note that $H_{WW} = \underline{1}'\Omega^{-1}\underline{1}X'X = \omega X'X$, where ω is the sum of all elements of Ω^{-1}. Differentiation of the loglikelihood gives, in a obvious notation,

$$\begin{aligned} \hat{\underline{\beta}}(\alpha) &= (H_{WW} - H_{WZ} H_{ZZ}^{-1} H_{ZW})^{-1} \sum_{i=1}^{n} (W_i' - H_{WZ} H_{ZZ}^{-1} Z_i') \Omega^{-1} \underline{y}_i , \\ \hat{\underline{\zeta}}(\alpha) &= H_{ZZ}^{-1} \left(\sum_{i=1}^{n} Z_i' \Omega^{-1} \underline{y}_i - H_{ZW} \hat{\underline{\beta}}(\alpha) \right) , \end{aligned} \tag{7}$$

assuming that all inverse matrices exist; hence $\hat{\sigma}^2(\alpha)$. The associated loglikelihood is

$$-\tfrac{1}{2} n T \log \hat{\sigma}^2(\alpha) + \tfrac{1}{2} n \log(1-\alpha^2) + \text{constant} , \tag{8}$$

which has to be maximized numerically. Although the computation is more involved than in the previous cases, it is not prohibitive because a $(p+q+2)$-dimensional search is reduced to a 1-dimensional search for which efficient algorithms exist. Moreover, computations are simplified if one takes advantage of the pattern of Ω^{-1} ; in fact, all products of the form $\underline{u}' \Omega^{-1} \underline{v}$, for any two vectors $\underline{u} = (u_1, \ldots, u_T)'$ and $\underline{v} = (v_1, \ldots, v_T)'$, can be written as

$$\sum_{t=1}^{T} u_t v_t - \alpha \sum_{t=2}^{T} (u_t v_{t-1} + u_{t-1} v_t) + \alpha^2 \sum_{t=2}^{T-1} u_t v_t$$

and the three sums have to be computed only once for all values of α.

The information matrix is block diagonal. The block for (σ^2, α) is as in Chapter 1. The block for $(\underline{\beta}, \underline{\zeta})$ is

$$\sigma^{-2} \begin{pmatrix} H_{WW} & H_{WZ} \\ H_{ZW} & H_{ZZ} \end{pmatrix} ,$$

as can be seen directly because, for any given α, the model is a linear one with design matrix

$$X_o = \begin{pmatrix} W_1 & Z_1 \\ W_2 & Z_2 \\ \cdot & \cdot \\ W_n & Z_n \end{pmatrix} .$$

Therefore, the inverse matrices in (7) exist simultaneously either for all values of α (if X_o has full rank), or for none. This ensures that (8) has no "holes".

Clearly, (7) applies to any Normal process for which Ω^{-1} can be computed. This is especially simple for autoregressive processes (Wise, 1955; Siddiqui, 1958). In (6) and (8), $\log(1-\alpha^2)$ must be replaced by $\log|\Omega^{-1}|$.

2.4 Series with different autoregressive parameters

In this section and the following one, we suppose that the expanatory variables influence the autoregressive parameter, rather than the mean. This makes sense especially when the Markov model has physical motivation and is not simply a mathematical convenience. If such physical motivation exists, it is conceivable that the fraction α for which y_t is indebted to y_{t-1} is influenced by covariates.

Taking the common mean as zero for the moment, we have

$$y_{i,t} = \alpha_i \, y_{i,t-1} + \varepsilon_{i,t}$$

retaining previous assumptions on Normality and stationarity. We want to write α_i as a function of a p-dimensional parameter $\underline{\beta}$ and a set of

covariates $\underline{x}_i'$ $(i=1,\ldots,n)$. Because of the constraint $|\alpha_i| < 1$, it is appealing to write

$$\alpha_i = \tanh(\tfrac{1}{2}\underline{x}_i'\underline{\beta}) = \frac{\exp(\underline{x}_i'\underline{\beta}) - 1}{\exp(\underline{x}_i'\underline{\beta}) + 1} \quad . \tag{9}$$

However, (9) does not allow a tractable solution, because the loglikelihood for $(\underline{\alpha},\sigma^2)$ is

$$\ell(\underline{\alpha},\sigma^2) = -\tfrac{1}{2}\Sigma T_i \log(2\pi\sigma^2) + \tfrac{1}{2}\Sigma_i \log(1-\alpha_i^2) +$$

$$- (2\sigma^2)^{-1} \Sigma_i \left(d_{11}^{(i)} - 2\alpha_i\, d_{12}^{(i)} + \alpha_i^2\, d_{22}^{(i)}\right) ,$$

where $\underline{\alpha} = (\alpha_1,\ldots,\alpha_n)'$ and

$$d_{rs}^{(i)} = \sum_{t=s}^{T_i-r+1} y_{i,t}\, y_{i,t+r-s} \quad .$$

When (9) is replaced into $\ell(\underline{\alpha},\sigma^2)$, no reduction to lower dimensional sufficient statistic is possible; a direct search of the parameter space is necessary for computing the MLE. An approximate solution to this problem is via the loglikelihood conditional on the initial values $y_{i,1},\ldots,y_{n,1}$

$$\ell_c(\underline{\alpha},\sigma^2) = -\tfrac{1}{2}\Sigma(T_i - 1)\log(2\pi\sigma^2) +$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(d_{11}^{(i)} - y_{i,1}^2 - 2\alpha_i\, d_{12}^{(i)} + \alpha_i^2\left(d_{22}^{(i)} + y_{i,1}^2\right)\right) .$$

The conditional MLE of $\underline{\beta}$ is obtained minimizing

$$\Sigma_i \left(-2\alpha_i(\underline{\beta})\, d_{12}^{(i)} + \alpha_i(\underline{\beta})^2 \left(d_{22}^{(i)} + y_{i,1}^2\right)\right) ,$$

that is minimizing

$$\Sigma_i\, d_i \{ a_i - \alpha_i(\underline{\beta})\}^2,$$

where $d_i = d_{22}^{(i)} + y_{i,1}^2$, $a_i = d_{12}^{(i)}/d_i$. This is a non-linear weighted least squares problem, for which routines are available. To initialize the algorithm, one possible way is as follows. For each series independently, compute an estimate $\tilde{\alpha}_i$ of α_i; inverting (9), compute

$$\tilde{\theta}_i = \log \frac{1 + \tilde{\alpha}_i}{1 - \tilde{\alpha}_i} \qquad (i=1,\ldots,n) ;$$

then, in an obvious notation, compute $\underline{\tilde{\beta}} = (X'X)^{-1} X'\,\underline{\tilde{\theta}}$.

If the series have unknown mean μ, we modify the definition of $d_{rs}^{(i)}$ as

$$d_{rs}^{(i)} = \sum_{t=s}^{T_i-r+1} (y_{i,t} - \mu)(y_{i,t+r-s} - \mu)$$

and a search for μ must be done.

An alternative approach is to replace (9) by the less ambitious $\alpha_i = \underline{x}_i' \underline{\beta}$, which is considered in the next section, where it also appears not unreasonable.

2.5 Non-stationarity caused by the autoregressive parameter

In section 3 we decomposed a non-stationary (in the mean) series as the sum of a zero-mean stationary process and a non-constant function of time. An alternative approach interpretes non-stationarity in the mean as the effect of time-dependence of the autoregressive parameter. For a single time series, we then suppose that y_1 is $N(\mu_1,\sigma_1^2)$ and

$$y_t = \alpha_{t-1}\, y_{t-1} + \varepsilon_t \qquad (t=2,\ldots,T) \ , \qquad (10)$$

where $\{\varepsilon_t\}$ are i.i.d. $N(0,\sigma^2)$ and α_t is a slow-varying function of t, parameterized in a few parameters. For instance, if $\mu_1 \gg \sigma_1 = \sigma$ and $\alpha_t = \alpha_0\, t^{-1}$, with α_0 around 5σ, the series will have a χ^2-like shape, which is fairly common in experimental studies. If we believe, say, that the effect of the treatment will be higher for individuals that, for any reason, at the beginning of the observation period are above the average and *vice versa* for individuals that are below the average, then (10) is the appropriate scheme, not the one of section 3.

This model, when α_t is a polynomial in t with coefficients to be estimated, has been considered by Wegman (1974), who has shown how to compute the MLE. A more general model is

$$\underline{\alpha} = (\alpha_1,\ldots,\alpha_{T-1})' = Z\underline{\zeta} \ ,$$

where Z is a $(T-1) \times q$ matrix ($T-1 \geq q$), whose entries are values of q functions of t and $\underline{\zeta}$ is a q-dimensional vector of coefficients to be estimated. We also assume that n replicates of the same process are available. For simplicity, we keep Z constant across series. Simple differentiation on the loglikelihood gives the MLE of $\underline{\zeta}$, which is

$$\underline{\zeta} = (Z' D_{d^*} Z)^{-1} Z' \underline{c}^* ,$$

where

$$D_{d^*} = \mathrm{diag}(d_1^*, \ldots, d_{T-1}^*), \qquad d_t^* = \sum_{i=1}^{n} y_{i,t}^2 ,$$

$$\underline{c}^* = (c_1^*, \ldots, c_{T-1}^*)' , \qquad c_t^* = \sum_{i=1}^{n} y_{i,t+1}\, y_{i,t} .$$

A more complex model allows α to vary with i and t simultaneously, i.e.

$$y_{i,t} = \alpha_{i,t-1}\, y_{i,t-1} + \varepsilon_{i,t} ,$$

with $\alpha_{i,t}$ decomposed in individual-effect and time-effect:

$$\begin{aligned} \alpha_{i,t} &= \alpha_i + \alpha_t^* \\ &= \underline{x}_i' \underline{\beta} + \underline{z}_t' \underline{\zeta} , \end{aligned}$$

where $\underline{x}_i'$ is p-dimensional vector of covariates of the i-th series, $\underline{\beta}$ is p-dimensional vector of parameters, $\underline{z}_t'$ is the t-th row of Z defined before. Define X as the $n \times p$ matrix with i-th row $\underline{x}_i'$.

The loglikelihood is

$$\begin{aligned} \ell(\mu_1, \sigma_1^2, \sigma^2, \underline{\beta}, \underline{\zeta}) = \text{constant} &- \tfrac{1}{2} n \log \sigma_1^2 - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (y_{i,1} - \mu_1)^2 + \\ &- \tfrac{1}{2} n (T-1) \log \sigma^2 + \\ &- \sum_{t=2}^{T} \sum_{i=1}^{n} \{ y_{i,t} - (\underline{x}_i' \underline{\beta} + \underline{z}_{t-1}' \underline{\zeta})\, y_{i,t-1} \}^2 / (2\sigma^2) \end{aligned}$$

with derivatives

$$\partial \ell / \partial \underline{\beta} = \sigma^{-2} (X' \underline{c} - X' D_d X \underline{\beta} - X' Y_2 Z \underline{\zeta}) ,$$

$$\partial \ell / \partial \underline{\zeta} = \sigma^{-2} (Z' \underline{c}^* - Z' Y_2' X \underline{\beta} - Z' D_{d^*} Z \underline{\zeta}) ,$$

where

$$D_d = \mathrm{diag}(d_1, \ldots, d_n) , \qquad d_i = \sum_{t=2}^{T} y_{i,t-1}^2 ,$$

$$\underline{c} = (c_1,\ldots,c_n)' , \qquad c_i = \sum_{t=2}^{n} y_{i,t}\, y_{i,t-1} ,$$

and Y_2 is a $n \times (T-1)$ matrix with (i,t)-th entry $y^2_{i,t}$. The solutions of the MLE equations are

$$\hat{\underline{\zeta}} = (Z'\, D_{d*}\, Z)^{-1}\, Z'\, (\underline{c}^* - Y_2'\, X\, \hat{\underline{\beta}}) ,$$

$$\hat{\underline{\beta}} = \{X'(D_d - Y_2\, Q\, Y_2')X\}^{-1}\, X'\, (\underline{c} - Y_2\, Q\, \underline{c}^*) ,$$

with $Q = Z(Z'D_d\ Z)^{-1}\, Z'$, assuming that the inverse matrices exist. The MLE of μ_1 and σ_1^2 are the sample mean and variance of $y_{1,1},\ldots,y_{n,1}$. However, it is not necessary to assume $E(y_{i,1})$ constant with i. We can write $E(y_{i,1}) = \underline{u}_i'\,\underline{\beta}_1$ and use least squares techniques, provided $\underline{\beta}_1$ is unrelated with $\underline{\beta},\underline{\zeta}$.

The parameter $(\underline{\beta},\underline{\zeta})$ is orthogonal to the others, with observed information matrix

$$\sigma^{-2} \begin{pmatrix} X'\, D_d\, X & X'\, Y_2\, Z \\ Z'\, Y_2'\, X & Z'\, D_{d*}\, Z \end{pmatrix} .$$

2.6 Testing between two models. An example

In section 2 and 3, two different models for non-stationary series have been described. In certain cases, both can be used for fitting a set of data, and a method for testing between them is necessary. Since the models are not nested, the likelihood ratio criterion described by Cox (1961,1962) is one appropriate statistic.

We consider here a very special case as an example. Denote by U and V the two alternative hypotheses. Under hypothesis U, the data are generated by

$$y_{i,t} = \mu_t + \varepsilon_{i,t} , \quad \mu_t = \mu\, u_t \qquad (i=1,\ldots,n;\ t=1,\ldots,T)$$

where μ is an unknown parameter and $u_1,\ldots,u_T$ is a known sequence. Under hypothesis V, $y_{i,1}$ is $N(\mu_1,\sigma_1^2)$ and

$$y_{i,t} = \nu_{t-1}\, y_{i,t-1} + \varepsilon_{i,t}\,, \qquad \nu_{t-1} = \nu\, v_{t-1} \qquad (i=1,\ldots,n;\ t=2,\ldots,T)$$

with ν unknown parameter and $v_1,\ldots,v_{T-1}$ a given sequence. For both hypotheses, $\{\varepsilon_{i,t}\}$ are i.i.d. N(0,1). This is a restrictive assumption but it can be defended as follows. Suppose that both u_t and v_t tend to zero as $t \to T$; we could think that the actual data go well beyond $t = T$, where they become useless for estimating the parameters as well as for discriminating between the hypotheses, except for estimating $\mathrm{var}(\varepsilon_{i,t})$ which therefore can be taken as known and equal to 1.

The loglikelihood for hypothesis U is, up to a constant term,

$$\ell_U(\mu) = -\tfrac{1}{2}\, \Sigma_i\, \Sigma_t (y_{i,t} - \mu\, u_t)^2$$

with associated MLE of μ

$$\hat{\mu} = \frac{\Sigma_t\, u_t\, \bar{y}_{\cdot t}}{\Sigma_t\, u_t^2} \qquad \text{where} \quad \bar{y}_{\cdot t} = \frac{1}{n}\, \Sigma_i\, y_{i,t}\,.$$

The loglikelihood for hypothesis V is, up to a constant term,

$$\ell_V(\mu_1,\sigma_1^2,\nu) = -\tfrac{1}{2}\, n \log(\sigma_1^2) \;- (2\sigma_1^2)^{-1} \textstyle\sum_i (y_{i1} - \mu_1)^2 \;+$$

$$-\tfrac{1}{2} \textstyle\sum_i \sum_t (y_{i,t} - \nu\, v_{t-1}\, y_{i,t-1})^2$$

with MLE of ν

$$\hat{\nu} = \frac{\sum_{t=2}^{T} v_t \sum_i y_{i,t}\, y_{i,t-1}}{\sum_{t=2}^{T} v_t^2 \sum_i y_{i,t-1}^2}\,.$$

For computing Cox's test T_U we need

$$E_\mu\{\, \ell_U(\hat{\mu})\,\} = -\tfrac{1}{2}(nT - 1)\,,$$

$$E_\mu\{\, \ell_V(\hat{\mu}_1,\hat{\sigma}_1^2,\hat{\nu})\} = -\tfrac{1}{2}\, n \left(T + \mu^2 \sum_{t=2}^{T} u_t^2 - \nu_\mu\, \mu^2 \sum_{t=2}^{T} u_t\, u_{t-1}\, v_{t-1} \right),$$

where

$$\nu_\mu = \frac{\mu^2 \sum_{t=2}^{T} u_t\, u_{t-1}\, v_{t-1}}{\sum_{t=2}^{T} (\mu^2\, u_{t-1}^2 + 1)\, v_{t-1}^2}\,.$$

For standardizing T_U we uses Cox's (1962) formula (17). Hence the

test function for U versus V is defined. The symmetric test T_V is more complicated, but not impossible to compute.

A more informal procedure, applicable more generally than the one just described, is to plot the residual values against time, for each of the two models separately. Residuals for model U are expected to have constant dispersion over time, while for model V dispersion will in general be a function of the v_t's, and hence of time.

Chapter 3

STOCHASTIS MODELS FOR NON-IDENTICALLY DISTRIBUTED MARKOV SERIES

In this chapter, we consider random-effect models to interpret the variability of parameters between series, as opposed to systematic effect models of the previous chapter.

3.1 Empirical Bayes approach for the mean

Similarly to section 2.2, we assume that the following model holds

$$y_{i,t} - \mu_i = \alpha (y_{i,t-1} - \mu_i) + \varepsilon_{i,t} , \qquad (1)$$

where $|\alpha| < 1$, $\{\varepsilon_{i,t}\}$ are white noise $N(0,\sigma^2)$, $y_{i,1}$ is sampled from the stationary distribution $N(\mu_i,\sigma^2/(1-\alpha^2))$. However, here we assume either that the vector $\underline{\mu} = (\mu_1,\ldots,\mu_n)'$ is unrestricted or, alternatively, that $\mu_1,\ldots,\mu_n$ are i.i.d. random variables sampled from some Normal distribution $N(\bar{\mu},\sigma^2_\mu)$. On writing

$$y_{i,t} = \mu_i + \eta_{i,t} , \qquad \eta_{i,t} = \alpha \eta_{i,t-1} + \varepsilon_{i,t} ,$$

with $\eta_{i,1}$ sampled from $N(0,\sigma^2/(1-\alpha^2))$, we see that the scheme is a random-effect one-way classification with autocorrelated errors . This problem has received some attention in the statistical and econometric literature. In the statistical one, it has been considered by Tiao and Tan (1966) from a Bayesian point of view, the main purpose being the computation of the *a posteriori* distribution of σ^2, σ^2_μ, α using vague

information *a priori* distribution. Tiao and Tan stress that, when the unbiased estimate σ_μ^2 for the usual random-effect anlysis of variance is negative, a possible explanation may be negative autocorrelation of the errors not accounted for. In the econometric literature, the interest in the topic seems to begin with a paper of Balestra and Nerlove (1966). However, all of them work with the likelihood conditional on the initial values. Besides a certain lack of efficiency, this approach seems to lead to additional problems; see later.

We shall follow an empirical Bayes approach. By integrating with respect to μ, the marginal p.d.f. of a single series $\underline{y} = (y_1, \dots, y_T)$ is

$$\begin{aligned} f(\underline{y};\bar{\mu},\sigma_\mu^2,\sigma^2,\alpha) &= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-T/2}\,(1-\alpha^2)^{\frac{1}{2}}\,\exp\Big\{-\frac{1}{2\sigma^2}\Big[d_{11} - 2\alpha d_{12} + \alpha^2 d_{22} + \\ &\quad -2\alpha(1-\alpha)(d_1 - \alpha d_2)\mu + (1-\alpha)\big(T-(T-2)\alpha\big)\mu^2\Big]\Big\} \times \\ &\quad \times (2\pi\sigma_\mu^2)^{-\frac{1}{2}}\,\exp\Big(-\frac{1}{2\sigma_\mu^2}(\mu-\bar{\mu})^2\Big)\,d\mu \\ &= (2\pi\sigma^2)^{-T/2}\,(1-\alpha^2)^{\frac{1}{2}}\,\exp\Big[-\frac{1}{2\sigma^2}(d_{11} - 2\alpha d_{12} + \alpha^2 d_{22})\Big] \times \\ &\quad \times (\sigma_\mu^2 B)^{-\frac{1}{2}}\,\exp\Big[\frac{C^2(y)}{2B} - \frac{\bar{\mu}^2}{2\sigma_\mu^2}\Big] \qquad (2) \end{aligned}$$

where

$$B = (1-\alpha)\big(T-(T-2)\alpha\big)/\sigma^2 + 1/\sigma_\mu^2\,,$$

$$C(\underline{y}) = (1-\alpha)(d_1 - \alpha d_2)/\sigma^2 + \bar{\mu}/\sigma_\mu^2\,,$$

$$d_r = \sum_{t=r}^{T-r+1} y_t\,, \qquad d_{rs} = \sum_{t=s}^{T-r+1} y_t\, y_{t+r-s}\,.$$

Some simple algebra gives the *a posteriori* distribution of μ given $\underline{y}$, that is $N(C(\underline{y})/B,\ 1/B)$ assuming $\bar{\mu}$, σ_μ^2, σ^2, α known. Clearly, if n series $\underline{y}_1, \dots, \underline{y}_n$ (possibly of different length) have been observed, then the distribution of μ_n given all n series is $N(C(\underline{y}_n)/B,\ 1/B)$, since the μ_i are independently sampled.

Look for estimates of $\bar{\mu}$, σ_μ^2, σ^2, α. Using (2), the loglikelihood

for a set of n series is readily written down, namely

$$\ell(\bar{\mu},\sigma^2_\mu,\sigma^2,\alpha) = \text{constant} - \tfrac{1}{2} n T \log \sigma^2 + \tfrac{1}{2} n \log(1-\alpha^2) +$$
$$- \frac{1}{2\sigma^2}(d_{11} - 2\alpha d_{12} + \alpha^2 d_{22}) + \qquad (3)$$
$$- \tfrac{1}{2} n \log \sigma^2_\mu - \tfrac{1}{2} n \log B + \frac{1}{2B} \sum_i C(\underline{y}_i)^2 - \frac{n\,\bar{\mu}^2}{2\,\sigma^2_\mu}$$

where

$$d_{rs} = \sum_{i=1}^{n} \sum_{t=s}^{T-r+1} y_{i,t}\, y_{i,t+r-s} \,,$$

restricting ourselves to series of equal length T, for the sake of simplicity. For maximizing ℓ, notice that the first two lines of (3) do not depend on μ, σ^2_μ. Therefore, we maximize with respect to $\bar{\mu}$, σ^2_μ first, for any fixed value of σ^2 and α. For, define

$$d_r^{(i)} = \sum_{t=r}^{T-r+1} y_{i,t} \,, \qquad u_i = \frac{d_1^{(i)} - \alpha\, d_2^{(i)}}{T - (T-2)\,\alpha} \,,$$

$$\bar{u} = \sum_i u_i/n \,, \qquad s_u^2 = \sum_i (u_i - \bar{u})^2/n \,.$$

Equating to zero the derivatives of (3), we find the solution $\bar{u}$ for $\bar{\mu}$, and

$$s_u^2 - \frac{\sigma^2}{(1-\alpha)\{T - (T-2)\alpha\}} \qquad (4)$$

for σ^2_μ, which can be negative. Assuming (4) positive and replacing into ℓ, one has to maximize numerically with respect to σ^2, α to get the MLE. The possibility that (4) takes negative values is disturbing, at least from a technical theoretical point of view, because $\sigma^2_\mu = 0$ is not a point of the parameter space. However, Nerlove (1978, p.15) commenting on the existence of boundary solutions when maximizing the likelihood conditional on the initial values says that "the prevalence of boundary solutions in the maximum likelihood case disappears when the likelihood function is based on an approximation to the unconditional probability (*sic*) of the sample observations rather than on the probability conditional on the initial values of the dependent variable".

Quick estimates of the parameters can be obtained as follows. Note that the distribution of $y_{i,t} - y_{i,t-1}$ does not depend on μ_i. Then

$$E\{(y_{i,t} - y_{i,t-1})^2\} = \frac{2\sigma^2}{1+\alpha} ,$$

$$E\{(y_{i,t} - y_{i,t-1})(y_{i,t-1} - y_{i,t-2})\} = -\sigma^2 \frac{1-\alpha}{1+\alpha} .$$

Replacing the left hand sides by sample moments and solving for σ^2, α, estimates are obtained. Clearly, we estimate $\bar{\mu}$ by $\sum_i \bar{y}_i/n$, where $\bar{y}_i$ is the sample mean of the i-th series. For σ_μ^2 use

$$\mathrm{var}(\bar{y}_i) = \sigma_\mu^2 + \nu ,$$

where

$$\nu = \mathrm{var}(\bar{y}_i | \mu_i) = \frac{\sigma^2}{T(1-\alpha^2)}\left[\frac{1+\alpha}{1-\alpha} - \frac{2\alpha(1-\alpha^T)}{T(1-\alpha)^2}\right] , \qquad (5)$$

which can be estimated using previous estimates of α, σ^2.

Goodrich and Caines (1979) have considered a much more general form of the above model and have given an algorithm, based on the Kalman filter, for evaluating the likelihood, although not in closed form like (3). In a very recent paper, T.W.Anderson and Cheng Hsiao (1981) discuss thoroughly various assumptions on the distribution of the initial values and their effect on the consistency of the MLE and alternative estimates.

3.2 Means with drift

We still assume that the data satisfy equation (1), but we change our hypothesis about the μ_i's. When the subscript i is not arbitrary, but corresponds to some natural ordering, it might well be that the μ_i's are autocorrelated. For instance, consider

$$\mu_n - \bar{\mu} = \rho_\mu (\mu_{n-1} - \bar{\mu}) + \xi_n , \qquad (6)$$

where the usual subscript i has been changed to n, $\{\xi_n\}$ are i.i.d. $N(0,\sigma_\xi^2)$, $|\rho_\mu| < 1$, and the initial value μ_1 is sampled from the stationary distribution $N(\bar{\mu},\sigma_\xi^2/(1-\alpha^2))$. Re-write the observations as

$$\underline{y}_n = \mu_n \underline{1} + \underline{n}_n \tag{7}$$

where $\underline{1}$ is a T-dimensional vector of 1's, and $\{\underline{n}_n\}$ is a sequence of i.i.d. $N_T(\underline{0}, \sigma^2 \Omega)$, $\underline{0}$ being a T-dimensional vector of 0's and Ω a $T \times T$ matrix with (i,j)-th entry $(1-\alpha^2)^{-1} \alpha^{|i-j|}$; $|\alpha| < 1$.

The *a posteriori* distribution of μ_n given $\underline{y}_1,\ldots,\underline{y}_n$ and the predictive one of μ_{n+1} given the same data can be computed recursively using the Kalman filter (Kalman, 1963). They are both Normal ones; if we denote

$$\hat{\mu}_{n|n-1} = E(\mu_n|\underline{y}_1,\ldots,\underline{y}_{n-1}) , \qquad \sigma^2_{n|n-1} = \mathrm{var}(\mu_n|\underline{y}_1,\ldots,\underline{y}_{n-1}) ,$$

and analogously for $\hat{\mu}_{n|n}$ and $\sigma^2_{n|n}$, then it is known that the following recurrence relationships hold

$$\hat{\mu}_{n+1|n} = \bar{\mu}(1-\rho_\mu) + \rho_\mu \hat{\mu}_{n|n-1} + \underline{k}'_n(\underline{y}_n - \hat{\mu}_{n|n-1}\underline{1}) , \tag{8}$$

$$\underline{k}'_n = \rho_\mu \sigma^2_{n|n-1} \underline{1}'(\sigma^2_{n|n-1}\underline{1}\underline{1}' + \sigma^2\Omega)^{-1} ,$$

$$\sigma^2_{n+1|n} = \rho^2_\mu \sigma^2_{n|n-1} \{1 - \sigma^2_{n|n-1} \underline{1}'(\sigma^2_{n|n-1}\underline{1}\underline{1}' + \sigma^2\Omega)^{-1}\underline{1}\} + \sigma^2_\xi ,$$

from which the parameters of the *a posteriori* distribution are easily obtained, namely

$$\hat{\mu}_{n+1|n+1} = (\sigma^{-2}_{n+1|n} + \sigma^{-2}\underline{1}'\Omega^{-1}\underline{1})^{-1} (\sigma^{-2}_{n+1|n}\hat{\mu}_{n+1|n} + \sigma^{-2}\underline{1}'\Omega^{-1}\underline{y}_{n+1}) ,$$

$$\sigma^2_{n+1|n+1} = (\sigma^{-2}_{n+1|n} + \sigma^{-2}\underline{1}'\Omega^{-1}\underline{1})^{-1} .$$

For initializing the procedure, one takes the parameters of the stationary distribution, that is $\hat{\mu}_{1|0} = \bar{\mu}$, $\sigma^2_{1|0} = \sigma^2_\xi/(1-\rho^2_\mu)$.

The same formulae can be used for computing the likelihood, hence for making inference about the parameters. In fact, $\underline{e}_n = (\underline{y}_n - \hat{\mu}_{n|n-1}\underline{1})$ for $n = 1,2,\ldots$ is a sequence of independent Normal random variables with mean $\underline{0}$ and variance matrix $\sigma^2_{n|n-1}\underline{1}\underline{1}' + \sigma^2\Omega$, for which the likelihood

can be computed. Moreover, since the sequence $\{\sigma^2_{n|n-1}\}$ converges rapidly to a stationary value σ^2_*, and so does $\{\underline{k}'_n\}$, then left multiplying (8) by $\underline{1}$ we have

$$\underline{y}_n - \rho_\mu \underline{y}_{n-1} = \bar{\mu}(1-\rho_\mu)\underline{1} + \underline{e}_n + (\underline{1}'\underline{k}_* - \rho_\mu I)\underline{e}_{n-1}$$

where $\{\underline{e}_n\}$ is a sequence of independent and approximately identically distributed $N_T(\underline{0}, \sigma^2_* \underline{1}\underline{1}' + \sigma^2\Omega)$. This device for representing $\{\underline{y}_n\}$ as a multivariate ARMA(1,1) process is commonly used among control engineers.

Quick estimates of the parameters can be obtained as follows. The technique of the previous section for estimating σ^2, α and hence ν, defined by (5), still holds. Defining $\bar{y}_n = T^{-1}\underline{1}'\underline{y}_n$, we have

$$\operatorname{cov}(\bar{y}_n, \bar{y}_{n-1}) = \rho_\mu \frac{\sigma^2_\xi}{1-\rho^2_\mu}, \qquad \operatorname{var}(\bar{y}_n) = \frac{\sigma^2_\xi}{1-\rho^2_\mu} + \nu ;$$

replacing the left hand side by sample moments, we get estimates of ρ_μ, σ^2_ξ.

If one is not interested in α or σ^2 separately, but only in ν, then it is possible to consider the process $\{\underline{y}_n\}$ which is exactly a univariate ARMA(1,1) process with parameters related to ρ_μ, σ^2_ξ, ν by functions given in Chapter 5.

There are various possible generalizations of model (6) - (7): (a) the process $\{\mu_n\}$ does not need to be stationary, provided information about the initial distribution is available; (b) the length of the series may vary with n; (c)instead of $\mu_n \underline{1}$, the mean vector of the n-th series may be $X_n \underline{\beta}_n$, where X_n is a $T \times p$ matrix of covariates associated with the n-th series, and $\underline{\beta}_n$ is a p-dimensional linear Markov process; (d) $\sigma^2\Omega$ may be any symmetric positive definite matrix, possibly parameterized in a few parameters to facilitate inference. As all these generalizations are included in the general Kalman filter and inference in possible via the sequence $\{\underline{e}_n\}$, we shall not give details.

3.3 Empirical Bayes approach for the autoregressive parameter

We suppose that the n series are stationary and identically distributed except for the autoregressive parameter, i.e.

$$y_{i,t} - \mu = \alpha_i (y_{i,t-1} - \mu) + \varepsilon_{i,t} \qquad (9)$$

where, as usual, $\{\varepsilon_{i,t}\}$ are i.i.d. $N(0,\sigma^2)$ and the initial values are sampled from the stationary distribution $N(\mu,\sigma^2/(1-\alpha_i^2))$. By contrast with section 2.4, we regard the α_i's as a sample from some *a priori* distribution in $(-1, 1)$, say $\pi(\alpha)$. Nicholls and Quinn (1980) have studied a model (for one series) where α varies with time; here α is constant over time and varies across series.

This model has been considered by Robinson (1978), who has studied the asymptotic properties of the sample covariances and the associated estimates of σ^2 and of the moments of $\pi(\alpha)$.

For this reason, we shall concentrate on the evaluation of the *a posteriori* distribution of α given $\pi(\alpha)$ and a series $\underline{y} = (y_1,\ldots,y_T)'$, having dropped the suffix i. We assume that $\pi(\alpha)$ is a Beta(p,q) density in $(-1, 1)$, i.e.

$$\pi(\alpha) = \pi(\alpha;p,q) = \frac{(1+\alpha)^{p-1}\,(1-\alpha)^{q-1}}{B(p,q)\;2^{p+q-1}} \qquad (-1 < \alpha < 1)\ . \qquad (10)$$

The marginal density function of $\underline{y}$ is

$$f(\underline{y}) = \int_{-1}^{1} (2\pi\sigma^2)^{-T/2} \exp\{-\tfrac{1}{2}(d_{11} - 2\alpha d_{12} + \alpha^2 d_{22})\}\ \pi(\alpha;p-\tfrac{1}{2},q-\tfrac{1}{2})\ d\alpha$$

$$= \frac{\exp(-\frac{1}{2}d_{11})\ \ I}{(2\pi\sigma^2)^{T/2}\ \ B(p,q)\ \ 2^{p+q-1}}\ ,$$

where

$$I = \int_{-1}^{1} \exp\{-\tfrac{1}{2}(2\alpha d_{12} + \alpha^2 d_{22})\}\ (1+\alpha)^{p-\frac{1}{2}}\ (1-\alpha)^{q-\frac{1}{2}}\ d\alpha\ ,$$

$$d_{rs} = \sum_{t=s}^{T-r+1} (y_t - \mu)(y_{t+r-s} - \mu)/\sigma^2\ ,$$

regarding μ, σ^2 as known since the argument is for large n.

The integral I cannot be solved explicitly for the general case. Consider for the moment the special case T = 2, for which $d_{22} \equiv 0$. Using the equality

$$\int_{-1}^{1} \exp(\alpha z)\ \pi(\alpha;p,q)\ d\alpha = e^{-z}\ M(p,p+q,2z)$$

where M is Kummer's function as defined by Abramowitz and Stegun (1964, formula 13.1.2, see also 13.2.1), we get

$$f(y_1,y_2) = \frac{B(p+\frac{1}{2},q+\frac{1}{2})}{B(p,q)\ \pi\ \sigma^2} \exp\{-\tfrac{1}{2}(d_{11} + 2\,d_{12})\}\ M(p,p+q,2\,d_{12})\ .$$

Hence the *a posteriori* distribution of α is

$$\pi_c(\alpha|y_1,y_2) = \frac{\exp\{(1+\alpha)d_{12}\}\ \ \pi(\alpha;p+\frac{1}{2},q+\frac{1}{2})}{M(p+\frac{1}{2},p+q+1,\,2\,d_{12})}\ ,$$

from which

$$E(1+\alpha|y_1,y_2) = \frac{M(p+3/2,\ p+q+2,\,2\,d_{12})\ \ (p+\frac{1}{2})}{M(p+\frac{1}{2},\ p+q+1,\,2\,d_{12})\ \ (p+q+1)}\ ,$$

$$E(1-\alpha^2|y_1,y_2) = \frac{M(p+3/2,\ p+q+3,\,2\,d_{12})(p+\frac{1}{2})\ \ (q+\frac{1}{2})}{M(p+\frac{1}{2},\ p+q+1,\,2\,d_{12})\ \ (p+q+1)\ \ (p+q+2)}\ .$$

For T > 2, we find an approximate value of I for $T \to \infty$, breaking the rule of keeping T fixed we have followed so far. For, write

$$I = \exp(\tfrac{1}{2}\,d\,T\,a^2) \int_{-1}^{1} \exp\{-\tfrac{1}{2}\,d\,T(\alpha-a)^2\}\ \ (1+\alpha)^{p-\frac{1}{2}}\ \ (1-\alpha)^{q-\frac{1}{2}}\ d\alpha$$

$$= \exp(\tfrac{1}{2}\,d\,T\,a^2) \int_{-1-a}^{1-a} \exp\{-\tfrac{1}{2}\,d\,T\ \theta^2\}\ \ (1+\theta+a)^{p-\frac{1}{2}}\ \ (1-\theta-a)^{q-\frac{1}{2}}\ d\theta$$

where

$$d = d_{22}/T\ , \qquad a = d_{12}/d_{22}\ ,$$

which are of order 1 in probability as $T \to \infty$. For sufficiently large T, (-1-a, 1-a) is a neighbourhood of 0. Expanding

$$g(\theta) = (1+\theta+a)^{p-\frac{1}{2}}\ \ (1-\theta-a)^{q-\frac{1}{2}}$$

in Taylor series from 0 and using the Laplace method for integrals, we

have

$$I \doteq \exp(\tfrac{1}{2} d\,T a^2) \left(\frac{2\pi}{d\,T}\right)^{\frac{1}{2}} g(0) \left(1 + \frac{g''(0)}{2 d\,T\, g(0)}\right),$$

from which

$$\pi_c(\alpha|\underline{y}) \doteq \left(\frac{d\,T}{2\pi}\right)^{\frac{1}{2}} \exp\{-\tfrac{1}{2} d\,T(\alpha - a)^2\} \left(\frac{1+\alpha}{1+a}\right)^{p-\frac{1}{2}} \left(\frac{1-\alpha}{1-a}\right)^{q-\frac{1}{2}} \left(1 - \frac{g''(0)}{2 d\,T\, g(0)}\right)$$

where the leading term is a N(a, 1/(dT)) density. A further asymptotic argument gives

$$E(\alpha|\underline{y}) \doteq a + \frac{1}{d\,T}\left(\frac{p-\frac{1}{2}}{1+a} - \frac{q-\frac{1}{2}}{1-a}\right),$$

$$\mathrm{var}(\alpha|\underline{y}) \doteq 1/(dT) .$$

Obviously one can look at $\pi(\alpha)$ and $\pi_c(\alpha|\underline{y})$ from a Bayesian point of view, regarding $\underline{y}$ as the only series available. The choice $p = q = \frac{1}{2}$ takes then special relevance, because it corresponds to Jeffreys uninformative prior as $T \to \infty$. In this case, for $d_{22} > 0$,

$$I = \exp(\tfrac{1}{2} a\, d_{22}) \left(\frac{2\pi\sigma^2}{d_{22}}\right)^{\frac{1}{2}} \left(\Phi\{(1-a)/d_{22}^{\frac{1}{2}}\} - \Phi\{-(1+a)/d_{22}^{\frac{1}{2}}\}\right)$$

where Φ is the standard Normal integral. Therefore $\pi_c(\alpha|\underline{y})$ is the $N(a, d_{22}^{-1})$ density truncated outside (-1, 1). Also Thornber (1967) has considered $\pi(\alpha; \frac{1}{2}, \frac{1}{2})$, but regarding σ unknown with improper prior density σ^{-1}; the corresponding posterior density of α is then a truncated Student's t . The associated Bayes estimate seems, from Thornber's Monte Carlo study, to have appreciably smaller mean square error than the exact MLE when $\frac{1}{2} < \alpha < 1$, and slightly larger mean square error when $0 < \alpha < \frac{1}{2}$.

After this thesis had already been completed, the author got to know the paper of Liu and Tiao (1980), who also consider model (9) with α_i's sampled from distribution (10). They shown that the exact posterior density $\pi_c(\alpha|\underline{y})$ has a unique mode inside (-1, 1) for p, $q > \frac{1}{2}$, and derive an approximation to the posterior distribution different from

our, although equivalent as $T \to \infty$. Furthermore, after changing parameterization from (p,q) to alternative parameters (u,v), they regard (u,v) as "secondary parameters", for which uninformative prior distributions are introduced, and an approximation to the corresponding posterior distribution of the α_i's is derived.

Chapter 4

APPROXIMATE FILTERING OF PARAMETER DRIVEN PROCESSES

4.1 Parameter driven processes

Much time series has concentrated on Normal processes or, at least, on processes with Normal increments or on associated second order theory. Considerably less attention has been paid to processes which do not have Normal increments. The present chapter fits into the second class of problems and, in particular, is concerned with filtering and prediction for some of these processes.

Let $\{\theta_t\}$ be a discrete-time stationary Markov process which is of interest but cannot be observed directly. Instead, at time t, we observe y_t which is sampled from a distribution determined by the current value of θ_t, say $f(y_t|\theta_t)$, and which is conditionally independent of its predecessors and the past values of $\{\theta_t\}$. We shall say that $\{y_t\}$ is a parameter driven process in the terminology of Cox (1981). Note that, in general, $\{y_t\}$ is not a Markov process. Our aim is to produce, for all values of t, an estimate of the current value θ_t given $y^{(t)}$, where $y^{(t)} = (y_1,\ldots,y_t)$. The problem of estimating the next value θ_{t+1} given $y^{(t)}$ is closely related. In a different context, $\{\theta_t, y_t\}$ would be called a partially observable process and the term state estimation possibly used instead of filtering.

An example of the above set-up is as follows. Batches of items are received at each time unit. We are interested in evaluating the proportion of defective items of each batch, θ_t say. From each batch we

take a sample of size N and denote by y_t the number of defective items of the sample. If the size of the batches is sufficiently large, y_t is approximately distributed as a Binomial r.v. with index N and probability of "success" θ_t. We can obviously estimate θ_t by y_t/N, or $(y_t + p)/(N + p + q)$ if θ_t has a Beta(p,q) *a priori* distribution, but we can expect that a better estimate is obtained if an allowance is made for the autocorrelation of the θ_t's and also past values $y^{(t-1)}$ are used.

This kind of problem has been studied by Bather (1965) who, among other results, has given conditions for the existence of a one-dimensioanl sufficient statistic $u_t = u(u_{t-1}, y_t)$ such that the distribution of θ_t given $y^{(t)}$ depends on $y^{(t)}$ only via u_t. For further work, see Sawitzki (1979, 1981).

Unfortunately, there are many non-pathological problems which do not admit such a simple sufficient statistic u_t. We shall deal with some of these problems. In section 2 and 4, an approximate method of filtering is considered which is applicable when the marginal distribution of θ_t and $f(y_t|\theta_t)$ are conjugate exponential densities. Section 5 deals with the estimation of the fixed parameters, that is parameters which are constant throughout time.

4.2 Gamma-Poisson process

Although the method we are going to describe has some generality, we introduce it by an example. We say that a r.v. has a Gamma(ν,λ) distribution for ν, $\lambda > 0$ if it continuous with density function at x

$$\pi(x;\nu,\lambda) = \begin{cases} \lambda^{\nu} x^{\nu} e^{-\lambda x}/\Gamma(\nu) & (x > 0) , \\ 0 & (x \leq 0) . \end{cases} \tag{1}$$

Consider the autoregressive process $\{\theta_t\}$ satisfying

$$\theta_{t+1} = \rho\, \theta_t + \varepsilon_{t+1} \qquad (0 < \rho < 1,\ t = 0, \pm 1, \ldots), \tag{2}$$

where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed r.v.'s such that $\{\theta_t\}$ is stationary with Gamma(ν,λ) marginal distribution. Gaver and Lewis (1980) have shown that, for integer ν, ε_t is a mixture of Gamma$(1,\lambda),\ldots,$Gamma(ν,λ) r.v.'s and a r.v. degenerate at 0.

At time t, we observe y_t which is sampled from a Poisson distribution with mean θ_t. For all t, we want to compute the conditional densities of θ_t and θ_{t+1} given $y^{(t)}$, say $\omega_t(\cdot|y^{(t)})$ and $\omega_t'(\cdot|y^{(t)})$.

First of all, we note that no reduction to one-dimensional sufficient statistic is possible. In fact, Sawitzki (1981) has shown, generalizing a result of Bather (1965), that for such a reduction both ω_t and ω_t' should be exponential family densities with ω_t' of dimension 1. This cannot hold because, even if ε_t satisfied the condition, ω_t' would be a mixture of densities, due to the nature of ε_{t+1} in (2).

We look for a method of approximating ω_t which allows simple, recursive computation. Suppose that the problem has been solved up to time t, for which $\omega_t(\cdot|y^{(t)})$ is Gamma(ν_t,λ_t) and that we want to compute $\omega_{t+1}(\cdot|y^{(t+1)})$.

The first step is to compute $\omega_t'(\cdot|y^{(t)})$. The exact conditional density is very complicated and needs numerical integration. However, from

$$m_t \equiv E(\theta_t|y^{(t)}) = \nu_t/\lambda_t , \qquad v_t \equiv \mathrm{var}(\theta_t|y^{(t)}) = \nu_t/\lambda_t^2 ,$$

we obtain

$$m'_t \equiv E(\theta_{t+1}|y^{(t)}) = \rho m_t + E(\varepsilon),$$

$$v'_t \equiv \text{var}\,(\theta_{t+1}|y^{(t)}) = \rho^2 v_t + \text{var}\,(\varepsilon)$$

where

$$E(\varepsilon) = (1-\rho)\,\nu/\lambda, \qquad \text{var}\,(\varepsilon) = (1-\rho^2)\,\nu/\lambda^2.$$

We approximate $\omega'_t(.|y^{(t)})$ by a Gamma distribution with mean m'_t and variance v'_t. In order words, we say that $\omega'_t\,(.|y^{(t)})$ is Gamma (ν'_t, λ'_t), where

$$\nu'_t = (m'_t)^2/v'_t, \qquad \lambda'_t = m'_t/v'_t. \tag{3}$$

Now make use of y_{t+1}. Since the Gamma and the Poisson are conjugate exponential families and since the y_t's are conditionally independent, we have at once that $\omega_{t+1}(.|y^{(t+1)})$ is Gamma $(\nu_{t+1},\lambda_{t+1})$, where

$$\nu_{t+1} = \nu'_t + y_{t+1}, \qquad \lambda_{t+1} = \lambda'_t + 1, \tag{4}$$

which completes the updating cycle. The procedure can be started with $\nu_o = \nu$, $\lambda_o = \lambda$.

For studying the behaviour of the filter, attention has been concentrated on the loss function $E(\theta_t - \hat{\theta}_t)^2$, for which $\hat{\theta}_t = m_t$ is the obvious choice. This loss function may not be the most appropriate one in some cases, but alternative choices depend on the particular application.

An algorithm for computing $E(\theta_t - m_t)^2$ in the steady state has been devised but its implementation would have been too complicated and it

Table 4.1: Relative loss of the filter for the Gamma-Poisson case

ν	1	1	1	2	2	5	5
λ	2	1	1	1	1	1	1
ρ	0.8	0.5	0.8	0.5	0.8	0.5	0.8
$L(\nu,\lambda)$, M = 2	0.69	0.80	0.64	0.92	0.72	0.90	0.74
$L(\nu,\lambda)$, M = 4	2.37	0.79	0.63	0.92	0.71	0.90	0.72
$L_e(\nu,\lambda)$, M = 2	0.74	0.82	0.66	0.95	0.74	0.92	0.76

Table 4.2: Relative loss of the filter for the Beta-Binomial case

p	2	2	2	2	5	5
q	10	10	5	5	5	5
α	0.5	0.8	0.8	0.95	0.8	0.95
L(p,q)	0.83	0.67	0.79	0.36	0.72	0.49
$L_e(p,q)$	0.89	0.77	0.81	0.50	0.75	0.58

has not been pursued. Alternatively, a simulation study has been carried out, based on runs of length 500 of the process. The results of some of these simulations are summarized in Table 4.1; the first three rows are parameter levels, the fourth row gives the sample value of the relative loss

$$L(\nu,\lambda) = \frac{E(\theta_t - m_t)^2}{\nu/\{\lambda(\lambda + 1)\}}$$

where the denominator is $E(\theta_t - m_t^*)^2$ for $m_t^* = (y_t + \nu)/(1 + \lambda)$, i.e. the Bayes estimate which accounts only for the marginal distribution of θ_t. Note that $m_t \to m_t^*$ as $\rho \to 0$. The results consistently indicate that the method works satisfactorily; in no case (including those not reported here), did it exhibit unstable behaviour. The last two rows of Table 4.1 and the meaning od "M=2" will be explained later.

4.3. Higher order approximation

The crucial point of the method is the approximation of $\omega_t'(\cdot|y^{(t)})$ by a member of a two-parameter family. In this section we consider a M-parameter approximation. To this end, let us summarize some results of Khamis (1960).

Denote by $L_r(\cdot)$, $(r = 0,1,\ldots)$, the set of Laguerre polynomials which are orthogonal with respect to $\pi(\cdot;\nu',\lambda')$ defined by (1). For a density function $p(\cdot)$ with known M moments and support on the positive half-line, define the approximation

$$p_M(x) = \sum_{r=0}^{M} c_r L_r(x) \pi(x;\nu',\lambda')$$

where c_r $(r = 0,\ldots,M)$ are such that p and p_M have the same first M moments $(M > 1)$; p_M may not be a proper density function (we shall

return to this point). It turns out that

$$p_M(x) = \sum_{r=0}^{M} B_r \, \pi(x;\nu'+r,\lambda') \tag{5}$$

where B_r $(r = 0,\ldots,M)$ are coefficients defined by Khamis (1960).

The form (5) has a useful feature. If Y is a Poisson r.v. whose means is *a priori* distributed like p_M, then we can easily compute the "a posteriori" density, which is

$$p_M(x|Y=y) = \sum_{r=0}^{M} \frac{B_r}{\Delta} \frac{\Gamma(\nu' + r + y)}{\Gamma(\nu' + r)} \left(\frac{\lambda'}{\lambda'+1}\right)^r \pi(x;\, \nu'+r+y,\lambda'+1) \tag{6}$$

where Δ is a norming factor.

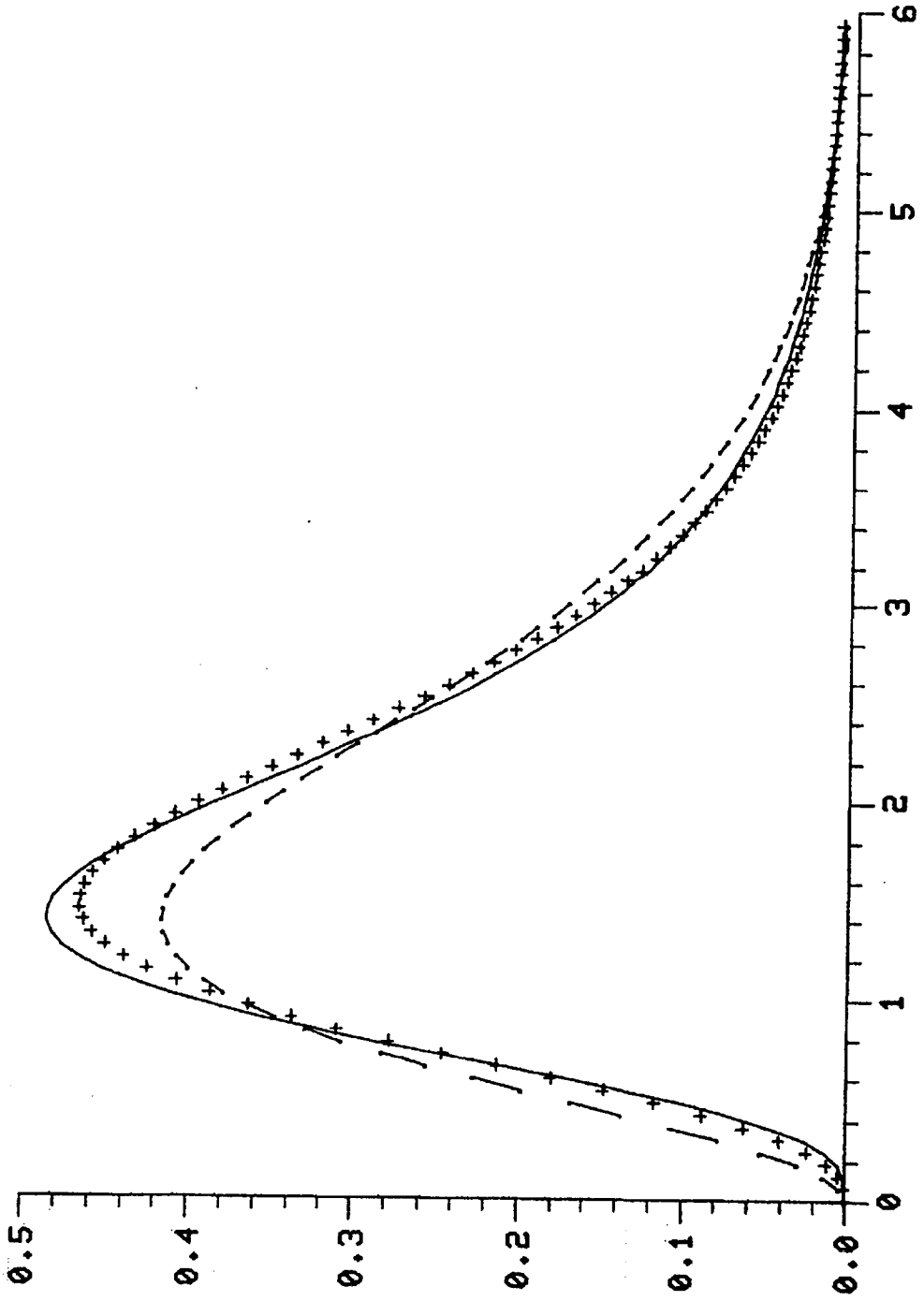
A generalization of the procedure of section 2 is therefore as follows: (i) given M moments of $\omega_t(\cdot|y^{(t)})$, compute the first M moments of $\omega_t'(\cdot|y^{(t)})$; (ii) approximate ω_t' by a distribution of form (5) a possible choice of ν',λ' is via (3); (iii) compute $\omega_{t+1}(\cdot|y^{(t+1)})$ using (6), hence its first M moments. Note that this procedure reduces to the previous one when M = 2.

To examine the improvement of fitting of ω_t' using M > 2, the following computations have been carried out. Suppose that $\omega_t(\cdot|y^{(t)})$ is Gamma (ν_t,λ_t) and the marginal distribution of θ_t is Gamma (2,1). Using numerical integration, we can compute $\omega_t'(\cdot|y^{(t)})$ exactly and compare it with its approximations. Various plots have been produced and one is reported in Figure 4.1, for which $\rho = 0.8$, $\nu_t = 5$, $\lambda_t = 2.5$; this choice of ν_t,λ_t corresponds approximately to the stationary distribution of ω_t for the above parameter levels. The improvement of M = 4 over M = 2 is appreciable.

The filter with M = 4 was applied to the same data used for the simulations described before. The results are given in the fifth row

Figure 4.1: Plot of $\omega_t'(\cdot|y^{(t)})$ and approximations when $\omega_t(\cdot|y^{(t)})$ is Gamma(5,2.5), $\nu = 2$, $\lambda = 1$, $\rho = 0.8$;

—— exact, ---- approximate, M = 2, ++++ approximate, M = 4.



of Table 4.1. There is a minor reduction of the loss function with respect to M = 2, except for the aberrant value of the first column which requires an explanation. As mentioned before, (5) may take negative values. This does not matter in most cases because negative values are taken only in the extreme tail and the absolute value is very small anyway. But, in one case, when ω_t' was all concentrated in the interval (0,4), $y_{t+1} = 9$ was observed, giving all weight to the extreme tail of the approximate ω_t', where it took values about -10^{-7}. This caused the filter to produce absurd values for two time units, which spoiled irreparably the estimate of $E(\theta_t - \hat{\theta}_t)^2$.

These simulations suggest that the use od M > 2 can improve only marginally on M = 2, at least for quadratic loss, and it can occasionally produce nonsense. Therefore, in the following we restrict ourselves to the method of section 2, which is computationally much simpler. Of course, we are not saying that another technique (in particular the exact filtering solution) cannot be appreciably better than the one of section 2.

4.4. Other examples

The method of section 2 can be easily extended to other cases provided: (i) $\{\theta_t\}$ is a stationary Markov process with marginal distribution specified by K moments; (ii) we can compute, or approximate, $E(\theta_{t+1}^r | y^{(t)})$ given $E(\theta_t^r | y^{(t)})$, r = 1,...,K; (iii) the marginal distribution of $\{\theta_t\}$ and $f(y_t|\theta_t)$ are conjugate. It is worth enumerating the more easily workable cases.

Normal-Normal process. Suppose that $\{\theta_t\}$ satisfies (2) but $\{\varepsilon_t\}$ are independent and identically distributed $N(0,\sigma_\varepsilon^2)$ and $f(y|\theta)$ is $N(\theta,\sigma_\eta^2)$. In this case the formulae of the Kalman filter are obtained, without any approximation.

Gamma-Gamma process. The $\{\theta_t\}$ process is the same as in section 2 but $f(y|\theta)$ is Gamma (p,θ). The only change for the filter is that (4) is replaced by $\nu_{t+1} = \nu_t' + p$, $\lambda_{t+1} = \lambda_t' + y_{t+1}$.

Beta-Binomial process. Suppose that $\{\theta_t\}$ has a Beta(p,q) marginal distribution, i.e. with density proportional to $x^p(1-x)^{q-1}$ for $x \in (0,1)$. For such a process, see Appendix B. Assume $f(y|\theta)$ is Binomial(N,θ). The argument for deriving the filter is similar to the Gamma-Poisson case. Suppose that $\omega_t(\cdot|y^{(t)})$ is Beta (p_t, q_t). Then

$$m_t \equiv E(\theta_t|y^{(t)}) = p_t/(p_t + q_t),$$

$$v_t \equiv \text{var}\ (\theta_t|y^{(t)}) = m_t(1 - m_t)/(p_t + q_t + 1).$$

Using (B.1) of Appendix B,

$$m_t' \equiv E\ (\theta_{t+1}|y^{(t)}) = E(\theta_t^{\alpha}\ \varepsilon_t|y^{(t)}) = \frac{B(p_t + \alpha,\ q_t)}{B(p_t,\ q_t)} E(\varepsilon)\ ,$$

$$v_t' \equiv \text{var}\ (\theta_{t+1}|y^{(t)}) = \frac{B(p_t + 2\alpha,\ q_t)}{B(p_t,\ q_t)} E(\varepsilon^2) - (m_t')^2\ ,$$

where $E\ (\varepsilon^k)$ is given by (B.3). Now choose the element of the Beta family with mean m_t' and variance v_t', that is with parameters

$$p_t' = m_t'\left(\frac{m_t'\ (1 - m_t')}{v_t'} - 1\right)\ , \qquad q_t' = \frac{(1 - m_t')}{m_t'}\ p_t'\ ,$$

and approximate $\omega_t'\ (\cdot|\ y^{(t)})$ by the Beta (p_t', q_t') density. After observing y_{t+1}, the posterior distribution $\omega_{t+1}(\cdot|y^{(t+1)})$ is

Beta(p_{t+1}, q_{t+1}), where

$$p_{t+1} = p'_t + y_{t+1}, \qquad q_{t+1} = q'_t + N - y_{t+1}, \tag{7}$$

which completes the updating cycle. The procedure can be started with $p_o = p$, $q_o = q$. Note that is not essential that N is constant over time.

Some simulations were run for this filter and the results are summarized in Table 4.2. The first three rows are parameter levels, in all cases being N = 10 and the length of the realization 500; the fourth row gives estimates of the relative loss

$$L(p,q) = \frac{E\,(\theta_t - m_t)^2}{pq/\{(p+q)(p+q+1)(p+q+N)\}}$$

where the denominator is the expected loss of the standard Bayes estimate $m_t^* = (p+y_t)/(p+q+N)$. Again $m_t \to m_t^*$ as $\alpha \to 0$. The behaviour of the filter appears satisfactory, leading to a gain with respect to the standard Bayes estimate comparable to that in the Gamma-Poisson case.

Beta-Negative Binomial process. Let $\{\theta_t\}$ be as in the previous case and $f(y|\theta) \propto \theta^N (1-\theta)^y$, for y = 0,1,... and $N \geq 1$. The formulae of the filter are the same of the previous case except that (7) is replaced by $p_{t+1} = p'_t + N$, $q_{t+1} = q'_t + y_{t+1}$.

4.5. Estimating the fixed parameters

Here we deal with the estimation of the parameters which are constant throughout time, such as ν,λ,ρ for the Gamma-Poisson process. Since the method of the maximum likelihood is not feasible, we consider the method of moments. Assume that there exists a function g such that

$$E\{g(y_t)|\theta_t\} = \theta_t \quad .$$

Then

$$E\{g(y_t)\} = E(\theta_t),$$

$$\text{var}\{g(y_t)\} = \text{var}(\theta_t) + E\{\text{var}(g(y_t)|\theta_t)\},$$

$$\text{cov}\{g(y_t), g(y_{t+h})\} = \text{cov}(\theta_t, \theta_{t+h}), \qquad (h \neq 0),$$

from which estimates of the fixed parameters can be obtained.

For most of the examples mentioned above, it is trivial to find such a function g, except perhaps the Beta-Negative Binomial case for which $g(y) = (N-1)/(N-1+y)$ provided $N > 1$. For the Gamma-Gamma process, $g(y) = (p-1)/y$ if p is known and greater than 1. If p is unknown, an additional statistic must be considered, for instance y_t, whose marginal distribution is a Pearson type VI distribution, or the third moment of g(y). If p is known to be 1 (which corresponds to the important case of exponential distribution for $f(y|\theta)$), then it appears that ν,λ,ρ cannot be estimated with the method of moments; however, the problem can be tackled with the method described in the following chapter.

It is not difficult to prove that, if $\{\theta_t\}$ is strongly mixing, then so is $\{y_t\}$. For, let $A \in \mathcal{B}_{-\infty}^{t}$, $B \in \mathcal{B}_{t+\tau}^{\infty}$, where $\mathcal{B}_{\alpha}^{\beta}$ is the σ-algebra generated by the $\{y_t\}$ process from $t=\alpha$ to $t=\beta$; analogously, $\mathcal{D}_{\alpha}^{\beta}$ is the σ-algebra generated by the $\{\theta_t\}$ process. Then

$$\begin{aligned} |P\{A\cap B\} - P\{A\}\,P\{B\}| &= |\,E(P\{A\cap B|\mathcal{D}_{-\infty}^{t}, \mathcal{D}_{t+\tau}^{\infty}\}) - P\{A\}\,P\{B\}| \\ &= |\,E(P\{A|\,\mathcal{D}_{-\infty}^{t}\}\;\;P\{B|\mathcal{D}_{t+\tau}^{\infty}\}) - P\{A\}\,P\{B\}| \\ &= |\,E(f_A\,f_B) - E(f_A)\;\;E(f_B)| \end{aligned}$$

where $f_A = P\{A|\mathcal{D}_{-\infty}^{t}\}$, similarly for f_B. Now apply theorem 17.2.1 of Ibragimov and Linnik (1971, p.306) which says that the above quantity is less than or equal to $4\,\phi(\tau)$, where $\phi(\tau)$ is the mixing coefficient

of $\{\theta_t\}$. This ends the proof.

An analogous argument applies for weak mixing. Therefore, almost sure convergence of the sample moments is ensured under very broad conditions. Confidence intervals could be computed in principle, but the formulae are too complicated to be useful and are not given.

The filters for the Gamma-Poisson and the Beta-Binomial process were run using estimates of the fixed parameters. The first 50 data of each run were not used for filtering but only for fixed parameter estimation. After t = 50, simultaneous estimation and filtering were applied; if the estimates of the fixed parameters were not in the admissible range, they were forced to be. The results relative to the same data used before are given in the last row of Table 4.1 and 4.2, respectively; it appears that there is only a minor increase of the loss function with respect to the case of known parameters.

As a final remark, a deeper understanding of the behaviour of this method, based on analytic results rather than synthetic sampling, would be desirable. However, in view of its potential practical usefulness, the method may be of interest in its present state.

Chapter 5

m-ORDER MAXIMUM LIKELIHOOD ESTIMATION FOR STATIONARY STOCHASTIC PROCESSES

5.1 Introduction

The MLE is consistent and efficient in a broad class of situations. Unfortunately, there are cases where it is not feasible, because the likelihood cannot be written down, or computed, even if we are ready to ignore differences of order $o_p(T^{-1})$ between the exact and the computed likelihood, T being the sample size. This occurs with some frequency in inference from stochastic processes; an example is mentioned in the previous chapter.

A possible way to overcome the problem is to replace the likelihood by another function, easier to determine. Any function $\ell(\theta)$ which asymptotically has its maximum at the true parameter point, say θ_o, is a potential candidate. However, heuristically the closer $\ell(\theta)$ is to the exact likelihood, at least near θ_o, the better it is, because one expects that the relative efficiency of the associated estimators will be determined by the curvature of $\ell(\theta)$ near θ_o.

For discrete-time stochastic processes, a possible choice of $\ell(\theta)$ is motivated by the fact that the exact loglikelihood can be written as a sum whose t-th term is the logarithm of the p.d.f. of the corresponding sample element conditional on all previous observations. D.R. Cox and P.A.W. Lewis have suggested replacing the conditioning on all previous observations by only the m most recent ones, for some $m \geq 0$.

For this choice of $\ell(\theta)$, we shall study the asymptotic properties

of $\hat{\theta}_m$, a value which maximizes $\ell(\theta)$, when sampling from a stationary stochastic process. Section 2 gives some general results about strong consistency and asymptotic normality. Section 3 and 4 compare $\hat{\theta}_m$ with other estimators for special cases.

5.2 Consistency and asymptotic normality

Let $\{Y_t\}$ be a stationary stochastic process whose h-coordinates distribution is indexed by θ_o in each of the families $\mathcal{P}^{(h)} = \{P_\theta^{(h)}, \theta \in \Theta\}$ for $h = 1,2,\ldots$; assume $\mathcal{P}^{(h)} << \mu^{(h)}$, for some appropriate measure $\mu^{(h)}$, and that Θ is an interval in the k-dimensional Euclidean space. For a sample $(y_{-m+1},\ldots,y_T)$ of size $m+T$ from $\{Y_t\}$ define the m-order log-likelihood

$$\ell(\theta) = \sum_{t=1}^{T} \ell_t(\theta) = \sum_{t=1}^{T} \log f(y_t | y_m^{(t-1)}; \theta) \qquad (1)$$

where $y_m^{(t-1)} = (y_{t-m},\ldots,y_{t-1})$ and $f(\cdot|y_m^{(t-1)})$ is the p.d.f. of Y_t given $Y_m^{(t-1)} = y_m^{(t-1)}$ with respect to some measure which does not depend on θ; if $m = 0$, the marginal p.d.f. of Y_t has to be intended. Define $\hat{\theta}_m$ as a value which maximizes (1).

To prove consistency, it is easy to see that, for $\phi, \theta \in \Theta$,

$$E\{\ell_t(\theta)\} \geq E\{\ell_t(\phi)\} \qquad (2)$$

where the expectations are taken with respect to $P_\theta^{(m+1)}$; the equality sign hold if and only if $\mu^{(m+1)}\{\ell_t(\theta) \neq \ell_t(\phi)\} = 0$. An inequality of type (2) is the key fact for the proof of strong consistency of the ordinary MLE in the i.i.d. case. A perusal of the proof of theorem 5.3.1 of Zacks (1971, p.233-5) reveals that it holds also in the present case, provided

(a) conditions (i) to (v) of Zacks's theorem apply; μ is $\mu^{(m+1)}$ in our

notation;

(b) $\{Y_t\}$ is ergodic.

Under these conditions, $\hat{\theta}_m \to \theta_o$ almost surely (a.s.) as $T \to \infty$. We stress the importance of Zacks's condition (i) which ensures identifiability. This is typically achieved choosing m sufficiently large. Strictly speaking, what one needs is stationarity and ergodicity of $\{\ell_t(\theta)\}$ rather than of $\{Y_t\}$.

Consider now the asymptotic distribution of $\hat{\theta}_m$. Define

$$u(\theta) = \sum_{t=1}^{T} u_t(\theta) = \sum_{t=1}^{T} \frac{d}{d\theta} \ell_t(\theta) .$$

Provided the absolute value of the first and second derivatives of f have finite integral a.s., we have as usual $E\{u_t(\theta)\} = 0$ and

$$E\{u_t(\theta)\, u_t'(\theta)\} = -E\left\{\frac{d}{d\theta'} u_t(\theta)\right\} = V(\theta) , \quad \text{say,}$$

provided the expectations exist. By contrast with the usual i.i.d. case, $V(\theta)$ is different from the variance matrix of $T^{-\frac{1}{2}} u(\theta)$. In fact

$$T^{-1} E\{u(\theta)\, u'(\theta)\} = E\{u_1(\theta)\, u_1'(\theta)\} + \sum_{s=1}^{T-1} \left(1 - \frac{s}{T}\right) E\{u_1(\theta)\, u_{1+s}'(\theta) + u_{s+1}(\theta)\, u_1'(\theta)\}$$

which converges to

$$W(\theta) = E\{u_1(\theta)\, u_1'(\theta)\} + \sum_{s=1}^{\infty} E\{u_1(\theta)\, u_{s+1}'(\theta) + u_{s+1}(\theta)\, u_1'(\theta)\}$$

as $T \to \infty$, provided the series on the right hand side converges.

The argument used in proving theorem 5.5.3 of Zacks (1971, p.246-7) applies, provided his conditions (i) - (iv) hold and $u(\theta_o)$ is asymptotically Normal; $I(\theta_o)$ is $V(\theta_o)$ here. Then

$$T^{\frac{1}{2}} (\hat{\theta}_m - \theta_o) \xrightarrow{d} N_k(0,\, V(\theta_o)^{-1}\, W(\theta_o)\, V(\theta_o)^{-1}) \tag{3}$$

which simplifies to

$$T^{\frac{1}{2}} (\hat{\theta}_m - \theta_o) \xrightarrow{d} N\left(0, \frac{\sum_{-\infty}^{\infty} \operatorname{corr}\{u_1(\theta_o), u_{1+s}(\theta_o)\}}{V(\theta_o)}\right) \tag{4}$$

in the scalar case. Note that, if $\{Y_t\}$ is a sequence of i.i.d. random

variables, the sum in (4) is equal to 1, and the denominator is the usual Fisher information.

The asymptotic normality of $u(\theta_o)$ depends on the fulfilling of a suitable mixing condition. The book of Ibragimov and Linnik (1971, p. 315-364 and p.420-424) contains various limit theorems for stationary sequences. We only point out that $\{u_t\}$ is a function of a finite number of co-ordinates of $\{Y_t\}$; therefore, if $\{Y_t\}$ satisfies some mixing condition, $\{u_t\}$ satisfies the same mixing condition for an asymptotically equivalent mixing sequence.

5.3 Efficiency for Normal Markov process plus noise

In this section, we compare $\hat{\theta}_m$ with other estimators in a special case. Suppose that the process $\{X_t\}$ is stationary and satisfies

$$X_t = \rho X_{t-1} + \varepsilon_t , \qquad |\rho| < 1, \ t = 0, \pm 1, \ldots ,$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. $N(0,\sigma_\varepsilon^2)$. We observe $Y_t = X_t + \eta_t$ where $\{\eta_t\}$ is a sequence of i.i.d. $N(0,\sigma_\eta^2)$ which is independent of $\{\varepsilon_t\}$.

This model was mentioned in the previous chapter in a filtering framework. However, here we are concerned with estimation of ρ. A simple estimate, based on the fact that $\gamma_h \equiv \mathrm{cov}(Y_t, Y_{t+h}) \propto \rho^h$ for $h \neq 0$, is

$$\tilde{\rho} = \left[T^{-1} \sum_{t=3}^{T} Y_t Y_{t-2}\right] \Big/ \left[T^{-1} \sum_{t=2}^{T} Y_t Y_{t-1}\right] = c_2/c_1 , \quad \text{say,}$$

which is easily extended to the multivariate case (Anderson *et al.*, 1969). Since also of independent interest, we digress to study the asymptotic variance of $\tilde{\rho}$ and its ARE. Using a standard asymptotic argument, we have

$$T \ \mathrm{var}(\tilde{\rho}) \ \sim \ \gamma_1^{-2} \mathrm{var}(c_2) - 2\gamma_2 \gamma_1^{-3} \mathrm{cov}(c_1,c_2) + \gamma_2^2 \gamma_1^{-4} \mathrm{var}(c_1)$$

which, after some algebra, gives

$$T \operatorname{var}(\tilde{\rho}) \sim \frac{1-\rho^2}{\rho^2}\{1 + 2(1-\rho^2)\omega + (1-\rho^4)\omega^2\} \tag{5}$$

where $\omega = \sigma_\eta^2/\sigma_\varepsilon^2$, using known formulae for the asymptotic covariances of sample covariances (Anderson, 1971, p.466).

Consider now the variance of $\hat{\rho}_\infty$, the exact MLE, or any efficient estimator. The computation of efficient estimates for the case when $\{X_t\}$ is a p-order autoregressive process has been studied by Pagano (1974). Here we are concerned with their variances for the case $p = 1$. The process $\{Y_t\}$ has a ARMA(1,1) representation, i.e. of the form

$$Y_t - \rho Y_{t-1} = e_t - \delta e_{t-1} \tag{6}$$

where $\{e_t\}$ is a sequence of i.i.d. $N(0,\sigma_e^2)$; see e.g. Pagano (1974). We determine δ and σ_e^2 equating expressions for γ_0 and γ_1 based on (6) and the previous representation. Then

$$\delta = 2\rho \Big/ \left\{ 1 + \rho^2 + \omega^2 + \left[(1-\rho^2)^2 + 2\omega(1+\rho^2) + \omega^2\right]^{\frac{1}{2}} \right\} ,$$

$$\sigma_e^2 = \sigma_\varepsilon^2 / \{(1-\rho\delta)(1-\delta/\rho)\} ,$$

discarding a root which gives a non-invertible process in (6). It is easy to check that $0 < \delta/\rho < 1$.

Define $\psi = (\rho,\delta,\sigma_e^2)'$, $\phi = (\rho,\sigma_\varepsilon^2,\sigma_\eta^2)'$, whose information matrices are related by $I_\phi = A' I_\psi A$, where $A = ((a_{i,j})) = ((\partial\psi_i/\partial\phi_j))$. The first row of A is (1,0,0) and so is the first row of A^{-1}, which exists. Therefore, the entry in the top left corner of I_ϕ^{-1} is equal to the corresponding entry of I_ψ^{-1}, which is known (Box and Jenkins, 1970, p.248). Hence

$$T \operatorname{var}(\hat{\rho}_\infty) \sim \frac{1-\rho^2}{\rho^2}\left(\frac{1-\delta\rho}{1-\delta/\rho}\right)^2 . \tag{7}$$

The plot of the ratio of (5) and (7) versus ρ for various values of ω is given in Figure 5.1. Unless ω is large, the efficiency of $\tilde{\rho}$ is high. Since neither the variance of the moments nor that of the MLE is affected to the first order of approximation if the mean is unknown, then the above conclusions hold also for the case of unknown mean.

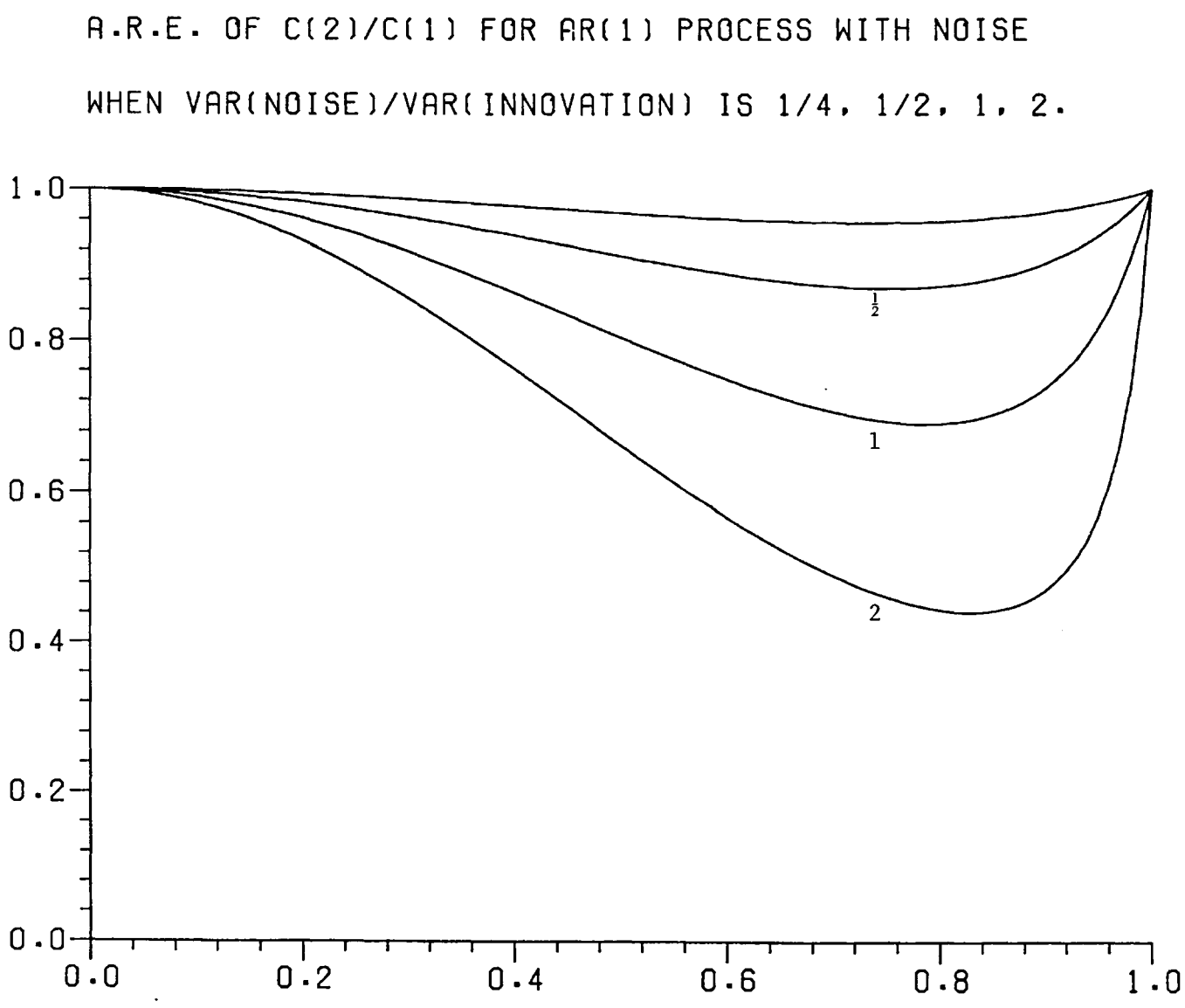


Figure 5.1: ARE of c_2/c_1 for AR(1) Normal process plus noise, when ω = 1/4, ½, 1, 2.

Table 5.1: Asymptotic variance of MLE of ρ and ARE of $\tilde{\rho}$ and $\hat{\rho}_1$ for AR(1) process plus noise, (a) I_ρ^{-1}, (b) ARE of $\tilde{\rho}$, (c) ARE of $\hat{\rho}_1$.

ρ	$\omega = 1/4$			$\omega = \frac{1}{2}$			$\omega = 1$			$\omega = 2$		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
0.01	1.562	1.000	1.000	2.249	1.000	1.000	3.998	1.000	1.000	8.994	1.000	1.000
0.10	1.535	0.998	0.998	2.191	0.993	0.996	3.835	0.985	0.990	8.454	0.974	0.983
0.20	1.456	0.991	0.994	2.026	0.975	0.984	3.393	0.946	0.965	7.064	0.907	0.940
0.30	1.333	0.981	0.988	1.783	0.949	0.969	2.798	0.892	0.934	5.355	0.820	0.888
0.40	1.175	0.968	0.982	1.499	0.920	0.955	2.176	0.836	0.907	3.761	0.735	0.844
0.50	0.995	0.956	0.977	1.205	0.892	0.945	1.610	0.789	0.982	2.488	0.666	0.819
0.60	0.801	0.947	0.975	0.922	0.873	0.944	1.135	0.759	0.894	1.562	0.625	0.821
0.70	0.601	0.942	0.976	0.660	0.867	0.950	0.754	0.754	0.913	0.928	0.623	0.855
0.80	0.400	0.947	0.981	0.422	0.880	0.965	0.452	0.784	0.946	0.504	0.674	0.922
0.90	0.200	0.964	0.989	0.204	0.921	0.981	0.209	0.862	0.979	0.216	0.797	0.992
0.95	0.100	0.979	0.994	0.101	0.955	0.988	0.102	0.923	0.986	0.103	0.889	1.000

Now suppose that only ρ is unknown. The Fisher information is $I_\rho = a' I_\psi a$ where a is the first column of A. The 1-order loglikelihood is

$$\ell(\rho) = \sum_{t=1}^{T} \{-\tfrac{1}{2} \log \sigma_c^2 - \frac{1}{2\sigma_c^2}(Y_t - \rho_Y Y_{t-1})^2\} \qquad (8)$$

where

$$\rho_Y = \frac{\rho}{1 + \omega(1-\rho^2)}, \qquad \sigma_c^2 = (1-\rho_Y^2)\ \mathrm{var}(Y_t). \qquad (9)$$

The computation of (4) is conceptually straightforward but algebraically messy and no closed form for $\mathrm{var}(\hat{\rho}_1)$ could be found. Numerical values of the ARE are given in Table 5.1, which also gives the corresponding value for the moment estimate based on the first relationship of (9), say $\tilde{\rho}$. The behaviour of $\hat{\rho}_1$ is satisfactory over the whole range of the parameter values considered. Its ARE is often half way between the corresponding one of $\tilde{\rho}$ and 1, and sometimes is much higher.

5.4 Efficiency for second-order Markov chain

Let $\{Y_t\}$ be a stationary binary second-order Markov chain, with transition probabilities

$$p_{ij} = P\{ Y_t = 1 \mid Y_{t-2} = i,\ Y_{t-1} = j \}$$

for $i,j \in \{0,1\}$. The associated first-order Markov chain with states $\{(0,0),(0,1),(1,0),(1,1)\}$ has transition matrix Ω and stationary distribution $\pi = (\pi_1,\pi_2,\pi_3,\pi_4)'$, where

$$\pi_2 = \pi_3 = \left\{\frac{1-p_{10}}{p_{00}} + 2 + \frac{p_{01}}{1-p_{11}}\right\}^{-1},$$

$$\pi_1 = \pi_3\ (1-p_{10})/p_{00}, \qquad \pi_4 = \pi_3\ p_{01}/(1-p_{11}).$$

If $\omega_{ij} = \omega_{ij}(\theta)$ where θ is an unknown parameter, the variance of the

exact MLE is asymptotically proportional to

$$\left(\sum_{\omega_{ij}>0} \frac{\pi_i}{\omega_{ij}} \left(\frac{d\omega_{ij}}{d\theta} \right)^2 \right)^{-1}$$

(Billingsley, 1961). For the approximate MLE, consider $m = 0$ and $m = 1$. For $m = 0$, we approximate the loglikelihood by that of a sequence of i.i.d. Bernoulli r.v.'s with probability of success $\phi_0 = \pi_3 + \pi_4$. For $m = 1$, consider

$$\ell(\theta) = \sum_{t=1}^{T} \log \phi_1(y_{t-1}, y_t)$$

where $\phi_1(i,j) = P\{Y_t = j \mid Y_{t-j} = i\}$. In particular $\phi_1(0,1) = \pi_3/(\pi_1 + \pi_3)$, $\phi_1(1,1) = \pi_4/(\pi_2 + \pi_4)$.

Numerical work has been done using the parameterization

$$p_{ij}(\theta) = i\theta(1-\theta) + \tfrac{1}{2} j (1-\theta) + \delta , \qquad (0 < \theta < 1),$$

where δ is a constant regarded as known, $0 < \delta < 7/16$. Results are given in Table 5.2. Their interpretation is not so simple as for Table 5.1. For $\delta = 0.2$, the ARE varies appreciably as θ ranges between 0 and 1, while for $\delta = 7/16$ it is smooth and moderately high. Notice that, near $\delta = 0$, $\hat{\theta}_0$ is more efficient than $\hat{\theta}_1$; this illustrates the point that, although the ARE must increase as $m \to \infty$, it may not do so monotonically.

In a general way, the two examples suggest that the method works reasonably well and is worth considering in any more complex case where the "exact" MLE cannot be computed.

Table 5.2: Asymptotic variance of $\hat{\theta}_2$ and ARE of $\hat{\theta}_0$ and $\hat{\theta}_1$ for second order binary Markov chain,

(a) T var($\hat{\theta}_2$) as $T \to \infty$, (b) ARE of $\hat{\theta}_0$, (c) ARE of $\hat{\theta}_1$.

θ	$\delta = 0.2$			$\delta = 7/16$		
	(a)	(b)	(c)	(a)	(b)	(c)
0.01	0.867	0.268	0.185	0.254	0.881	0.756
0.10	1.497	0.192	0.170	0.255	0.828	0.805
0.20	2.809	0.041	0.210	0.251	0.772	0.833
0.24	3.459	0.002	0.307	0.250	0.755	0.840
0.26	3.734	0.002	0.382	0.250	0.748	0.844
0.30	4.030	0.056	0.569	0.251	0.737	0.850
0.40	3.261	0.359	0.935	0.262	0.729	0.862
0.50	2.236	0.559	0.994	0.285	0.749	0.871
0.60	1.657	0.634	0.939	0.318	0.781	0.873
0.70	1.313	0.631	0.849	0.351	0.801	0.859
0.80	1.054	0.569	0.739	0.370	0.780	0.822
0.90	0.813	0.459	0.609	0.366	0.706	0.759
0.95	0.687	0.389	0.534	0.354	0.650	0.717
0.99	0.580	0.327	0.467	0.338	0.598	0.677

Appendix A

PROVING UNIMODALITY OF THE LIKELIHOOD OF STATIONARY MARKOV PROCESSES

We prove here that equation (9) of Chapter 1 has one and only one real root in the open interval $(-1, 1)$. Since with probability 1

$$C(1) = -4\{d_{11} - 2d_{12} + d_{22}\}/T < 0,$$

$$C(-1) = 4(T-1)\{(T-1)(d_{11} + 2d_{12} + d_{22}) - (b_{11} + 2b_{12} + b_{22})\}/T > 0,$$

then there is one real root in $(-1, 1)$. For the case $T = 2$, the proof is complete. Assume then $T > 2$ and prove that there is only one real root. Define

$$\ell_{max}(\alpha) = \ell(\hat{\mu}(\alpha), \hat{\sigma}^2(\alpha), \alpha)$$

$$= -\tfrac{1}{2} n T \log \hat{\sigma}^2(\alpha) + \tfrac{1}{2} n \log(1 - \alpha^2) + \text{constant}$$

with $\hat{\mu}(\alpha)$, $\hat{\sigma}^2(\alpha)$ defined by (8) of Chapter 1. It suffices to prove that ℓ_{max} has only one stationarity point, and therefore it is a maximum. It is sufficient to prove that ℓ_{max} is concave at points such that $\ell'_{max}(\alpha) = 0$, for which

$$\frac{d}{d\alpha}\hat{\sigma}^2(\alpha) = -\frac{2\alpha\,\hat{\sigma}^2(\alpha)}{T(1-\alpha^2)}.$$

Using the last relationship in ℓ''_{max}, we see that it is sufficient to prove that $\hat{\sigma}^2(\alpha)$ is convex. Now

$$\frac{d^2}{d\alpha^2}\hat{\sigma}^2(\alpha) = \frac{2N(\alpha)}{n^2 T \{T - (T-2)\alpha\}^3 (T-2)}$$

where

$$N(\alpha) = \{n(T-2)d_{22} - d_2^2\}\{T - (T-2)\alpha\}^3 + 2\{T d_2 - (T-2) d_1\}^2$$

which is positive with probability one.

This ends the proof. As a curiosity, note that $c_5 > 0$ with probability one; therefore, there is another real root on the right of $\alpha = 1$, and a

third real root on the left of $\alpha = -1$. The remaining two roots could not be located; however, they have been computed numerically for a large number of sets of data: in all cases, these roots were on the right of $\alpha = 1$, possibly complex.

Appendix B

A MARKOV PROCESS WITH BETA MARGINAL DISTRIBUTION

Look for a stationary Markov process $\{X_t\}$ with Beta(p,q) marginal distribution, i.e. with density at x proportional to $x^{p-1}(1-x)^{q-1}$ for $0 < x < 1$, $p > 0$, $q > 0$. D.R.Cox has suggested the model

$$X_t = X_{t-1}^{\alpha} \varepsilon_t \qquad (0 < \alpha < 1,\ t = 0, \pm 1, \ldots) \quad (B.1)$$

for a sequence of i.i.d. random variables $\{\varepsilon_t\}$ to be determined. Taking logarithms, we have

$$Z_t = \alpha Z_{t-1} + \xi_t$$

in an obvious notation. Clearly, if the problem has a solution, the characteristic function of ξ is

$$\phi_\xi(u) = E(e^{iuZ}) \,/\, E(e^{iu\alpha Z})$$

$$= \frac{\Gamma(p+iu)\ \Gamma(p+q+iu\alpha)}{\Gamma(p+iu\alpha)\ \Gamma(p+q+iu)} \qquad (B.2)$$

having dropped the suffix t. Consider (B.2) for $q = 1$. Then

$$\phi_\xi(u) = \frac{p+iu}{p+iu} = \frac{1-\alpha}{1+iu/p} + \alpha \ ,$$

which is the characteristic function of

$$\zeta_{p,\alpha} = \begin{cases} 0 & \text{with probability } \alpha, \\ -e_p & \text{with probability } 1-\alpha, \end{cases}$$

where e_p is an Exponential r.v. with mean $1/p$. Note the close similarity with the problem of Gaver and Lewis (1980). For general q integer,

$$\phi_\xi(u) = \frac{(p+iu\alpha)\ (p+1+iu\alpha)\ \ldots\ (p+q-1+iu\alpha)}{(p+iu)\ (p+1+iu)\ \ldots\ (p+q-1+iu)} \ ,$$

so that

$$\xi = \sum_{j=1}^{q} \zeta_{p+j-1,\alpha}$$

where all components of the right hand side are independent.

Using the fact

$$E(\varepsilon^k) = \frac{\Gamma(p+k)\ \Gamma(p+q+k\alpha)}{\Gamma(p+k\alpha)\ \Gamma(p+q+k)} \quad , \tag{B.3}$$

we get

$$E(X_t\ X_{t+h}) = E\left(X_t^{1+\alpha^h}\ \varepsilon_{t+1}^{\alpha^{h-1}} \cdots \varepsilon_{t+h-1}^{\alpha}\ \varepsilon_{t+h}\right)$$

$$= \frac{\Gamma(p+\alpha^h+1)\ \Gamma(p+q)}{\Gamma(p)\ \Gamma(p+q+\alpha^h+1)} \prod_{j=1}^{h} \frac{\Gamma(p+\alpha^{j-1})\ \Gamma(p+q+\alpha^j)}{\Gamma(p+\alpha^j)\ \Gamma(p+q+\alpha^{j-1})}$$

$$= \frac{p(p+\alpha^h)}{(p+q)(p+q+\alpha^h)}$$

for $h > 0$. Hence, for general h,

$$\mathrm{corr}(X_t, X_{t+h}) = \frac{\alpha^{|h|}\ (p+q+1)}{p+q+\alpha^{|h|}} \quad .$$

Using the fact that $1-X$ is a Beta(q,p) r.v. if X is a Beta(p,q) r.v., we can generate processes with q fractional and p integer. The problem of both p and q fractional remains unsolved, as well as the problem of negative autocorrelation.

Consider now the generation of a stationary Markov process of form (B.1) such that the marginal distribution is Gamma(p,1). For $p = 1$, Shanbag and Sreehari (1977) have proved that $\varepsilon = S_\alpha^{-\alpha}$, where S_α is a positive stable r.v. with characteristic exponent α. A method for generating S_α has been given by Kanter (1975). For $p \neq 1$, the problem is unsolved, as it has been pointed out by McKenzie (1981). An approximate solution is to generate a process of type (B.1) with Beta(p,q) marginal distribution and multiply by q, for $q \gg 1$.

REFERENCES

Abramowitz, M. and Stegun, I.A., Editors (1964) Handbook of mathematical functions. Dover. New York.

Andersen, E.B. (1973) Conditional inference and models for measuring. Mentalhygiejnisk Forlag. Køben haven.

Anderson, T.W. (1971) The statistical analysis of time series. Wiley. New York.

——— (1978) Repeated measurements on autoregressive processes. J.Amer. Statist.Assoc., 73, 371-378.

——— and Cheng Hsiao (1981) Estimation of dynamic models with error components. J.Amer.Statist.Assoc., 76, 598-606.

Anderson, W.N. *et al.* (1969) Consistent estimates of the parameters of a linear system. Ann.Math.Statist., 40, 2064-2075.

Balestra, P. and Nerlove, M. (1966) Pooling cross section and time series data in the estimation of a dynamic model: the demand for natural gas. Econometrica, 34, 585-612.

Bather, J.A. (1965) Invariant conditional distributions. Ann.Math.Stat., 36, 829-846.

Berk, R.H. (1972) Consistency and asymptotic normality of MLE's for exponential models. Ann.Math.Statist., 43, 193-204.

Billingsley, P. (1961) Statistical inference for Markov processes. Chicago Univ. Press.

Borpujari, A.S. (1977) An empirical Bayes approach for estimating the mean of N stationary time series. J.Amer.Statist.Assoc., 72, 397-402.

Box, G.E.P. (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance and correlation between errors in two-way classification. Ann.Math.Statist., 25, 484-498.

——— and Jenkins, G.M. (1970) Time series analysis, forecasting and control. Holden Day. San Francisco.

Brillinger, D.R. (1980) Analysis of variance and problems under time series models. In: Handbook of Statistics 1; Analysis of Variance. Edited by P.R. Krishnaiah. North-Holland. Amsterdam.

Chipman, J.S. *et al.* (1968) Efficiency of the sample mean when residuals follow a first-order stationary Markoff process. J.Amer.Statist. Assoc., 63, 1237-1246.

Cox, D.R. (1961) Tests of separate families of hypotheses. Proc. 4th Berkeley Symp., 1, 105-123.

——— (1962) Further results on tests of separate families of hypotheses. J.Roy.Statist.Soc., 24, 406-424.

——— (1981) Statistical analysis of time series: some recent developments. Scand.J.Statist., 8, 93-115.

Davies, R.B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 64, 247-254.

Farrow, M. (1979) Some stochastic models and their application in aluminium reduction-cell control studies. Ph.D. thesis. University of Newcastle upon Tyne.

Gaver, D.P. and Lewis, P.A.W. (1980) First order autoregressive Gamma sequences and point processes. Adv.Appl.Prob., 12, 727-745.

Goodrich, R.L. and Caines, P.E. (1979) Linear system identification from nonstationary cross-sectional data. IEEE Trans.Autom.Contr., AC-24, 403-411.

Heckmann, J.J. (1978) Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state depencence. Ann. INSEE, 30-31, 227-270.

Ibragimov, I.A. and Linnik, Yu.V. (1971) Independent and stationary sequences of random variables. Wolters-Noordhoff. Groningen.

Jöreskog, K.G. (1978) An econometric model for multivariate panel data. Ann.INSEE, 30-31, 355-366.

Kalman, R.E. (1963) New methods in Wiener filtering theory. In: Proceedings of the first symposium on engineering applications of random functions and probability. Edited by J.L. Bogdanoff and K. Kozin. Wiley. New York.

Kanter, M. (1975) Stable densities under change of scale and total variation inequalities. Ann.Prob., 3, 697-707.

Khamis, S.H. (1960) Incomplete Gamma functions expansions of statistical distribution functions. Bull.Inst.Int.Statist., 37, 385-396.

Larsen, W.A. (1969) The analysis of variance for two-way classification model with observation within a row serially correlated. Biometrika, 56, 509-515.

Liu, L.M. and Tiao, G.C. (1980) Random coefficient first-order autoregressive model. J.Econometrics, 13, 305-325.

Ljung, G.M. and Box, G.E.P. (1980) Analysis of variance with autocorrelated observations. Scand.J.Statist., 7, 172-180.

Marodier, P. and Trognon, A. (1978) Heteroscedasticity and statification in error component models. Ann.INSEE, 30-31, 451-482.

McKenzie, E. (1981) Product autoregression: a time series characterization of the Gamma distribution. To appear in J.Appl.Prob.

Nerlove, M. (1978) Econometric analysis of longitudinal data: approaches, problems and prospects. Ann.INSEE, 30-31, 9-22.

Nicholls, D.F. and Quinn, B.G. (1980) The estimation of random coefficient autoregressive models, I. J.Time Series Analysis, 1, 37-46.

Pagano, M. (1974) Estimation of models of autoregressive signal plus noise. Ann.Statist., 2, 99-108.

Robinson, P.M. (1978) Statistical inference for a random coefficient autoregressive model. Scand.J.Statist., 5, 163-168.

Sawitzki, G. (1979) Exact filtering in exponential families: discrete time. In: Stochastic control theory and stochastic differential equations. Edited by M. Kohlmann and W. Vogel. Springer.

——— (1981) Finite dimensional filter systems in discrete time. Stochastics, 5, 107-114.

Shanbhag, D.N. and Sreehari, M. (1977) On certain self-decomposable distributions. Z.Wahrsheinlichkeit. verw. Geb., 38, 217-222.

Shumway, R.H. (1970) Applied regression and analysis of variance for stationary time series. J.Amer.Statist.Assoc., 65, 1527-1546.

Siddiqui, M.M. (1958) On the inversion of the sample covariance matrix in a stationary autoregressive process. Ann.Math.Statist., 29, 585-588.

Thornber, H. (1967) Finite Monte-Carlo studies: an autoregressive illustration. J.Amer.Statist.Assoc., 62, 801-818.

Tiao, G.C. and Tan, W.Y. (1966) Bayesian analysis of random-effect models in the analysis of variance, II. Effect of autocorrelated errors. Biometrika, 53, 477-495.

Wansbeek, T.J. and Kaptein, A. (1978) The separation of individual variation and systematic change in the analysis of panel data. Ann.INSEE, 30-31, 659-680.

Wegman, E.J. (1974) Some results on non stationary first order autoregression. Technometrics, 16, 321-322.

White, J.S. (1961) Asymptotic expansions for the mean and variance of the serial correlation coefficient. Biometrika, 48, 85-94.

Wise, J. (1955) The autocorrelation function and the spectral density function. Biometrika, 42, 151-159.

Zacks, S. (1971) The theory of statistical inference. Wiley. New York.

Biometrika (1981), 68, 1, *pp.* 326–8
Printed in Great Britain

A note on the estimation of a distribution function and quantiles by a kernel method

BY A. AZZALINI
Istituto di Statistica, Università di Padova, Italy

SUMMARY

A distribution function is estimated by integrating a kernel estimator of the density. Quantiles are estimated by inverting the estimate of the distribution function. Second-order properties of both these estimators are studied.

Some key words: Distribution function; Kernel method; Nonparametric estimation; Quantiles; Simulation.

There has recently been extensive work on the estimation by kernel methods of probability densities and their derivatives; for a review, see Wertz (1978). For estimating the cumulative distribution function, the empirical distribution function, F_n, is in a sense already quite smooth, although it is known that further smoothing can be an advantage. The object of the present note is to report briefly an investigation of kernel estimation applied to distribution function, and quantiles. Further details are available from the author.

For independent and identically distributed random variables $X_1, \ldots, X_n$ the usual kernel estimator of the density $f(.)$ is

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} w\left(\frac{x-X_i}{b}\right),$$

where $w(.)$ is some bounded density function with $w(t) = w(-t)$ for all real t; $b > 0$. The corresponding estimate for the distribution function $F(.)$ is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} W\left(\frac{x-X_i}{b}\right),$$

where

$$W(t) = \int_{-\infty}^{t} w(u)\,du.$$

Nadaraya (1964) has proved under mild conditions that $\hat{F}$ has asymptotically the same mean and variance as F_n. We consider here second-order properties of $\hat{F}$ under the assumption that $b \to 0$ as $n \to \infty$, that $w(t) = 0$ if $|t| > h$ for some positive h, that $f(.)$ is continuous in $(x-hb, x+hb)$ and that $f'(x)$ exists. Then it can be proved that, as $n \to \infty$,

$$E\{\hat{F}(x)-F(x)\}^2 \sim F(x)\{1-F(x)\}/n - ub/n + vb^4, \tag{1}$$

where

$$u = f(x)\left\{h - \int_{-h}^{h} W^2(t)\,dt\right\}, \quad v = \left\{\tfrac{1}{2}f'(x)\int_{-h}^{h} t^2 w(t)\,dt\right\}^2.$$

Note that $u > 0$ if $f(x) > 0$. We get the minimum of the right-hand side of (1) at

$$b = \left(\frac{u}{4vn}\right)^{1/3}, \tag{2}$$

provided $f'(x) \neq 0$, for which

$$E\{\hat{F}(x)-F(x)\}^2 \sim \frac{F(x)\{1-F(x)\}}{n}-\frac{3}{v^{1/3}}\left(\frac{u}{4n}\right)^{4/3}.$$

If we write $b = Kn^{-\delta}$, then the asymptotically optimum choice for δ is $\delta = \frac{1}{3}$. This is different from the optimum value for density estimation, which is $\delta = \frac{1}{5}$. If we choose $\delta \leqslant \frac{1}{4}$ then (1) shows that the mean squared error of $\hat{F}$ has an order of magnitude greater than that of the mean squared error of F_n.

Some numerical work has been done using Epanechnikov's kernel, the integral of which is

$$W_0(t) = \begin{cases} 0 & (t \leqslant -\sqrt{5}), \\ \frac{1}{2}+3t/(4\sqrt{5})-t^3/(20\sqrt{5}) & (-\sqrt{5} < t < \sqrt{5}), \\ 1 & (t \geqslant \sqrt{5}), \end{cases}$$

which is known to have optimal properties for estimating the density function. The value of K in $b = Kn^{-\delta}$ has been chosen as follows. An invariance argument applied to (2) shows that K has to be proportional to σ, the standard deviation of X_1. Numerical computation of the optimum K shows that it ranges between σ and 2σ for a large number of distributions, including those of Table 1, when we are estimating the long tail of F. The value $K = 1{\cdot}3\sigma$ has been found a satisfactory compromise. When we are estimating F not in the long tail, $K = 0{\cdot}5\sigma$ is a more reliable choice.

One relevant application of $\hat{F}$ is the estimation of the distribution of some statistic via simulation; in this case it is fairly common that σ is known exactly; however, some evidence can be given that the results are hardly affected if σ is estimated from the sample.

Table 1 reports some computations of the mean squared error of $\hat{F}$, when F is estimated at the pth quantile ζ_p, defined as $p = F(\zeta_p)$ $(0 < p < 1)$; the exact mean and variance of $\hat{F}$ have been computed by numerical quadrature of appropriate integrals. We notice that $\hat{F}$ is generally rather better than F_n, whose mean squared error is $p(1-p)/n$. This is not so for the gamma($\frac{1}{2}$) distribution at the 0·05 quantile, because the density is unbounded in $(x-hb, x+hb)$.

Now consider the estimation of ζ_p, assumed to be unique, with x_p defined as the root of $p = \hat{F}(x_p)$. The asymptotic distribution of x_p has been considered by Nadaraya (1964) to the first order of approximation; under certain regularity conditions, x_p has the same asymptotic distribution as the sample quantile.

Table 1. *Estimation of distribution function at the* pth *quantile* $(n = 1000)$: $n \times$ *mean squared error*

	$p = 0{\cdot}05$, $p(1-p) = 0{\cdot}0475$		$p = 0{\cdot}90$, $p(1-p) = 0{\cdot}09$		$p = 0{\cdot}95$, $p(1-p) = 0{\cdot}0475$	
Distribution	$K = 1{\cdot}3\sigma$	$K = 0{\cdot}5\sigma$	$K = 1{\cdot}3\sigma$	$K = 0{\cdot}5\sigma$	$K = 1{\cdot}3\sigma$	$K = 0{\cdot}5\sigma$
Normal	0·04305	0·04477	0·08193	0·08526	0·04305	0·04477
Gamma($\frac{1}{2}$)	8·91245	1·97219	0·08445	0·08755	0·04472	0·04633
Gamma(1)	0·71826	0·03334	0·08388	0·08724	0·04430	0·04612
Gamma(2)	0·09075	0·03976	0·08335	0·08690	0·04391	0·04590
Gamma(5)	0·05030	0·04269	0·08279	0·08648	0·04349	0·04562
Beta(5, 5)	0·04254	0·04434	0·08070	0·08583	0·04254	0·04434
Beta(2, 5)	0·04716	0·04162	0·08179	0·08613	0·04272	0·04519
Beta(2, 10)	0·05782	0·04074	0·08264	0·08657	0·04334	0·04559
Beta(1, 1)	0·02976	0·03920	0·06842	0·08170	0·02976	0·03920
Beta(1, 4)	0·19292	0·02995	0·08271	0·08678	0·04323	0·04560

We consider here second-order approximations. Under the previous assumptions, it can be proved that (2) is again the asymptotically optimum choice of b and, correspondingly,

$$E(x_p) \sim \zeta_p - \frac{\operatorname{sgn}(f')}{f v^{1/6}} \left(\frac{u}{4n}\right)^{2/3}, \quad E(x_p - \zeta_p)^2 \sim \frac{p(1-p)}{nf^2} - \frac{3}{f^2 v^{1/3}} \left(\frac{u}{4n}\right)^{4/3},$$

where the argument of all functions is ζ_p.

Some numerical comparison of x_p with the sample quantile has been carried out but is not reported here. The relative performances of x_p and the sample quantile are similar to those of $\hat{F}$ and F_n.

The value of x_p can be found by Newton–Raphson search starting from the sample quantile. This proved to be very effective with some simulated data.

The author is grateful to Dr A. C. Atkinson and Professor D. R. Cox for several comments and suggestions; his work at Imperial College, London was supported by the Italian Ministry of Education.

References

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory Prob. Applic.* **15**, 497–500.

Wertz, W. (1978). *Statistical Density Estimation: A Survey*. Göttingen: Vandenhoeck & Ruprecht..

[*Received July* 1980. *Revised October* 1980]