Time Series Models for Discrete Data

by

Iain Lachie MacDonald

Thesis Presented for the Degree of
Doctor of Philosophy
in Mathematical Statistics.
University of Cape Town
April 1992

air a choisrigidh do m' phàrantan

Abstract

Time Series Models for Discrete Data, by Iain Lachie MacDonald. Thesis presented for the degree of Doctor of Philosophy, University of Cape Town, April 1992.

This thesis considers the statistical problem of the modelling of discrete-valued time series. Such series arise in many scientific contexts, and although they are often represented by models based on the normal distribution, there are circumstances in which the discrete nature of the observations should be respected. The approach followed in this thesis is to model such series by sequences of dependent discrete random variables.

Chapter 1 consists of a survey of the models of this kind that are available for discrete-valued series. These models include Markov chains of first order or higher, several classes of models that parallel the familiar (Gaussian) autoregressive moving average processes, Markov regression models, the family of models known as 'parameter-driven' processes, and state-space models.

After reviewing relevant aspects of the theory of hidden Markov models as used in speech processing, Chapter 2 introduces hidden Markov models as general-purpose models for discrete-valued time series. These models are based on an underlying and unobserved stationary Markov chain and either a Poisson or a binomial distribution. The key property is that, conditional on the underlying Markov chain, the observations are assumed to be independent random variables with distributions specified by the current state of the Markov chain. Correlation properties are derived, an algorithm for evaluating the likelihood function is described, and direct numerical maximization of the likelihood is proposed as the most practical means of parameter estimation. Marginal, joint and conditional distributions are derived, and applications of these results to forecasting and the treatment of missing data are described.

Chapter 3 presents extensions and modifications of the basic hidden Markov models introduced in Chapter 2. These extensions include models for categorical series, multivariate models and models which cater for time trend, seasonality or, more generally, dependence on covariates.

Chapter 4 describes seven illustrative applications of the models introduced in Chapters 2 and 3. These applications are to: the durations of eruptions of a geyser; time series of births; locomotory behaviour of locusts; wind direction; evapotranspiration; thinly traded shares listed on the Johannesburg Stock Exchange; and homicide and suicide statistics.

Acknowledgements

It is customary for Ph.D. candidates, when submitting their theses, to say complimentary things about their supervisors. I believe that in this case, however, there is far more justification than usual. Prof. Walter Zucchini has been an inspiring supervisor, and it has been a privilege to have been given the benefit of his knowledge, his wide experience, his enthusiasm and his patience. I am most grateful to him. I owe a debt of gratitude also to Prof. Geoff Pegram of the University of Natal, who was co-supervisor in the early stages of this work and later commented on a draft of this thesis. The Friday afternoons I spent in the Civil Engineering department in Durban were time well spent!

Other members of the U.C.T. department of Statistical Sciences were most encouraging. In this regard it is difficult not to name the whole department, but I shall take the risk of naming only two: Prof. June Juritz, who was my first lecturer in the subject and has since continued to show interest in my work, and Dr Mary Lou Thompson, who never missed an opportunity to offer encouragement.

I am grateful to Dr Brian Leroux for sending me his work before publication; to Prof. Adelchi Azzalini for answering my enquiries about some of his work; to Mr Peter Digby for his advice on computing; and to those who provided the unpublished data analyzed in Chapter 4 and in some cases spent many hours discussing it with me: Dr Linda Haines, Dr David Raubenheimer, Mr Frikkie Potgieter, Mr Graham Fick, Mr David Bowie and Dr Len Lerer. It must however be understood that none of the above people bear any responsibility for the opinions expressed in this thesis: such responsibility is entirely mine.

My colleagues in Actuarial Science have all been most supportive, especially Prof. Rob Dorrington, who has in my interest made considerable sacrifices of his own time. Dr Eric Martens, my colleague until recently, contributed both mathematical and musical stimulation (and thereby provided a counterexample to Halmos).

I wish to acknowledge the financial support of the University of Cape Town as my employer: in particular for granting me six months' study and research leave, and for Staff Development awards which provided personal computing facilities.

My friends have given me all the encouragement and assistance that they possibly could while I was working on this thesis, but none more so than Dr Bice Martincigh, for whose support I am most grateful.

Finally I express my thanks to my parents, who have done more for me than I can begin to recount, and to whom this thesis is therefore dedicated.

Contents

List of Tables

Chapter 1

A survey of models for discrete-valued time series

1.1 Introduction: the need for discrete-valued time series models

Many of the time series which occur in practice are by their very nature discrete-valued, although it is often quite adequate, and obviously very convenient, to represent them by means of models based on the normal distribution. Some examples of discrete-valued series are:

(i) the numbers of defective items found in successive samples taken from a production line;

(ii) the sequence of wet and dry days at some site;

(iii) the numbers of cases of some notifiable disease in a given area in successive months;

(iv) the numbers of births, and the numbers of deliveries by various methods, at a hospital in successive months;

(v) road accident or traffic counts;

(vi) base sequences in DNA;

(vii) the presence or absence of trading in a particular share on consecutive trading days;

(viii) the numbers of firearm homicides and suicides in successive weeks in a given area; and

(ix) the behaviour category of an animal observed at regular intervals.

Although in some cases models based on the normal distribution will suffice, clearly this will not always be so. When the observations are categorical in nature, and when the observations are quantitative but fairly small, it is necessary to take into account explicitly the discrete nature of the data.

Furthermore, there are continuous-valued series in which the observations naturally fall in one of a small number of categories, for instance the series of lengths of eruptions of the 'Old Faithful' geyser analyzed by Azzalini and Bowman (1990). In that case most of the observations can be described as 'long' or 'short', with very few eruptions intermediate in length, and the pattern of long and short eruptions is the aspect of most scientific interest. It is therefore natural to treat this series as a binary time series. Another instance of continuous data being treated as discrete is the practice of classifying directional observations into the conventional sixteen points of the compass, even if more detailed information is available, because of the familiarity of this classification. The observations of wind direction analyzed in section 4.5 are an example of a time series of this kind.

The approach followed in this work is to seek models consisting of appropriately dependent sequences of discrete random variables, and to develop

means of fitting and selecting such models. There is however another, quite different, approach which will not be pursued here. This is to recognize that many discrete-valued time series arise as counts in a point process on the line, and to treat them accordingly. For instance, Guttorp (1986) describes the fitting of point process models to binary time series generated by a point process, and Guttorp and Thompson (1990) discuss several methods for estimating point process parameters, in particular the second-order product moment function, from equally spaced observations on the counting process. The approach followed in this thesis does have the advantage of allowing also for those discrete-valued time series which do not arise as counts of a point process, e.g. categorical series like example (ix) above.

The plan of the thesis is as follows. The rest of this chapter will survey the models that are available for discrete-valued series, stressing aspects of the models like marginal distributions, correlation structure and (where these have been discussed in the literature) parameter estimation techniques. We begin by discussing Markov chains and higher-order Markov chains, in particular the higher-order Markov chain models introduced by Pegram (1980) and generalized by Raftery (1985a). We then summarize (in sections 1.4–1.6) the work on three classes of models that may be described as attempts to provide for discrete-valued time series a broad class of models analogous to the familiar Gaussian ARMA models: models based on mixtures, models based on the idea of thinning a discrete random variable, and certain other bivariate geometric models of (loosely) autoregressive moving average structure. Markov regression models, an extension to the time series context of the ideas of generalized linear models, are discussed next. Section 1.8 presents some relevant examples of parameter-driven processes, i.e. processes in which there is an underlying and unobserved 'parameter process' which determines the distribution of a series of observations. In section 1.9 two

contrasting families of state-space models, those of Harvey and of West and Harrison, are described. After surveying further miscellaneous models for discrete-valued series, the chapter ends with a brief account of what has so far been achieved by the various different kinds of model.

Chapter 2 begins with a detailed review of certain results on hidden Markov models which are available in the speech-processing literature. Such models, which are examples of parameter-driven processes, have for some time been used in speech-recognition applications, but have only recently been considered as general-purpose models for discrete-valued time series. After this review we introduce the hidden Markov time series models to which the rest of the thesis is devoted. These models are based on an unobserved stationary Markov chain and either a Poisson or a binomial distribution. Correlation properties are derived, an algorithm for evaluating the likelihood function is described, and direct numerical maximization of the likelihood is proposed as the most practical means of parameter estimation. Marginal, joint and conditional distributions are derived and some applications of these results indicated, e.g. to forecasting and to the treatment of missing data. Reversibility of the observed process is shown to be implied by, but not equivalent to, reversibility of the underlying Markov chain. Finally some remarks are made on the way in which such processes may be used as statistical models.

In Chapter 3 these hidden Markov time series models are extended and modified in various useful ways. In one modification the underlying Markov chain is replaced by a higher-order Markov chain, and (inter alia) an algorithm for computing the likelihood in this case is derived. In another modification the models based on the binomial distribution are generalized by replacing that distribution by the multinomial. This yields a model for

categorical time series. More general multivariate models of several kinds are then discussed, and results on the likelihood function and cross-correlations of such models are derived. Two different methods of incorporating trend or seasonality or dependence on covariates other than time are also introduced in this chapter. One final variation discussed modifies the 'hidden Markov-binomial' model by assuming that the number of trials at each stage is not a known constant but is supplied instead by some further random process.

Chapter 4 presents examples of applications of the models of Chapters 2 and 3 to data of various types and from a variety of subjects, and makes comparisons with competing models such as Markov chains of order one or higher.

Some remarks on the notation to be used throughout may be helpful at this stage. With few exceptions, the model for the observations will be denoted by $\{S_t\}$. Unless otherwise indicated, vectors are row vectors. Transposition of matrices is denoted by the symbol $'$.

1.2 Saturated Markov chains as time series models

Since the Markov property is a simple, and mathematically tractable, relaxation of the assumption of independence, it is natural to consider discrete-time Markov chains on a finite state-space as possible models for time series taking values in that space. Although some of the models considered in later sections of this chapter are Markov chains with a specific structure, we confine our attention here to fully parametrized, or saturated, Markov chain models, by which is meant Markov chains which have $m^2 - m$ independent transition probabilities when m is the number of states. The states are as-

sumed to be either quantitative or ordered categories. We review here the following aspects of such Markov chains, on state-space $\{1, 2, \ldots, m\}$: the autocorrelation function (ACF), the partial autocorrelation function (PACF), and estimation of the transition probabilities by maximum likelihood. We do so in some detail because of the close links to the 'hidden Markov' models to be introduced in Chapter 2. Unless it is otherwise noted, Markov chain terminology is taken from Grimmett and Stirzaker (1982). Apart from some observations concerning the partial autocorrelations, the results of this section are not new, but some appear not to be readily accessible in the literature. Billingsley (1961) and Chapter 4 of Basawa and Prakasa Rao (1980) present extensive accounts of statistical methods in finite state-space Markov chains.

Let $\{S_t : t \in \mathbf{N}\}$, then, be an irreducible homogeneous Markov chain on the first m positive integers, with transition probability matrix Γ. That is, $\Gamma = (\gamma_{ij})$, where for all states i and j and times t:

$$\gamma_{ij} = \mathrm{P}(S_t = j \mid S_{t-1} = i).$$

(In this work, unless it is otherwise indicated, Markov chains are assumed to be homogeneous.) By the irreducibility, there exists a unique, strictly positive, stationary distribution, which we shall denote by the vector $\delta = (\delta_1 \ \delta_2 \ \ldots \ \delta_m)$. Suppose that $\{S_t\}$ is stationary, so that δ is for all t the distribution of S_t.

The irreducibility of $\{S_t\}$ implies also that 1 is a *simple* eigenvalue of Γ and the corresponding right and left eigenvectors are unique up to constant multiples (Seneta, 1981, Theorem 1.5). It then follows that such eigenvectors are multiples of the column vector $\mathbf{1}' = (1 \ 1 \ \ldots \ 1)'$ and δ respectively.

If we define $v = (1\ 2\ \ldots\ m)$, $V = \mathrm{diag}(v)$ (i.e. the diagonal matrix with v on the principal diagonal), and $\gamma_{ij}(k) = (\Gamma^k)_{ij}$, we have the following results for the mean of S_t and the covariance of S_t and S_{t+k}, for $k \in \mathbf{N}_0$, the set of all nonnegative integers:

$$\begin{aligned}
\mathrm{E}(S_t) &= \sum_{i=1}^{m} i\delta_i \\
&= \delta v'; \\
\mathrm{E}(S_t S_{t+k}) &= \sum_{i=1}^{m}\sum_{j=1}^{m} ij\delta_i \mathrm{P}(S_{t+k}=j \mid S_t=i) \\
&= \sum_{i,j} (i\delta_i)\gamma_{ij}(k)j \\
&= \delta V \Gamma^k v'; \\
\mathrm{Cov}(S_t, S_{t+k}) &= \delta V \Gamma^k v' - (\delta v')^2.
\end{aligned}$$

Even if Γ is not diagonalizable, some simplification of this expression for the covariance may be achieved by writing Γ in Jordan canonical form. Details of the Jordan canonical form may be found, for instance, in section 11.6 of Noble (1969) or on pp. 121–122 of Cox and Miller (1965), but for our purpose it will be sufficient to note that Γ may be written as $U\Omega U^{-1}$, where U, U^{-1} and Ω are of the following forms: $U = (\mathbf{1}'\ R)$, $U^{-1} = \begin{pmatrix} \delta \\ W \end{pmatrix}$ and $\Omega = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0}' & \Psi \end{pmatrix}$. (The matrix Ψ is band-diagonal, with the eigenvalues of Γ other than 1 on the diagonal, ones or zeroes on the superdiagonal, and zeroes elsewhere.) Hence

$$\Gamma^k = U\Omega^k U^{-1} = \mathbf{1}'\delta + R\Psi^k W,$$

and

$$\mathrm{Cov}(S_t, S_{t+k}) = \delta V(\mathbf{1}'\delta + R\Psi^k W)v' - (\delta v')^2 = (\delta V R)\Psi^k(Wv').$$

Since this is true for $k = 0$, we have also the variance of S_t, and hence the

ACF of $\{S_t\}$:

$$\rho_k = \frac{\text{Cov}(S_t, S_{t+k})}{\text{Var } S_t} = \frac{(\delta VR)\Psi^k(Wv')}{\delta VRWv'}.$$

The resulting expression involves powers up to the kth of the eigenvalues of Γ.

If Γ is diagonalizable, a much neater structure emerges. If the eigenvalues (other than 1) of Γ are denoted by $\omega_2, \omega_3, \ldots, \omega_m$, Ω can be taken to be diag(1 ω_2 ... ω_m), the columns of U are corresponding right eigenvectors of Γ, and the rows of U^{-1} corresponding left eigenvectors. We then have, for $k \in \mathbf{N}_0$:

$$\begin{aligned} \text{Cov}(S_t, S_{t+k}) &= \delta VU\Omega^k U^{-1}v' - (\delta v')^2 \\ &= a\Omega^k b' - a_1 b_1 \\ &= \sum_{i=2}^{m} a_i b_i \omega_i^k, \end{aligned}$$

where $a = \delta VU$ and $b' = U^{-1}v'$. Hence $\text{Var}(S_t) = \sum_{i=2}^{m} a_i b_i$, and for $k \in \mathbf{N}_0$

$$\rho_k = \text{Corr}(S_t, S_{t+k}) = \sum_{2}^{m} a_i b_i \omega_i^k / \sum_{2}^{m} a_i b_i.$$

This is a linear combination of the kth powers of the eigenvalues $\omega_2, \ldots, \omega_m$, and (if these eigenvalues are distinct) somewhat similar to the ACF of a Gaussian autoregressive process of order $m - 1$. As will shortly be seen, however, the analogy breaks down when one considers the PACF. We note in passing that $\rho_k = \rho_1^k$ for $k \in \mathbf{N}_0$ in the case $m = 2$, and also in certain other cases, e.g. if all the eigenvalues ω_i are equal, or if $a_i b_i = 0$ for all but one value of i. For m equal to 2, ρ_1 is just the eigenvalue of Γ other than 1.

Before considering the PACF of a Markov chain, we note first certain general facts concerning partial autocorrelations. For any second-order stationary process $\{S_t\}$, the partial autocorrelation of order k, denoted by ϕ_{kk}, is the correlation of the residuals obtained by regressing S_t and S_{t+k} on the

intervening $k-1$ variables. As Lawrance (1976) demonstrates, this is by no means always the same as the conditional correlation of S_t and S_{t+k} given the intervening variables, although some authors, e.g. Kendall and Stuart (1979, section 27.1), do not make this at all clear. In order to derive the partial autocorrelations from the autocorrelations ρ_k we may use the standard relation (Brockwell and Davis, 1987, p. 102):

$$\phi_{kk} = |P_k^*|/|P_k|, \tag{1.1}$$

where

$$P_k = \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{k-1} \\ \rho_1 & 1 & \dots & \rho_{k-2} \\ \vdots & \vdots & & \vdots \\ \rho_{k-1} & \rho_{k-2} & \dots & 1 \end{pmatrix}$$

and P_k^* is identical to P_k except that the last column of P_k is replaced by the vector $(\ \rho_1 \ \ \rho_2 \ \ \dots \ \ \rho_k\ )'$. It should be noted that equation (1.1) holds for any second-order stationary process: see equation (23.4.2) of Cramér (1946), which implies the above equation. In particular, equation (1.1) does not merely hold for processes with a specific distribution or structure such as Gaussian autoregressive or moving average processes. As Brockwell and Davis remark after their definition 3.4.2, this property provides an alternative definition of the partial autocorrelations.

Now consider the m-state Markov chain $\{S_t\}$ as defined above. By using equation (1.1) it can be verified that for $m = 2$ (and any other case in which $\rho_k = \rho_1^k$ for all $k \in \mathbf{N}$) ϕ_{rr} is zero for all r exceeding 1 — as is true also of the standard (Gaussian) AR(1) process. The example below shows, however, that for a three-state Markov chain ϕ_{33} may be nonzero. This seems to contradict a statement in the first paragraph of Pegram (1980), and is of course quite different behaviour from that of the standard AR(2) process, for which

$\phi_{rr} = 0$ for $r \geq 3$.

Example Consider the stationary Markov chain on $\{1, 2, 3\}$ with transition probability matrix

$$\Gamma = \begin{pmatrix} 0 & 1 & 0 \\ \frac{13}{16} & 0 & \frac{3}{16} \\ 1 & 0 & 0 \end{pmatrix}.$$

The ACF is $\rho_k = \frac{75}{124}(-\frac{3}{4})^k + \frac{49}{124}(-\frac{1}{4})^k$, and the first three autocorrelations are $\rho_1 = -137/248$, $\rho_2 = 181/496$, and $\rho_3 = -1037/3968$. Hence $|P_3^*| = -0.02055$, and ϕ_{33} is nonzero. □

In order to estimate the $m^2 - m$ parameters γ_{ij} $(i \neq j)$ of the Markov chain $\{S_t\}$ from a realization $s_1, s_2, \ldots, s_T$, we consider first the likelihood conditioned on the first observation. This is

$$\prod_{i=1}^{m} \prod_{j=1}^{m} \gamma_{ij}^{f_{ij}},$$

where f_{ij} is the number of transitions from state i to state j (and hence $\sum_{i,j} f_{ij} = T - 1$). Since $\gamma_{ii} = 1 - \sum_{k \neq i} \gamma_{ik}$, differentiating the logarithm of this likelihood with respect to γ_{ij} and equating the derivative to zero yields

$$\frac{f_{ii}}{1 - \sum_{k \neq i} \gamma_{ik}} = \frac{f_{ij}}{\gamma_{ij}}.$$

The intuitively plausible estimator $\hat{\gamma}_{ij} = f_{ij} / \sum_{k=1}^{m} f_{ik}$ may thereby be seen to be a conditional maximum likelihood estimator of γ_{ij}. (Note that the assumption of stationarity of the Markov chain was not actually used in the above derivation.) The unconditional likelihood of a stationary Markov chain $\{S_t\}$ is the conditional likelihood as above, multiplied by δ_{s_1}, and it or its logarithm may be maximized numerically, subject to nonnegativity and row-sum constraints, in order to estimate the transition probabilities γ_{ij}. In some nontrivial special cases of the two-state Markov chain explicit results

are available for unconditional maximum likelihood estimators: see Bisgaard and Travis (1991).

Even for an observed series which appears to satisfy the Markov property, a general Markov chain on m states, with its $m^2 - m$ parameters, may not be a very useful model, however, because of the large number of parameters involved. Other, more parsimonious, models which can also be easily fitted to data are therefore needed.

1.3 Higher-order Markov chains

In cases where the observations on a finite state-space process appear not to satisfy the Markov property, one possibility that suggests itself is to fit an lth-order Markov chain, i.e. a model $\{S_t\}$ satisfying the following generalization of the Markov property for some $l \geq 2$:

$$\mathrm{P}(S_t \mid S_{t-1}, S_{t-2}, \ldots) = \mathrm{P}(S_t \mid S_{t-1}, \ldots, S_{t-l}).$$

An account of such higher-order Markov chains may be found, for instance, in Lloyd (1980), section 19.9. Although such a model is not in the usual sense a Markov chain, i.e. not a 'first-order' Markov chain, we can redefine the model in such a way as to produce an equivalent process which is. If we let $X_t = (S_{t-l+1}, S_{t-l+2}, \ldots, S_t)$, then $\{X_t\}$ is a (first-order) Markov chain on M^l, where M is the state-space of $\{S_t\}$. Although some properties are more awkward to establish, no essentially new theory is therefore involved in analyzing an lth-order Markov chain rather than a first-order one. For instance, a stationary distribution for $\{S_t\}$, if it exists, may be found by determining the stationary distribution of the Markov chain $\{X_t\}$ and deducing from it the implied marginal distribution for any one of the l components of X_t.

Of course the use of a general higher-order Markov chain (instead of a first-order one) greatly increases the problem of overparametrization: an lth-order Markov chain on m states has $m^l(m-1)$ independent transition probabilities. (Although $\{X_t\}$ is a Markov chain on m^l states, the number of independent transition probabilities is less than $m^{2l}-m^l$ because many of the entries in its transition probability matrix are identically zero.) Pegram (1980) and Raftery (1985a, 1985b) have therefore proposed certain classes of parsimonious models for higher-order chains which are far more suitable for practical application. For $m = 2$ the models of Raftery are equivalent to those of Pegram, but for $m > 2$ those of Raftery are more general. Pegram's models have $m + l - 1$ parameters, and those of Raftery $m(m-1) + l - 1$, where m and l have the same meanings as above. This appears to contradict a remark of Li and Kwok (1990), who cite Pegram's work but state that Raftery's proposal 'results in a model that is much more parsimonious than all previously proposed'. If by parsimony one means fewer parameters, clearly Pegram's are more parsimonious. Raftery's models (for $m > 2$) can however represent a wider range of dependence patterns and autocorrelation structures. In both cases an increase of one in the order of the Markov chain requires only one additional parameter.

Raftery's models, which are an example of what have been termed 'linear conditional probability models' (Martin and Raftery, 1987), are defined as follows. The process $\{S_t\}$ takes values in $M = \{1, 2, \ldots, m\}$ and satisfies

$$\mathrm{P}(S_t = j_0 \mid S_{t-1} = j_1, \ldots, S_{t-l} = j_l) = \sum_{i=1}^{l} \lambda_i q_{j_i j_0}, \tag{1.2}$$

where $\sum_{i=1}^{l} \lambda_i = 1$, and $Q = (q_{jk})$ is an $m \times m$ matrix with nonnegative entries and row sums equal to one, such that the right-hand side of equation (1.2) is bounded by zero and one for all $j_0, j_1, \ldots, j_l \in M$. This last requirement, which generates m^{l+1} pairs of constraints nonlinear in the parameters,

ensures that the conditional probabilities in equation (1.2) are indeed probabilities, and the condition on the row sums of Q ensures that the sum of these probabilities over j_0 is one. Note that Raftery does not assume that the parameters λ_i are nonnegative. If, however, one does make that assumption, the right-hand side of equation (1.2) is a convex linear combination of probabilities and thereby automatically bounded by 0 and 1, which renders the nonlinear constraints redundant. The assumption that the parameters λ_i are nonnegative greatly simplifies the computations involved in parameter estimation, and in practice it seems to be necessary to make that assumption unless m and l are very small.

In the case $l=1$ this model is a first-order Markov chain with transition probability matrix Q. Another interesting case of the model (1.2) may be obtained by taking

$$Q = \theta I + (1-\theta)\mathbf{1}'\pi,$$

where π is a positive vector with $\sum_{j=1}^{m} \pi_j = 1$. This yields the models of Pegram. The vector π is then the stationary distribution corresponding to the t.p.m. Q (although in Pegram's treatment π is actually defined as the limiting distribution, as $t \to \infty$, of S_t).

Raftery proves the following limit theorem, on the assumption that the elements of Q are positive: if π is the stationary distribution corresponding to the t.p.m. Q, $S_t = j$ has limiting probability π_j independent of the initial conditions. That is,

$$\lim_{t\to\infty} \mathrm{P}(S_t = j \mid S_1 = i_1, \ldots, S_l = i_l) = \pi_j$$

for all $j, i_1, \ldots, i_l \in M$. It is therefore reasonable to restrict one's consideration to stationary models $\{S_t\}$. For such stationary models, Raftery

shows that the joint distribution of S_t and S_{t+k} satisfies a system of linear equations similar to the Yule-Walker equations. More precisely, if we define $P(k) = (p_{ij}(k))$ by

$$p_{ij}(k) = \mathrm{P}(S_t = i, S_{t+k} = j) \quad i, j, \in M; k \in \mathbf{Z} \ ,$$

with $P(0) = \mathrm{diag}(\pi)$ being the case $k=0$, we have for $k \in \mathbf{N}$

$$P(k) = \sum_{g=1}^{l} \lambda_g P(k-g)Q. \tag{1.3}$$

It is not always possible to solve these equations uniquely, but Raftery gives (separately for $l = 2, 3$ and ≥ 4) sufficient conditions on the parameters λ_i and q_{jk} for uniqueness. He derives from the equations (1.3) a system of equations for the autocorrelations $\rho_k = \mathrm{Corr}(S_t, S_{t+k})$ which resembles the Yule-Walker equations to some extent and may be solved uniquely in certain special cases only. He considers in detail the autocorrelation behaviour of his model when $m = 3$, $l = 2$, $\pi = \frac{1}{3}\mathbf{1}$ and Q has special structure implying that the autocorrelations do satisfy precisely a set of Yule-Walker equations, so that

$$\rho_1 = \phi_1 + \phi_2 \rho_1$$
$$\rho_2 = \phi_1 \rho_1 + \phi_2$$

for certain quantities ϕ_i not depending on ρ_1 or ρ_2. It emerges that the set of correlations (ρ_1, ρ_2) allowed by this model, although quite complicated in shape, is contained in and not much smaller than the corresponding set for the usual Gaussian AR(2) models.

Raftery (1985a) fits his models to three data sets, relating to wind power, interpersonal relationships, and occupational mobility. Parameters are estimated by direct numerical maximization of the logarithm of the conditional likelihood, i.e. conditional on the first l observations if l is (as above) the order. Model order selection and comparisons with competing models are done

on the basis of Schwarz's 'Bayesian information criterion' BIC (Schwarz, 1978). In all three applications the models of Raftery appear to combine parsimony with fidelity to data more successfully than do the alternatives available.

Subsequently Adke and Deshmukh (1988) have generalized Raftery's limit theorem to linear conditional probability models for higher-order Markov chains on a countable (possibly infinite) state-space. In fact they extend the result even further, to similar higher-order Markov models on arbitrary state-space. They do, however, state an assumption not made by Raftery: they assume in general that the coefficients λ_i are positive.

Li and Kwok (1990) discuss the merits of various estimation techniques applicable to Raftery's models in several different sampling situations. For a single realization, Raftery's conditional maximum likelihood estimation procedure and a minimum chi-squared procedure are discussed, and compared by means of a simulation experiment involving models in which $m = 3$, $l = 2$ and $\pi = \frac{1}{3}\mathbf{1}$. The two methods produce comparable estimates of Q, but the minimum chi-squared method appears to produce better estimates of λ_1. For 'macro' data, i.e. data which aggregate the observations on several (presumably independent) copies of a Raftery model, a nonlinear least squares method is proposed and investigated by simulation, but the results suggest that the estimators may be inconsistent in some circumstances. Finally Li and Kwok consider unaggregated data from a 'panel' of subjects assumed to follow Raftery models, but with the parameters possibly differing between subjects. On the evidence provided by another simulation experiment Li and Kwok recommend an empirical Bayes technique for the estimation of the parameters for each subject.

Azzalini (1983), in his investigation of maximum likelihood estimation 'of order m', uses a binary second-order Markov chain in order to study the efficiency of such estimation (of order zero or one) in a case where it can be compared with 'exact' maximum likelihood. His conclusion is that the method works well enough to be worth considering in cases where the maximum likelihood estimator is not available. Azzalini and Bowman (1990) report the fitting of a second-order Markov chain model to the binary series they use to represent the lengths of successive eruptions of the 'Old Faithful' geyser. Their analysis, and some alternative models, will be dicsussed in some detail in section 4.2.

Mehran (1989) proposes a generalization of Raftery's models which he terms an 'infinite-lag Markov model'. It allows for an infinite number of terms $\lambda_i q_{j_i j_0}$ rather than the l terms appearing in equation (1.2), with $\sum \lambda_i = 1$ as before. In this case, however, the coefficients λ_i are given by some simple decreasing parametric function of i, e.g. $\lambda_i = \varepsilon^{i-1}(1-\varepsilon)$ for some $\varepsilon \in (0,1)$. Such models, it is claimed, are particularly useful in circumstances of missing data items or nonconsecutive sampling schemes. A finite sequence is merely treated as an infinite sequence in which all the observations after a certain point are missing. Mehran describes an application to labour statistics collected according to a specific nonconsecutive sampling scheme, the 4-8-4 rotation sampling scheme of the U.S. Current Population Survey.

1.4 Models based on mixtures: the DARMA models of Jacobs and Lewis

The earliest attempt to provide a class of discrete-valued time series models analogous to the familiar Gaussian ARMA models appears to be that of

Jacobs and Lewis (1978a, 1978b, 1978c, 1983). Their two classes of discrete autoregressive moving average models, abbreviated DARMA and NDARMA, are formed by taking probabilistic mixtures of i.i.d. (independent and identically distributed) discrete random variables having the required marginal distribution. Applications have been published by Buishand (1978), Chang, Kavvas and Delleur (1984a, 1984b), and Chang, Delleur and Kavvas (1987).

In order to specify the models, we need first the following definitions. Let $\{Y_t\}$ be a sequence of i.i.d. random variables on some countable subset E of the real line, with $\mathrm{P}(Y_t = i) = \pi(i)$ for all $i \in E$. Let $\{U_t\}$ and $\{V_t\}$ be independent sequences of i.i.d. binary random variables with $\mathrm{P}(U_t = 1) = \beta \in [0,1]$ and $\mathrm{P}(V_t = 1) = \rho \in [0,1)$. Let $\{D_t\}$ and $\{A_t\}$ be sequences of i.i.d. random variables with

$$\begin{aligned} \mathrm{P}(D_t = n) &= \delta_n \quad n = 0, 1, \ldots, N \\ \mathrm{P}(A_t = n) &= \alpha_n \quad n = 1, 2, \ldots, p, \end{aligned}$$

where $N \in \mathbf{N}_0$ and $p \in \mathbf{N}$.

The DARMA(p,N+1) process $\{S_t\}$ is then formed as follows:

$$\begin{aligned} S_t &= U_t Y_{t-D_t} + (1 - U_t) Z_{t-(N+1)} \quad t = 1, 2, \ldots, \\ Z_t &= V_t Z_{t-A_t} + (1 - V_t) Y_t \qquad\quad t = -N, -N+1, \ldots . \end{aligned}$$

The process $\{Z_t\}$ is the DAR(p) process, i.e. the discrete autoregressive process of order p, as defined by Jacobs and Lewis (1978c). This process may be described informally as follows: with probability ρ, Z_t is one of the p previous values $Z_{t-1}, \ldots, Z_{t-p}$, and otherwise it is Y_t. The DARMA process $\{S_t\}$ may be described similarly: with probability β, S_t is one of the values $Y_t, \ldots, Y_{t-N}$, and otherwise it equals the 'autoregressive tail' $Z_{t-(N+1)}$. Clearly the above definition of the DARMA process requires specification of the joint distribution of the p-dimensional random vector $(Z_{-N-p}, \ldots, Z_{-N-1})$. In Jacobs

and Lewis (1978c) it is shown that there is a stationary distribution for that random vector. If the DAR(p) process $\{Z_t\}$ is started with that stationary distribution, the DARMA process $\{S_t : t \in \mathbf{N}\}$ is then a stationary process with marginal distribution π. We shall henceforth assume that $\{S_t\}$ is indeed stationary.

The cases $\beta = 0$ and $\beta = 1$ of the above general DARMA(p,$N+1$) model are of course simpler in structure than the general model. If $\beta = 0$, we have $S_t = Z_{t-(N+1)}$, i.e. the DAR(p) model already described. If $\beta = 1$, we have $S_t = Y_{t-D_t}$, which is the model termed DMA(N) (not DMA($N+1$)) by Jacobs and Lewis (1978a). In this case, S_t is a probabilistic mixture of the $N+1$ i.i.d. random variables $Y_t, Y_{t-1}, \ldots, Y_{t-N}$.

Jacobs and Lewis (1983) derive equations which enable one to find the ACF

$$\rho_k = \mathrm{Corr}(S_t, S_{t+k})$$

of a general DARMA(p,$N+1$) process. The ACF of $\{Z_t\}$, the DAR(p) process, satisfies equations of the same form as the Yule-Walker equations for a Gaussian AR(p) process. The ACF of the DMA(N) process is given by:

$$\rho_k = \begin{cases} \sum_{j=0}^{N-k} \delta_j \delta_{j+k} & k = 1, 2, \ldots, N \\ 0 & k = N+1, \ldots . \end{cases}$$

It should be noted that the autocorrelations of any DARMA process are all nonnegative, and do not depend in any way on the marginal distribution π. The marginal distribution itself is completely general, but clearly particular distributions such as the Poisson will be of most interest.

As Jacobs and Lewis (1983) remark, the models of Pegram (1980) are a generalization of finite state-space DAR(p) processes, and unlike the latter

they do allow some negative correlation. How strong this negative correlation may be depends on the stationary distribution of the process.

The second class of discrete ARMA models introduced by Jacobs and Lewis (1983) is the class of NDARMA processes ('new' discrete ARMA, presumably). Let $\{Y_t\}$, $\{V_t\}$, $\{D_t\}$ and $\{A_t\}$ be as before. The NDARMA(p,N) process is defined as $\{S'_t\}$, where

$$S'_t = V_t S'_{t-A_t} + (1 - V_t)Y_{t-D_t}.$$

Hence S'_t is, with probability ρ, one of the p previous values $S'_{t-1}, S'_{t-2}, \ldots,$ S'_{t-p}. With probability $1-\rho$, it is one of the $N+1$ quantities $Y_t, Y_{t-1}, \ldots, Y_{t-N}$. As is true also for the DARMA models, special cases yield the DAR and DMA processes. To be specific, the case $\rho = 0$ yields the DMA(N) model, and the case $\delta_0 = \mathrm{P}(D_t = 0) = 1$ yields the DAR(p) model.

Jacobs and Lewis show that the NDARMA models have a stationary distribution with marginal distribution π. They derive equations from which the autocorrelations of any stationary NDARMA model may be determined, and note that these too are necessarily nonnegative. As in the case of DARMA models, the correlation structure does not depend on the marginal distribution π, and that marginal distribution is quite general.

Some specific examples of stationary DARMA and NDARMA models are now presented as illustration.

DAR(1) As noted above, $\{Z_t\}$ is a DAR(p) process. The DAR(1) process is particularly simple, and satisfies

$$Z_t = V_t Z_{t-1} + (1 - V_t)Y_t.$$

That is,

$$Z_t = \begin{cases} Z_{t-1} & \text{with probability } \rho \\ Y_t & \text{with probability } 1-\rho. \end{cases}$$

It is therefore a (stationary) Markov chain, with stationary distribution π. The transition probabilities are given by

$$\mathrm{P}(Z_t = i \mid Z_{t-1} = k) = \rho\delta_{ki} + (1-\rho)\pi_i,$$

where δ_{ki} is the Kronecker delta symbol. The ACF is simply ρ^k, for all $k \in \mathbf{N}$.

DMA(1) The DMA(1) model $\{S_t\}$ satisfies

$$S_t = \begin{cases} Y_t & \text{with probability } \delta_0 \\ Y_{t-1} & \text{with probability } \delta_1 = 1-\delta_0. \end{cases}$$

The ACF is

$$\rho_k = \begin{cases} \delta_0(1-\delta_0) & k = 1 \\ 0 & k = 2, 3, \dots, \end{cases}$$

and it is shown by Jacobs and Lewis (1978a) that $\{S_t\}$ is reversible.

DARMA(1,1) and NDARMA(1,1) The DARMA(1,1) model satisfies

$$S_t = \begin{cases} Y_t & \text{with probability } \beta \\ Z_{t-1} & 1-\beta, \end{cases}$$

where

$$Z_{t-1} = \begin{cases} Z_{t-2} & \text{with probability } \rho \\ Y_{t-1} & 1-\rho. \end{cases}$$

The NDARMA(1,1) model $\{S'_t\}$ has

$$S'_t = \begin{cases} S'_{t-1} & \text{with probability } \rho \\ Y_t & \delta_0(1-\rho) \\ Y_{t-1} & (1-\delta_0)(1-\rho). \end{cases}$$

The autocorrelation functions are:

$$\rho_k = \mathrm{Corr}(S_t, S_{t+k}) = \rho^{k-1}(1-\beta)\{\beta(1-\rho) + (1-\beta)\rho\}$$
$$\rho'_k = \mathrm{Corr}(S'_t, S'_{t+k}) = \rho^{k-1}\{\rho + (1-\rho)^2\delta_0(1-\delta_0)\}.$$

Jacobs and Lewis (1983) present graphs of the possible pairs (ρ_1, ρ_2) and (ρ'_1, ρ'_2), from which it is apparent that, although neither set of attainable correlations contains the other, that of the NDARMA(1,1) process is smaller. By comparison with Figure 3.10(b) of Box and Jenkins (1976), we see that the standard Gaussian ARMA(1,1) model has a much larger set of attainable correlations, mainly because negative correlations are possible.

Jacobs and Lewis (1983) consider the estimation of the autocorrelations of DARMA and NDARMA processes, in particular ρ in the case of a DAR(1) process. On the basis of a simulation study they conclude that the usual sample autocorrelation performs worst among the eight alternatives they present. An ad hoc estimator based on properties specific to DARMA and NDARMA models is found to perform well. Many other properties (e.g. the fact that the DARMA(1,$N+1$) process is ϕ-mixing) and some possible extensions (e.g. to negative correlations and bivariate processes) are described in the papers by Jacobs and Lewis already cited. We conclude this section instead by describing briefly the applications to hydrological problems that have been reported.

Jacobs and Lewis (1983) report the work of Buishand (1978), who used a binary DARMA(1,1) process as a model for sequences of wet and dry days. Chang et al. (1984a, 1984b, 1987) have applied various DARMA models to sequences of wet and dry days during some 'locally stationary season', to a three-state discretization of daily precipitation amounts, and to estimation of daily runoff. Estimation of the distribution π was done by utilizing the observed runlengths. Chang et al. state in all three applications that the other parameters were estimated by fitting the theoretical ACF to the sample

ACF, by nonlinear least squares, but presumably this refers only to the first few terms of the ACF.

1.5 Models based on thinning

A fairly broad class of models is that based on the idea of 'thinning'. Such models have been discussed by McKenzie (1985a, 1985b, 1986, 1987, 1988a, 1988b), Al-Osh and Alzaid (1987, 1988, 1991), Alzaid and Al-Osh (1988, 1990), and Du and Li (1991). Although the properties of the models have been studied extensively, it seems that no applications have yet been published. We introduce the models by considering first the case of those with geometric marginal distribution, and in particular the geometric AR(1) process of McKenzie.

1.5.1 Models with geometric marginal

Let the thinning operation '$*$' (also known as binomial thinning) be defined as follows. If X is any nonnegative integer-valued random variable and $0 \leq \alpha \leq 1$, $\alpha * X$ is defined as $\sum_{i=1}^{X} B_i$, where $\{B_i\}$ is a sequence of i.i.d. binary random variables, independent of X, with $\mathrm{P}(B_i = 1) = \alpha$. Conditional on X, therefore, $\alpha * X$ is distributed binomially with parameters X and α. Now suppose that

$$S_t = \alpha * S_{t-1} + R_t, \tag{1.4}$$

where the innovation R_t is independent of S_{t-1}, $0 \leq \alpha \leq 1$ and S_{t-1} has the geometric distribution with mean $\lambda^{-1} = \theta/(1-\theta)$, i.e. for nonnegative integer k:

$$\mathrm{P}(S_{t-1} = k) = (1-\theta)\theta^k. \tag{1.5}$$

Then S_{t-1} has the 'alternate probability-generating function' (a.p.g.f.)

$$\mathrm{E}((1-z)^{S_{t-1}}) = \lambda/(\lambda + z),$$

$\alpha * S_{t-1}$ has a.p.g.f. $\lambda/(\lambda + \alpha z)$, and the condition for S_t to have the same distribution as S_{t-1} is that R_t have a.p.g.f.

$$\frac{\lambda/(\lambda + z)}{\lambda/(\lambda + \alpha z)} = \alpha + (1 - \alpha)\frac{\lambda}{\lambda + z}.$$

That is, R_t either is zero (with probability α) or has the geometric distribution (1.5). Equivalently, R_t can be described as the product of (independent) geometric and binary random variables. If R_t satisfies this requirement and S_0 has the distribution (1.5), then S_t will also have that geometric distribution for all nonnegative integers t.

The correlation structure of the stationary sequence $\{S_t\}$ is simple: it can be shown that ρ_k, the autocorrelation of order k, is just α^k, as is the case for the usual continuous-valued AR(1) model

$$S_t = \alpha S_{t-1} + R_t. \tag{1.6}$$

Any random process $\{S_t\}$ satisfying equation (1.4) is a Markov chain, and the transition probabilities for this case are given explicitly in equation (2.8) of McKenzie (1986). The thinning operation '$*$' is a very natural discrete analogue of the scalar multiplication appearing in the model (1.6). Furthermore, the model $S_t = \alpha * S_{t-1} + R_t$ has the possibly useful interpretation that S_t consists of the survivors of those present at time $t-1$ (each with survival probability α) plus R_t new entrants entering between $t-1$ and t.

The above model is McKenzie's geometric autoregressive process of order one (McKenzie, 1985b), described also by Alzaid and Al-Osh (1988). It has been modified and generalized in several directions, mainly by McKenzie and Al-Osh and Alzaid, and we present in the rest of this section a summary of these developments. These models are designed to have a given marginal distribution (e.g. Poisson, binomial or negative binomial) and dependence

structure (autoregressive, moving average or ARMA). It needs to be emphasised, however, that the terms autoregressive and moving average are often used rather loosely in this context, merely to indicate some similarity in form to that of the standard (Gaussian) ARMA models. For example, Alzaid and Al-Osh (1990) state that their integer-valued autoregressive process of order p actually has autocorrelation function similar to that of a standard ARMA$(p, p-1)$ model.

McKenzie (1986) introduces moving average and autoregressive-moving average models with geometric marginal, all of which (like the geometric AR(1) model described above) are analogues of the corresponding exponential ARMA models of Lawrance and Lewis (1980), obtained by replacing scalar multiplication by thinning and exponential distributions by geometric. The most general model of this kind described by McKenzie is a geometric ARMA(p, q), but to convey the flavour of these models we present in detail only the geometric ARMA(1,1) process. First let $\{M_t\}$, $\{U_t\}$ and $\{V_t\}$ be independent sequences of i.i.d. random variables, where M_t is geometric with mean λ^{-1}, U_t is binary with $\mathrm{P}(U_t = 0) = \alpha$, and V_t is binary with $\mathrm{P}(V_t = 0) = \beta$. Now suppose that W_0 is geometric with mean λ^{-1}, and that

$$S_t = \beta * M_t + V_t W_{t-1}$$

and

$$W_t = \alpha * W_{t-1} + U_t M_t.$$

(Here and elsewhere we assume that each thinning operation is performed independently of all other aspects of the process under consideration, including other thinning operations.) Both W_t and S_t are then geometric with mean λ^{-1}, and $\{S_t\}$ is the geometric ARMA(1,1) process. The k-th order autocorrelation of $\{S_t\}$ is, for positive integers k:

$$\rho_k = \bar{\beta}(\bar{\alpha}\beta + \alpha\bar{\beta})\alpha^{k-1}.$$

(In general we denote $1-\gamma$ by $\bar{\gamma}$.) Clearly $\{W_t\}$ is a geometric AR(1), and in the special cases $\beta = 0$ and $\alpha = 0$ respectively $\{S_t\}$ is AR(1) and MA(1).

Several other models with geometric marginal have been described. McKenzie (1986) discusses also the geometric analogue of the NEAR(1) process of Lawrance and Lewis (1981), and displays simulations of such a process. It is defined by

$$S_t = (\beta U_t) * S_{t-1} + \{1 - V_t + \bar{\alpha}\beta V_t\} * M_t,$$

where $\{M_t\}$, $\{U_t\}$ and $\{V_t\}$ are independent sequences of i.i.d. random variables, M_t is geometric, U_t is binary with $\mathrm{P}(U_t=1) = \alpha$, and V_t is binary with $\mathrm{P}(V_t=1) = \alpha\beta/(1-\bar{\alpha}\beta)$. The case $\alpha=1$ yields the geometric AR(1) process already described above. McKenzie (1985a) gives the geometric analogue of the NEAR(2) model of Lawrance and Lewis (1985).

1.5.2 Models with negative binomial marginal

The geometric distribution is a special case of the negative binomial, and McKenzie (1986) considers also the construction of general negative binomial AR(1) processes. We shall say the random variable S has the negative binomial distribution with shape parameter β and scale parameter λ if for all nonnegative integers k:

$$\mathrm{P}(S=k) = \binom{\beta+k-1}{k} \left(\frac{\lambda}{1+\lambda}\right)^{\beta} \left(\frac{1}{1+\lambda}\right)^{k}.$$

The parameters β and λ are assumed only to be positive. Note in particular that the shape parameter β is not necessarily an integer. The a.p.g.f. of such a negative binomial distribution is $(\lambda/(\lambda+z))^{\beta}$, the Laplace transform of the gamma density with shape and scale parameters β and λ:

$$\lambda^{\beta} x^{\beta-1} e^{-\lambda x}/\Gamma(\beta) \quad (\text{for } x > 0).$$

This suggests that to define a negative binomial AR(1) process, all one has to do is to replace scalar multiplication by thinning, and gamma distributions by negative binomial, in the gamma AR(1) process of Gaver and Lewis (1980). The model which results is

$$S_t = \alpha * S_{t-1} + R_t,$$

with $0 < \alpha < 1$, S_t having the negative binomial (β, λ) distribution for all t, and the innovation R_t having a.p.g.f.

$$\left(\frac{\lambda + \alpha z}{\lambda + z}\right)^{\beta} = \left\{\alpha + (1-\alpha)\frac{\lambda}{\lambda + z}\right\}^{\beta}.$$

To construct a random variable having this a.p.g.f. for general β (and not merely integer values) requires considerable ingenuity. McKenzie (1987) describes such a construction, based on a shot-noise process, and notes that it is essentially the same as that devised by Lawrance (1982) to solve the corresponding problem for gamma processes.

The complexity of the innovation process led McKenzie, however, to define a different kind of negative binomial process (McKenzie, 1986). This is analogous to the gamma beta AR(1) process described by Lewis (1985), which is a random coefficient autoregression with gamma marginal. The resulting negative binomial AR(1) is defined as follows:

$$S_t = A_t * S_{t-1} + M_t,$$

where A_t has a beta distribution with parameters α and $\beta - \alpha$, M_t has a negative binomial distribution with parameters $\beta - \alpha$ and λ, $0 < \alpha < \beta$, $\lambda > 0$, and A_t, S_{t-1} and M_t are mutually independent. If S_0 has the negative binomial distribution with parameters β and λ, S_t then has that distribution in general, and for all nonnegative integers k

$$\rho_k = (\alpha/\beta)^k.$$

Notice that at time t there are three sources of randomness: the (unobserved) random variables A_t and M_t, and the thinning operation.

Clearly the special case $\beta = 1$ of this negative binomial AR(1) process provides a geometric AR(1) process other than the one described in subsection 1.5.1. The innovation of this new geometric AR(1) is neither geometrically distributed nor the product of a geometric and a binary random variable: it is negative binomial with parameters $1 - \alpha$ and λ, where $0 < \alpha < 1$ and $\lambda > 0$.

1.5.3 Models with Poisson marginal

A Poisson AR(1) process is discussed by McKenzie (1985b, 1988b), Al-Osh and Alzaid (1987), and Alzaid and Al-Osh (1988), who treat it as a special case of their INAR(1) model ('integer-valued autoregressive of order 1'). More general Poisson ARMA processes, and a multiple Poisson AR(1), are also introduced by McKenzie (1988b). Al-Osh and Alzaid (1988) discuss various properties of the Poisson case of their INMA(1) and INMA(q) models, INMA standing for 'integer-valued moving average'. Alzaid and Al-Osh (1990) describe properties of the Poisson INAR(p) process, and relate it to the multiple Poisson AR(1) of McKenzie.

McKenzie's Poisson AR(1) process is simply a stationary Poisson solution $\{S_t\}$ of the equation

$$S_t = \alpha * S_{t-1} + R_t,$$

with the usual assumptions that R_t and S_{t-1} are independent and $\{R_t\}$ is a sequence of i.i.d. random variables. What sets the Poisson AR(1) model apart from (e.g.) the geometric AR(1) is that in the Poisson case the marginal and innovation distributions belong to the same family of distributions. More

precisely, S_t is Poisson with mean λ if and only if R_t is Poisson with mean $(1-\alpha)\lambda$: see Al-Osh and Alzaid (1987), section 3. The role of the Poisson distribution relative to the above equation is therefore very much the same as that of the normal distribution relative to the autoregressive equation $S_t = \alpha S_{t-1} + R_t$.

Other interesting properties of the Poisson AR(1) process are that its autocorrelation function is α^k and that it is a reversible Markov chain with transition probabilities

$$\mathrm{P}(S_t = j \mid S_{t-1} = i) = \sum_{k=0}^{j} \{ \binom{i}{k} \alpha^k \bar{\alpha}^{i-k} \} \{ e^{-\lambda\bar{\alpha}} (\lambda\bar{\alpha})^{j-k} / (j-k)! \}.$$

Here λ is the mean of S_t, and we use the convention that $\binom{i}{k} = 0$ for $k > i$. The regression of S_t on S_{t-1} is linear (and vice versa, by the reversibility), but the variance of S_t given S_{t-1} is not constant with respect to S_{t-1}. This last property is one respect in which the Poisson AR(1) does differ from its Gaussian counterpart. Details may be found in McKenzie (1988b). Al-Osh and Alzaid (1987) describe four techniques for estimating α and the innovation mean in a Poisson AR(1), given a realization $s_0, s_1, \ldots, s_T$. These are Yule-Walker estimation, conditional least squares as proposed by Klimko and Nelson (1978), maximum likelihood conditional on the initial observation, and unconditional maximum likelihood. The first three are compared in a simulation experiment in which the initial observation (s_0) is apparently set equal to the integer part of the process mean. The conclusion drawn is that conditional maximum likelihood performs best of the three, in terms of bias and mean squared error.

The Poisson moving average process of order one, as defined by McKenzie

(1988b) and Al-Osh and Alzaid (1988), is a process $\{S_t\}$ satisfying

$$S_t = Y_t + \beta * Y_{t-1}$$

for $0 \leq \beta \leq 1$ and Y_t a sequence of i.i.d. Poisson random variables. If the mean of Y_t is $\lambda/(1+\beta)$, that of S_t is λ. The autocorrelation ρ_k is zero for $k \geq 2$, and $\rho_1 = \beta/(1+\beta)$. What is notable is that the joint distribution of S_{t-1} and S_t is of the same form as in the Poisson AR(1) case discussed above: for both the joint a.p.g.f. is

$$\mathrm{E}\left((1-u)^{S_{t-1}}(1-v)^{S_t}\right) = \exp\{-\lambda(u+v-\rho_1 uv)\}.$$

The Poisson moving average process of order q is a natural extension of that of order one:

$$S_t = Y_t + \sum_{i=1}^{q} \beta_i * Y_{t-i},$$

where $0 \leq \beta_i \leq 1$ for all i, $\beta = \sum_{i=0}^{q} \beta_i$, and $\{Y_t\}$ is a sequence of i.i.d. Poisson random variables with mean λ/β. (By convention $\beta_0 = 1$.) The distribution of S_t is Poisson with mean λ, as usual, and the ACF is :

$$\rho_k = \begin{cases} \sum_{i=0}^{q-k} \beta_i \beta_{i+k}/\beta & k = 1, 2, \ldots, q \\ 0 & k > q. \end{cases}$$

The Poisson ARMA$(1, q)$ process is $\{S_t\}$, where

$$S_t = Y_{t-q} + \sum_{k=1}^{q} \beta_k * W_{t+1-k}$$

and

$$Y_t = \alpha * Y_{t-1} + W_t.$$

The sequence $\{W_t\}$ is taken to be an i.i.d. sequence of Poisson random variables with mean $\bar{\alpha}\lambda$, and Y_0 is an independent Poisson of mean λ. It follows that $\{Y_t\}$ is Poisson AR(1), mean λ, and $\{S_t\}$ a stationary sequence

of Poisson random variables with mean $(1 + \bar{\alpha}b)\lambda$, where b is defined as $\sum_{k=1}^{q} \beta_k$. If $\alpha = 0$, $\{S_t\}$ is Poisson MA(q); if $\beta_k = 0$ for all k, $\{S_t\}$ is Poisson AR(1). McKenzie (1988b) gives the ACF of the general Poisson ARMA$(1, q)$ process, which has the property that, for $k \geq q$, $\rho_k = \alpha^{k-q}\rho_q$. (Since $\rho_q = (\alpha^q + \bar{\alpha}\sum_{i=1}^{q} \beta_i \alpha^{i-1})/(1 + \bar{\alpha}b)$, there appears to be a minor error in McKenzie's expression for the autocorrelation in the case $k > q$: it seems the first α should be α^q.)

McKenzie also presents results concerning the joint distribution of n consecutive observations from a Poisson ARMA$(1, q)$ process $\{S_t\}$, and uses them to draw conclusions about the reversibility or otherwise of various special cases of the process. The joint distribution of the n consecutive observations is shown to be given by the multivariate Poisson distribution of Teicher (1954). An interesting result is that the directional moments $\mathrm{Cov}(S_t^2, S_{t-k})$ and $\mathrm{Cov}(S_t, S_{t-k}^2)$ are in general equal, although the process may be irreversible. The AR(1), MA(1) and MA(2) processes are in general reversible, but for $q \geq 3$ the MA(q) process may be irreversible.

To define a multiple Poisson AR(1) process McKenzie (1988b) first defines $\alpha * Y$ for Y a random variable taking values in $\mathbf{N}_0$ and α a vector of probabilities whose sum does not exceed one. This is done by specifying that, conditional on Y, $\alpha * Y$ has a multinomial distribution with parameters Y and α. The operation '$*$' thus defined is described as multinomial thinning. For a $p \times p$ matrix $A = (\alpha_1\ \alpha_2\ \ldots\ \alpha_p)$ and a vector $Y = (Y_1, Y_2, \ldots, Y_p)$ we then define

$$A * Y = \sum_{i=1}^{p} \alpha_i * Y_i,$$

each multinomial thinning here being performed independently. The multiple

Poisson AR(1) is defined as a stationary solution $\{X_t\}$ of

$$X_t = A * X_{t-1} + E_t,$$

with $\{E_t\}$ being a sequence of i.i.d. p-dimensional random vectors and each vector X_t consisting of independent Poisson random variables. McKenzie (1988b) presents general results concerning the innovation distribution and correlation structure of such processes, of which the most notable is that the jth component of X_t (as defined above) has ACF $\{\rho_j(k)\}$ satisfying

$$\rho_j(k) = \begin{cases} (A^k)_{jj} & k = 1, 2, \ldots, p \\ \sum_{i=1}^{p} \phi_i \rho_j(k-i) & k \geq p \end{cases}$$

for certain constants $\phi_1, \ldots, \phi_p$. Each component therefore has ARMA$(p, p-1)$ structure, although the orders may obviously be lower for matrices A of a particular structure. This and other aspects McKenzie examines in detail for the case $p = 2$.

Finally McKenzie indicates possible extensions of the Poisson models in the following directions: compound Poisson marginal distributions, negative correlation (which is precluded by the structure of the above models), and allowance for trend or cyclical behaviour.

Alzaid and Al-Osh (1990) describe the Poisson case of their general INAR(p) process (see subsection 1.5.5), and show for $p = 2$ that there is embedded in it a multiple Poisson AR(1) process of the type discussed by McKenzie, i.e. with independence of components. This is the process $\{X_t\}$ defined by the state vector

$$X_t = \begin{pmatrix} S_t \\ \alpha_2 * S_{t-1} \end{pmatrix},$$

where α_2 is the 'coefficient' of S_{t-2} in the defining equation

$$S_t = \alpha_1 * S_{t-1} + \alpha_2 * S_{t-2} + R_t.$$

Because the embedded process is simple in structure, it can then be used to derive properties of $\{S_t\}$ such as the joint distribution of S_{t-1} and S_t and the fact that $\{S_t\}$ is reversible.

1.5.4 Models with binomial marginal

McKenzie (1985b) proposes a binomial AR(1) process $\{S_t\}$ satisfying:

$$S_t = \alpha * S_{t-1} + \beta * (N - S_{t-1}),$$

with S_t distributed binomial (N, θ) for all t, $0 < \alpha < 1$ and, unless this exceeds one, $\beta = \bar{\alpha}\theta/\bar{\theta}$. (A modification is possible in the case $\bar{\alpha}\theta/\bar{\theta} > 1$.) The ACF is given by $\rho_k = (\alpha - \beta)^k$. The usual construction, involving the equation $S_t = \alpha * S_{t-1} + R_t$, is not possible because, unlike the Poisson, negative binomial (and geometric) distributions, the binomial is not 'discrete self-decomposable' (as defined in Steutel and van Harn (1979)).

Al-Osh and Alzaid (1991) construct a rather different class of binomial ARMA models, which we can describe as being based on hypergeometric thinning. To define the models we need the following preliminaries. For a random variable S having a binomial distribution with parameters N and p, let $T(S)$ be a random variable whose distribution conditional on S is given by the hypergeometric distribution with parameters N, s and M:

$$\mathrm{P}(T(S)=k \mid S=s) = \frac{\binom{s}{k}\binom{N-s}{M-k}}{\binom{N}{M}},$$

for all appropriate values of k. It then follows that, if S is binomial (N, p), $T(S)$ is binomial (M, p), and $T(S)$ and $S - T(S)$ are independent.

The binomial AR(1) process of Al-Osh and Alzaid is $\{S_t\}$ defined by

$$S_t = T(S_{t-1}) + R_t,$$

where R_t is independent of S_{t-1} (and $T(S_{t-1})$) and has a binomial distribution with parameters $N - M$ and p. Hence if S_{t-1} is binomial (N, p), so also is S_t. Such a binomial AR(1) process, if stationary, has ACF given by $\rho_k = (M/N)^k$ for all $k \in \mathbf{N}_0$, and is in fact a reversible Markov chain. The regression of S_t on S_{t-1} is linear, but the corresponding conditional variance is not constant.

In very similar fashion the same authors define a stationary MA(q) process $\{S_t\}$ satisfying

$$S_t = R_t + \sum_{i=1}^{q} T_i(R_{t-i}),$$

where $\{R_t\}$ is an i.i.d. sequence of binomial $(N - n, p)$ random variables, the distribution of $T_i(R_{t-i})$, given R_{t-i}, is hypergeometric with parameters $N-n$, R_{t-i} and N_i, $\sum_{i=1}^{q} N_i = N$, and the various hypergeometric thinnings are performed independently. The ACF is:

$$\rho_k = \begin{cases} \frac{1}{N(N-n)} \sum_{i=0}^{q-k} N_i N_{i+k} & k = 1, 2, \ldots, q \\ 0 & k > q \end{cases},$$

N_0 being defined as 1. The process therefore has the usual moving average cut-off property. As is the case with almost all of the models based on thinning, however, the correlations are restricted to being nonnegative.

For the corresponding stationary binomial ARMA(1,q) model $\{S_t\}$ the defining equations are:

$$S_t = Y_{t-q} + \sum_{i=1}^{q} T_i(R_{t+1-i})$$

and

$$Y_t = T(Y_{t-1}) + R_t.$$

This structure is analogous to the construction of McKenzie's Poisson ARMA(1,q) process. The autocorrelation functions of the two processes are very similar too: for $k = 1, 2, \ldots, q$ the explicit expressions for ρ_k are similar in form, and there is in both cases some α such that for $k \geq q$ $\rho_k = \alpha^{k-q}\rho_q$. The joint distribution of S_{t-1} and S_t is easily derived for the binomial ARMA(1,1) process, and turns out to be of the same symmetric form as for the binomial AR(1) process. Hence the regression of S_t on S_{t-1} (and vice versa) is linear here too. The conclusion of Al-Osh and Alzaid (from the symmetry of the joint distribution of S_t and S_{t-1}) that $\{S_t\}$ is reversible in the binomial ARMA(1,1) case does not seem justified, however.

Al-Osh and Alzaid define also a multiple binomial AR(1) process very similar to the multiple Poisson AR(1) of McKenzie. This model shares with that of McKenzie the property that individual components of the model have ARMA($p, p-1$) correlation structure.

1.5.5 Results not based on any explicit distributional assumption

Some of the results quoted above in the context of particular marginal distributions are far more generally valid. For instance, the property $\rho_k = \alpha^k$ of geometric and Poisson AR(1) models is a property of any INAR(1) process, as defined by Al-Osh and Alzaid (1987): see their equation 3.3. Furthermore the estimation techniques developed in that paper, although described in detail for the Poisson case only, apply more generally. Further properties of general INAR(1) processes appear in Alzaid and Al-Osh (1988). Results for general INAR(p) processes appear in Alzaid and Al-Osh (1990) and Du and

Li (1991), and for general INMA(q) processes in Al-Osh and Alzaid (1988). We now discuss these models and results briefly.

The INAR(p) process $\{S_t\}$ defined by Alzaid and Al-Osh satisfies

$$S_t = \sum_{i=1}^{p} \alpha_i * S_{t-i} + R_t,$$

where $\{R_t\}$ is, as before, a sequence of i.i.d. random variables taking values in $\mathbf{N}_0$, $\sum_{i=1}^{p} \alpha_i < 1$, and the conditional distribution, given S_t, of the random vector

$$(\alpha_1 * S_t, \alpha_2 * S_t, \ldots, \alpha_p * S_t)$$

is multinomial with parameters S_t and $\alpha = (\alpha_1, \ldots, \alpha_p)$, independent of the history of the process: independent, that is, of S_{t-k} and all thinnings thereof, for $k > 0$. This particular structure (which, incidentally, is not the same as the definition given by Du and Li for their INAR(p) process) implies that the correlation structure is similar to that of a standard ARMA($p, p-1$) process, not an AR(p). The definition of Du and Li turns out to imply the standard AR(p) correlation structure. The two papers cited discuss the existence of a stationary or limiting distribution for their respective versions of the INAR(p) process, and derive sufficient conditions for such existence. The key condition in both cases is the familiar requirement that the roots λ of

$$\lambda^p - \alpha_1 \lambda^{p-1} - \cdots - \alpha_{p-1}\lambda - \alpha_p = 0$$

lie inside the unit circle. Al-Osh and Alzaid give a state-space representation of their model which can be used to find its joint distributions. Du and Li discuss Yule-Walker and conditional least squares estimation for their model, and derive a minimum variance prediction formula that is the same as for a standard AR(p).

The general INMA(q) model $\{S_t\}$ of Al-Osh and Alzaid (1988) can be described loosely as the Poisson MA(q) of subsection 1.5.3 minus the Poisson assumption. That is, it satisfies

$$S_t = Y_t + \sum_{i=1}^{q} \beta_i * Y_{t-i},$$

where $\{Y_t\}$ is a sequence of i.i.d. random variables and $0 \leq \beta_i \leq 1$ for all i. The above authors derive, inter alia, the ACF of the process, and determine the probability generating function of S_t, and the joint p.g.f. of S_t and S_{t+1}, in terms of the p.g.f. of Y_t. The ACF has the cut-off property (after lag q) that one would expect.

1.6 The bivariate geometric models of Block, Langberg and Stoffer

Langberg and Stoffer (1987) and Block, Langberg and Stoffer (1988) present accounts of the properties of certain bivariate models with geometric marginal distributions. No example of an application of these models to an observed time series is included in their papers, and the models have not been pursued in the literature. We therefore consider only the following aspects of such models: marginal distributions, correlation structure, and a brief comparison with other geometric models. In order to ease comparison of this section with the original papers we use the notation of those papers as far as possible, even though it differs from the notation used so far in this chapter.

1.6.1 Moving average models with bivariate geometric distribution

Langberg and Stoffer (1987) introduce a class of bivariate geometric moving average models $\{G(n,m) : n \in \mathbf{Z}\}$, where the positive integer m denotes the

order of dependence on the past. Before we define the models, however, it is worth noting that the above authors use the term 'geometric distribution' to mean a distribution on the *positive* integers with probability mass function of the form $p(1-p)^{k-1}$ ($k \in \mathbf{N}$) for some $p \in (0,1]$. The mean is then p^{-1} and the p.g.f. $ps/\{1-(1-p)s\}$. (Others, e.g. McKenzie, use the term for a distribution on the nonnegative integers.) A bivariate geometric distribution is any bivariate distribution with geometric marginals.

Let the column vectors $M(n) = (M_1(n), M_2(n))'$ be i.i.d. bivariate geometric random vectors, with common mean vector $(p_1^{-1}, p_2^{-1})'$. Let the column vectors $N(n) = (N_1(n), N_2(n))'$ be independent bivariate geometric with mean vectors $(\alpha_1(n)/p_1, \alpha_2(n)/p_2)'$, independent of all $M(n)$. (We suppose that $p_i \leq \alpha_i(n) \leq 1$ for $i = 1,2$ and all n.) Let $(J_1(n,j), J_2(n,j))$ be independent random vectors, independent of all $M(n)$ and $N(n)$, such that $J_i(n,j)$ is a binary random variable with probability $1-\alpha_i(n-j+1)$ of equalling 1. (This differs from the probability appearing in Langberg and Stoffer (1987): they have $1-\alpha_i(n)$. As is discussed further below, their definition appears to be in error.) Define

$$U_q(n,j) = \begin{pmatrix} \prod_{k=q}^{j} J_1(n,k) & 0 \\ 0 & \prod_{k=q}^{j} J_2(n,k) \end{pmatrix},$$

that is, $U_q(n,j) = J(n,q)J(n,q+1)\cdots J(n,j)$, where:

$$J(n,k) = \begin{pmatrix} J_1(n,k) & 0 \\ 0 & J_2(n,k) \end{pmatrix}.$$

For simplicity $U_1(n,j)$ is written $U(n,j)$, and equals $J(n,1)J(n,2)\cdots J(n,j)$. The bivariate geometric moving average model of order m, sometimes abbreviated to BGMA(m), is now defined by

$$G(n,m) = \sum_{r=0}^{m} U(n,r)N(n-r) + U(n,m+1)M(n-m).$$

It is worth noting here that there is no 'cross-coupling': the first component of $G(n,m)$, for instance, depends only on the first component of the vectors $N(n-r)$ and $M(n-m)$, and not on the second. In order to show that $G(n,m)$ has the required distribution, a bivariate geometric distribution with mean vector $(p_1^{-1}, p_2^{-1})'$, Langberg and Stoffer consider the more general random vector

$$H_q(n,m) = \sum_{r=0}^{m} U_q(n, r+q-1)N(n-r-q+1) + U_q(n, m+q)M(n-m-q+1),$$

of which $G(n,m)$ is the special case $q = 1$. They claim that $H_q(n,m)$ has the required distribution, but both steps of their inductive proof of this result (Lemma 3.7) seem to require the definition given above for $J_i(n,j)$ rather than the one they give.

The case $m = 1$, i.e. the moving average model of order one, will suffice to illustrate the structure of these processes:

$$\begin{aligned} G(n,1) &= N(n) + J(n,1)N(n-1) + J(n,1)J(n,2)M(n-2) \\ &= N(n) + J(n,1)\{N(n-1) + J(n,2)M(n-2)\}. \end{aligned}$$

Notice that, with the definition given above, all three of the following random vectors have a bivariate geometric distribution with mean vector $(p_1^{-1}, p_2^{-1})'$: $M(n-2)$, the vector in curly brackets, and $G(n,1)$.

Langberg and Stoffer give, in their equation (3.17), a general expression for the covariance structure at lag h, i.e. for the 2×2 matrix

$$\Gamma^m(n,h) = (\mathrm{Cov}(G_i(n,m), G_j(n+h,m))).$$

It should be noted that this is not in general a symmetric matrix. The case $m = 1$ of their result yields inter alia the following properties of the moving

average of order one, with $\Xi_N(n)$ denoting the covariance matrix of the vector $N(n)$:

$$\Gamma^1(n,1) = \Xi_N(n) \begin{pmatrix} 1-\alpha_1(n+1) & 0 \\ 0 & 1-\alpha_2(n+1) \end{pmatrix};$$

and, for $h = 2, 3, \ldots$:

$$\Gamma^1(n,h) = \mathbf{0}.$$

A similar cut-off property holds for higher-order models, i.e. $\Gamma^m(n,h) = \mathbf{0}$ for $h > m$.

Langberg and Stoffer also define a similar moving average model of infinite order and discuss its properties.

1.6.2 Autoregressive and autoregressive moving average models with bivariate geometric distribution

Block, Langberg and Stoffer (1988) define a bivariate geometric autoregressive model of general order m, denoted by BGAR(m), and two bivariate geometric autoregressive moving average models of orders m_1 and m_2, both denoted by BGARMA(m_1,m_2). In their definition of the BGAR process, the column vectors $M(n)$ and $N(n)$ are as in subsection 1.6.1. The binary random variables $J_i(n,m)$ are however defined differently from before. We suppose now that the $2m$-component random vectors $J(n)$, defined by

$$J(n) = (J_1(n,1), \ldots, J_1(n,m), J_2(n,1), \ldots, J_2(n,m)),$$

satisfy the requirements

$$\mathrm{P}((J_i(n,1), \ldots, J_i(n,m)) = \mathbf{0}) = \alpha_i(n)$$

and

$$\sum_{j=1}^{m} \mathrm{P}((J_i(n,1),\ldots,J_i(n,m))=e_j) = 1-\alpha_i(n)$$

for $i = 1,2$ and all n, where e_j has its jth component equal to one and the other $m-1$ components equal to zero. That is, $(J_i(n,1),\ldots,J_i(n,m))$ is either all zeroes (with probability $\alpha_i(n)$), or a single one and $m-1$ zeroes (with probability $1-\alpha_i(n)$). We suppose that the vectors $J(n)$ are independent of each other and of all vectors $M(n)$ and $N(n)$. Define also, for $q = 1,2,\ldots,m$:

$$C(n,q) = \begin{pmatrix} J_1(n,q) & 0 \\ 0 & J_2(n,q) \end{pmatrix}.$$

The BGAR(m) process is then defined by

$$G(n) = \begin{cases} M(n) & n = 0,1,\ldots,m-1 \\ \sum_{q=1}^{m} C(n,q)G(n-q) + N(n) & n = m, m+1,\ldots\ . \end{cases}$$

It follows, by induction on n, that $G(n)$ has a bivariate geometric distribution with mean vector $(p_1^{-1}, p_2^{-2})'$.

If we assume that $\alpha_i(n) = \alpha_i$ for all n and $i = 1$ and 2, so that the marginal processes $\{G_i(n)\}$ are stationary, we can derive equations for the autocorrelations

$$\rho_i(k) = \mathrm{Corr}(G_i(n), G_i(n+k)) \quad (i = 1,2;\ n = m, m+1,\ldots;\ k \in \mathbf{N}_0).$$

If we define $\gamma_i(q) = \mathrm{P}(J_i(n,q) = 1)$ for $i = 1,2$ and $q = 1,\ldots,m$, so that $\sum_{q=1}^{m} \gamma_i(q) = 1-\alpha_i$, the result is the following equation of Yule-Walker form, for $k = m, m+1,\ldots$:

$$\rho_i(k) = \gamma_i(1)\rho_i(k-1) + \gamma_i(2)\rho_i(k-2) + \cdots + \gamma_i(m)\rho_i(k-m).$$

Assuming $\alpha_i(n) = \alpha_i$ for all n, while sufficient for marginal stationarity, is not sufficient for stationarity of the bivariate process. Block et al. describe also a (bivariate-) stationary BGAR(1) model, and derive its covariance structure, i.e. the auto- and cross-covariances at a given lag $k \in \mathbf{N}_0$. The covariance structure turns out to be very similar to (but slightly simpler than) the covariance structure of a bivariate Gaussian AR(1) model: compare equation (4.12) of Block et al. with equation (9.4.7) of Priestley (1981). The model under discussion is simpler because there is no cross-coupling in its definition.

The two bivariate geometric ARMA models $\{L(n)\}$ defined by Block et al. may be described briefly as follows. Let $\{G(n)\}$ be a BGMA(m_1) model with mean vector $(\beta_1/\delta_1, \beta_2/\delta_2)'$, and $\{H(n)\}$ a BGAR(m_2) with mean vector $(\delta_1^{-1}, \delta_2^{-1})'$. Let $U_1(n)$ and $U_2(n)$ be binary variables with $\mathrm{P}(U_i(n) = 1) = 1 - \beta_i$. Then with appropriate independence assumptions we may define

$$L_i(n) = G_i(n) + U_i(n)H_i(n)$$

for $i = 1$ and 2, and so obtain a process $\{L(n)\}$ such that $L(n)$ has a bivariate geometric distribution. Alternatively, we may exchange the mean vectors of $G(n)$ and $H(n)$, and define instead

$$L_i(n) = H_i(n) + U_i(n)G_i(n).$$

Various other properties of all the bivariate geometric models defined above are considered in detail in the two papers cited, especially positive dependence properties. In the context of the present work, however, it may be more interesting to consider an example of one of the models and compare it with similar time series models with geometric marginal.

Let us therefore consider the first component, $\{G_1(n)\}$, of a BGAR(1) model $\{G(n)\}$. The fundamental property of the sequence $\{G_1(n)\}$ is that for $n \in \mathbf{N}$

$$G_1(n) = J_1(n,1)G_1(n-1) + N_1(n),$$

where the three random variables on the right-hand side are independent, $J_1(n,1)$ is binary, and $N_1(n)$ is a geometric 'innovation', but with mean differing from that of $G_1(n-1)$ and $G_1(n)$ (except in a trivial case). The most important difference between this model and the geometric AR(1) of McKenzie is that $G_1(n-1)$ is not here 'thinned': either the whole of $G_1(n-1)$ is included in $G_1(n)$ (along with $N_1(n)$), or it is not included at all. In McKenzie's model, each of the 'individuals' present at time $n-1$ is considered *separately* and, with the appropriate survival probability, included among the survivors to time n.

The geometric DAR(1) process of Jacobs and Lewis may be described as a process, with geometric marginal, such that the value at time n is either the same as the value at time $n-1$ (with probability ρ), or else an innovation having exactly the geometric distribution required as marginal. Therefore the construction of such a process also differs from that of $\{G_1(n)\}$. All three of these 'geometric AR(1)' models do however have ACF of geometrically decreasing form α^k — as does yet another geometric AR(1), the model described at the end of subsection 1.5.2.

1.7 Markov regression models

We now discuss a very useful class of models, described by Zeger and Qaqish (1988) as Markov regression models. Such models are essentially an extension to the time series context of the ideas of generalized linear models and quasi-likelihood. Their utility is certainly not confined to discrete-valued ob-

servations, although in this work we consider such applications only. The principal advantages of this class of models are their flexibility, especially as regards the inclusion of covariates, and the ease with which parameter estimates may be computed by standard software.

We begin by describing an example of such a model $\{S_t\}$ and the way in which it may be fitted to data. Suppose that S_t, the observation at time t, is the number of 'successes' in n_t trials and that, conditional on the history $S^{(t-1)} = \{S_k : 1 \leq k \leq t-1\}$, S_t has a binomial distribution with parameters n_t and p_t, where for some positive integers q and r:

$$\begin{aligned} \text{logit } p_t &= \alpha_1 + \alpha_2 t + \gamma_1 \sin(2\pi t/r) + \gamma_2 \cos(2\pi t/r) \\ &\quad + \beta_1(S_{t-1}/n_{t-1}) + \cdots + \beta_q(S_{t-q}/n_{t-q}). \end{aligned} \tag{1.7}$$

This model is similar to, but slightly more general than, the 'linear logistic autoregression' of Cox (1981), which is based on a Bernoulli rather than a binomial distribution and lacks the terms representing trend and seasonality. Model (1.7) makes allowance for dependence of the success probability on the proportion of successes at each of the previous q time points, time trend and r-period seasonality. Clearly it is straightforward to add further terms to the above expression for logit p_t to allow for the effect of any further covariates thought relevant. (There is a problem if, for example, $n_{t-1} = 0$, but it is not clear how one should modify the model to allow for such a case.) The model is an example of what Cox (1981) terms 'observation-driven' models: it is observation-driven in the sense that the distribution of the observation at a given time is specified in terms of the observations at earlier times. In the next section we shall consider some examples of a very different class of models, the processes described by Cox as 'parameter-driven'.

Suppose we have available a realization $\{s_t : 1 \leq t \leq T\}$ of model (1.7).

The likelihood of $s_{q+1}, \ldots, s_T$, conditional on the first q observations, is just

$$\prod_{t=q+1}^{T} \binom{n_t}{s_t} p_t^{s_t} (1-p_t)^{n_t - s_t}.$$

To estimate the parameters $\alpha_1, \alpha_2, \gamma_1, \gamma_2, \beta_1, \ldots, \beta_q$ we may maximize the conditional likelihood with respect to these parameters by performing a logistic regression of S_t on t, $\sin(2\pi t/r)$, $\cos(2\pi t/r)$, and $S_{t-1}/n_{t-1}, \ldots, S_{t-q}/n_{t-q}$. A program such as GLIM or GENSTAT may conveniently be used for this purpose.

More generally, let $\{S_t\}$ be the observed time series, and X_t a (row) vector of p covariates. Let D_t, the 'information set', consist of the past observations $S_1, \ldots, S_{t-1}$ and present and past covariate vectors $X_1, \ldots, X_t$. Let the conditional mean of S_t, given D_t, be denoted by μ_t and suppose that the corresponding conditional variance is $\phi V(\mu_t)$. (This relation between mean and variance is the usual quasi-likelihood assumption except that we are dealing here with conditional rather than marginal means and variances.) Suppose further that some link function g of the conditional mean is linear in the current covariates X_t and in q known functions of the past observations and covariates. That is, for some functions f_i, often of S_{t-i} and X_{t-i} only, and some vectors β and $\theta = (\theta_1, \ldots, \theta_q)$ of parameters, the following is true:

$$g(\mu_t) = \beta X_t' + \sum_{i=1}^{q} \theta_i f_i(D_t). \tag{1.8}$$

A simple example of such a model, given by Zeger and Qaqish (1988), is that for binary outcomes

$$g(\mu_t) = \text{logit } \mu_t = \beta X_t' + \sum_{i=1}^{q} \theta_i s_{t-i}. \tag{1.9}$$

In this case $V(\mu) = \mu(1-\mu)$ and $\phi = 1$.

For models satisfying (1.8) the estimation of the vector $\gamma = (\beta\ \theta)$ from a realization $s_1, \ldots, s_T$ may be accomplished by solving the (conditional) quasi-likelihood estimating equations

$$\sum_{t=q+1}^{T} \frac{\partial \mu_t}{\partial \gamma}(s_t - \mu_t)/V(\mu_t) = \mathbf{0} \tag{1.10}$$

by iterative weighted least squares, e.g. by using GLIM. These equations result from equating to zero the 'quasi-score function', i.e. the derivative with respect to γ of the log of the quasi-likelihood function. If we define

$$Z_t = (X_t, f_1(D_t), \ldots, f_q(D_t))$$

and note that

$$\frac{\partial \mu_t}{\partial \gamma} = \frac{1}{\dot{g}(\mu_t)} Z_t,$$

where $\dot{g}$ denotes the derivative of g, we see that the estimating equations (1.10) are equivalent to

$$\sum_{t=q+1}^{T} \frac{1}{\dot{g}(\mu_t)} Z_t(s_t - \mu_t)/V(\mu_t) = \mathbf{0}.$$

When $\dot{g}(\mu) = 1/V(\mu)$, g is described as a 'canonical' link, and the equations reduce to

$$\sum_{t=q+1}^{T} Z_t(s_t - \mu_t) = \mathbf{0}.$$

Since the term 'canonical link' is usually defined in the context of a distribution assumed to belong to the exponential family, the above meaning is a slight extension of the usual. The link function $g(\mu) = \text{logit}\ \mu$ in the model (1.9) is an example of a canonical link.

If $\hat{\gamma}$ denotes the parameter estimates thus obtained, the distribution of $\sqrt{T}(\hat{\gamma} - \gamma)$ converges (under appropriate regularity conditions) to multivariate normal with mean zero and covariance matrix ϕW^{-1}, where

$$W = \lim_{T \to \infty} T^{-1} \sum_{t=q+1}^{T} Z_t' \dot{g}(\mu_t)^{-2} V(\mu_t)^{-1} Z_t.$$

If g is a canonical link, W reduces to $\lim_{T\to\infty} T^{-1}\sum_{t=q+1}^{T} Z_t' V(\mu_t) Z_t$. The scale parameter ϕ may in general be estimated by

$$\hat{\phi} = T^{-1}\sum_{t=1}^{T} \hat{a}_t^2,$$

where $\hat{a}_t$ is the residual $(s_t - \hat{\mu}_t)/V(\hat{\mu}_t)^{1/2}$.

The above treatment is essentially that of Zeger and Qaqish (1988) and Li (1991), except that it ignores two complications discussed by the former authors. Firstly, the functions f_i may in fact not be known completely, but may depend on the parameters β. Zeger and Qaqish describe several useful models of this kind, and generalize the estimation algorithm to allow for this possibility. Secondly, it may be necessary to estimate parameters other than those included in β and θ. A modification of the estimation algorithm is also possible in that case. Zeger and Qaqish describe in detail an application of their methods, but to a continuous-valued series, which is not directly of interest here.

Li (1991) introduces two methods of assessing the adequacy of a Markov regression model. One is based on the residual autocorrelations

$$\hat{C}_k = T^{-1}\sum_{t=k+1}^{T} \hat{a}_t\hat{a}_{t-k}/\hat{\phi},$$

which are shown to have asymptotically a multivariate normal distribution with zero mean. The other method, which is simpler to use, is based on the score function (more pedantically, the quasi-score). The score statistic derived has a chi-squared asymptotic distribution. Li describes a simulation study of the behaviour of the score statistic in respect of three discrete-valued models without covariates, and reports 'quite reasonable' results.

Fahrmeir and Kaufmann (1987) and Kaufmann (1987) describe models for categorical time series which are similar to the models of Zeger and Qaqish except in two respects. Firstly, they are not set in a quasi-likelihood context. Secondly, in order to represent the m categories of possible outcome, the response is a vector of $m-1$ binary variables, and not a scalar as in the case of the models of Zeger and Qaqish. Kaufmann proves very general results relating to the asymptotic normality, consistency and efficiency of maximum likelihood estimators for (inter alia) these models, results which are in fact the basis of some of the conclusions drawn by Zeger and Qaqish. Fahrmeir and Kaufmann discuss the testing of linear hypotheses on the parameters of their models by means of three different (but asymptotically equivalent) statistics, and indicate some tests of particular practical significance, e.g. a test of independence of parallel series and a test of nonsignificance of the covariates. They report also a simulation study of the finite-sample properties of the estimators, in particular in the case of a binary-response model with first-order Markov dependence and two binary covariates. The results are promising, and they conclude that the asymptotic distributions seem to be sufficiently accurate for inference purposes, even at moderate sample sizes.

Liang and Zeger (1989) and Zeger and Liang (1991) discuss multivariate Markov regression models in which at least some of the components of the vector of responses may be discrete. The first paper relates specifically to multivariate binary series, and to conditional logistic regression models for such series which specify the distribution of one response variable given the past values of that response, current and past values of the other responses, and current values of covariates. The conditional log-likelihood approach is computationally burdensome, and a 'pseudolikelihood' is maximized instead. An application to health-care utilization by 300 families enrolled in a health-maintenance plan in Maryland is described.

The second paper cited above refers to situations (and models) in which the vector of responses may include both discrete and continuous components. For each response some link function of the conditional mean is taken to be linear in the current covariates, the current values of the other responses, and the past values of all the responses. The conditional variance of the jth response is assumed to be proportional to some known function V_j of the conditional mean. The approach already described, based on the quasi-likelihood 'estimating equations', may be generalized to provide an estimate of the vector of parameters. The distribution of this estimate converges, as before, to multivariate normal with the correct mean. An application to infectious diseases and vitamin A deficiency in Indonesian pre-school children is described in detail by Zeger and Liang.

The 'linear contagion model' of Holden (1987) for aircraft hijackings in the U.S.A. in the years 1968–72 can also be described as a Markov regression model. Conditional on the history, the number of hijacking events in period t is taken to be a Poisson random variable with mean λ_t, where λ_t is a linear function of the numbers of events in preceding periods, possibly with similar linear contributions from the histories of several covariates. The weight structure chosen by Holden implies that any past hijacking event makes a positive contribution to λ_t, a contribution that is initially low, with time increases to a peak, and then dies out. A similar model for inhibition rather than contagion would have negative weights rather than positive.

MacDonald (1989) presents an application of logistic-linear autoregressive models, similar to (1.7), to a data-set relating to births in successive months at Edendale hospital in Natal, South Africa. In that paper the number of deliveries by Caesarean section is modelled as a proportion of the total

number of deliveries, and related, inter alia, to the corresponding proportion in the previous month. A more detailed discussion appears in section 4.3 of this thesis.

1.8 Parameter-driven models

We now consider models of the kind described by Cox (1981) as 'parameter-driven'. Conditional on some unobserved 'parameter process', the observations in such a model are independent, with distribution determined by the current state of the parameter process. For instance, the conditional distribution of an observation could be Poisson, with mean λ_1 or λ_2 depending on whether a two-state parameter process is in state one or state two. Alternatively, the mean could be provided by some positive-valued process having (say) gamma or lognormal marginal distribution. As Cox explains, there may be circumstances in which a parameter-driven model is appropriate to, or even strongly suggested by, the data. Consider a very long binary series consisting almost entirely of zeroes, but with occasional bursts of ones fairly close together. Such observations suggest some underlying process which occasionally changes to a state in which one or zero is possible, and then reverts to its normal state, in which only zeroes are possible.

In a parameter-driven process (as defined above), the only source of dependence and autocorrelation between observations is the parameter process. If the parameter process happens to consist of independent random variables, the observations are independent, with distribution compounded by the marginal distribution of the parameter process.

The models of Keenan (1982) for binary time series are examples of parameter-driven processes: they can be described briefly as follows. Let

$\{X_t\}$ be an unobserved (completely) stationary process with state-space $\mathbf{R}$, and let the 'response function' $F : \mathbf{R} \to [0,1]$ be monotone. Suppose that, conditional on $\{X_t\}$, the random variables $\{S_t\}$ are independent, with (conditional) distribution specified by

$$\mathrm{P}(S_t=1) = 1 - \mathrm{P}(S_t=0) = F(X_t).$$

One model $\{S_t\}$ considered in particular is that which results if $\{X_t\}$ is a Gaussian process with mean zero and F the distribution function of a normal distribution with mean zero and variance b^2. In that case explicit expressions are found for the distributions of two and three consecutive observations in the process $\{S_t\}$. The joint distribution of more than three consecutive observations is not available in closed form. The one-step-ahead forecast distribution based on observations $s_1, s_2, \ldots, s_T$ must therefore be approximated if $T > 2$. Keenan presents and compares six different approximations to this distribution under the assumption that $\{X_t\}$ is a Gaussian AR(1). The use of the sample autocorrelations of $\{S_t\}$ to estimate the parameters of $\{X_t\}$ and the parameter b in the response function is also explored in some generality.

The models of Kedem (1980) for 'clipped' series, although similar to those of Keenan, do not really meet the definition of parameter-driven, because the original series, as well as the clipped version, is assumed to be available. Some details will be presented in section 1.10.

Azzalini (1982) discusses a model in which the parameter process $\{\theta_t\}$ is the gamma AR(1) process of Gaver and Lewis (1980) and provides the mean for observations which are conditionally Poisson. The three parameters of the gamma AR(1) are initially assumed known. A fairly simple recursive filtering procedure is proposed, i.e. a procedure for estimating the unobserved

state variable θ_t from $S^{(t)}$, the history up to and including time t. This Azzalini compares with a second filtering algorithm, which is very similar to the Kalman filter: a simulation experiment suggests that the first method is slightly better in terms of squared-error loss. Another possiblity investigated by Azzalini is to use the first fifty observations (say) of a realization to estimate the three parameters by method of moments, and thereafter to use the observations for both parameter estimation and filtering. Again the first filtering method seems slightly the better of the two.

Zeger (1988) introduces a class of parameter-driven loglinear regression models for time series of counts. They are defined as follows. Suppose that, conditional on the unobserved process $\{\varepsilon_t\}$, the observed process $\{S_t\}$ is a sequence of independent counts, with S_t having conditional mean and variance both equal to $\exp(\beta X_t')\varepsilon_t$. The (row) vector X_t contains the values at time t of the p covariates, and the coefficients β are the quantities of interest. Suppose further that $\{\varepsilon_t\}$ is a stationary process with $\mathrm{E}(\varepsilon_t)=1$ and

$$\mathrm{Cov}(\varepsilon_t, \varepsilon_{t+\tau}) = \sigma^2 \rho_\varepsilon(\tau)$$

for all $\tau \in \mathbf{N}_0$. The marginal properties of $\{S_t\}$ are then:

$$\mu_t = \mathrm{E}(S_t) = \exp(\beta X_t')$$
$$\mathrm{Var}(S_t) = \mu_t + \sigma^2 \mu_t^2$$

and, for $\tau \in \mathbf{N}$ but not $\tau = 0$,

$$\mathrm{Corr}(S_t, S_{t+\tau}) = \frac{\rho_\varepsilon(\tau)}{[1 + (\sigma^2\mu_t)^{-1}]^{1/2}[1 + (\sigma^2\mu_{t+\tau})^{-1}]^{1/2}}.$$

This model is similar to the conventional Poisson-loglinear model in that $\log \mu_t = \beta X_t'$, but different in that $\{S_t\}$ displays autocorrelation and overdispersion relative to the Poisson:

$$\mathrm{Var}(S_t) = \mu_t(1 + \sigma^2 \mu_t) > \mu_t.$$

Furthermore the extent of this overdispersion depends on μ_t. For estimation of β, from observations $s = (s_1, \ldots, s_T)'$, Zeger proposes that the solution $\hat{\beta}$ of the following quasi-likelihood estimating equations be obtained by iterative weighted least squares:

$$D'V^{-1}(s - \mu) = \mathbf{0},$$

where

$$\mu = (\mu_1, \ldots, \mu_t)',$$
$$V = \text{the covariance matrix of } S_1, \ldots, S_T, \text{ and}$$
$$D = \left(\frac{\partial \mu_i}{\partial \beta_j}\right).$$

Unlike the usual case of independence, or even the case of the Markov regression models, V is here non-diagonal. Inversion of V is difficult, and Zeger suggests as an alternative to $\hat{\beta}$ an estimate $\hat{\beta}_R$ based on V_R, an approximation to V which is easier to invert. Both algorithms require estimates of σ^2 and other nuisance parameters: Zeger proposes moments estimators for these. Asymptotic results for $\hat{\beta}$ and $\hat{\beta}_R$ are stated, and the efficiency of $\hat{\beta}_R$ relative to $\hat{\beta}$ is discussed for some simple models. A model of the above type is fitted to U.S. polio incidence data for the years 1970–1983, and used to estimate the time trend in incidence. Comparisons are made with two conventional loglinear models which assume independence.

The hidden Markov models of Chapters 2 and 3 of this thesis are parameter-driven processes, and fairly straightforward to use as statistical models. In such models the parameter process is a Markov chain, or possibly a higher-order Markov chain.

1.9 State-space models

The 'state-space' time series models of firstly Harvey and secondly West and Harrison do not fit into the framework of either of the two broad categories

described above, observation-driven and parameter-driven processes. Both of these state-space approaches are applicable to discrete-valued series, and have been developed at some length in the books of Harvey (1989) and West and Harrison (1989), as well as in the papers of Harvey and Fernandes (1989a, 1989b) and West, Harrison and Migon (1985), inter alia. We therefore give here only brief accounts of these approaches. A comparative review of the above books has been provided by Fildes (1991).

We illustrate the 'structural' models of Harvey by means of an example, the model in which the observations are Poisson with mean given by a gamma process. Similar models are available for other distributions: see Harvey (1989), section 6.6. Suppose then that, given μ_t, the observation S_t has a Poisson distribution with mean μ_t, and that the process $\{\mu_t\}$ evolves as follows. Conditional on the information available at time $t-1$, i.e. the history $S^{(t-1)}$, μ_{t-1} has a gamma distribution with shape and scale parameters a_{t-1} and b_{t-1} respectively. (These parameters are as defined in subsection 1.5.2.) Given the same information, μ_t is taken to have a gamma distribution with parameters

$$a_{t|t-1} = \omega a_{t-1}$$
$$b_{t|t-1} = \omega b_{t-1}$$

for some $\omega \in (0, 1]$. Updating this prior information with the observation S_t, we obtain the posterior for μ_t, which is gamma with parameters given by

$$a_t = a_{t|t-1} + S_t$$
$$b_t = b_{t|t-1} + 1.$$

(This makes use of the properties of the gamma distribution as natural conjugate prior for the Poisson.) Using the result of Lewis, McKenzie and Hugus (1989) on the beta-gamma transformation, we can see that the transition from μ_{t-1} (given $S^{(t-1)}$) to μ_t (also given $S^{(t-1)}$) could equivalently be de-

scribed by

$$\mu_t = \omega^{-1}\mu_{t-1}\eta_t$$

for some η_t independent of μ_{t-1} and having the beta distribution with parameters ωa_{t-1} and $(1-\omega)a_{t-1}$. Two consequences of the definition of $\{\mu_t\}$ are that:

$$\mathrm{E}\left(\mu_t \mid S^{(t-1)}\right) = a_{t|t-1}/b_{t|t-1} = \mathrm{E}\left(\mu_{t-1} \mid S^{(t-1)}\right)$$

and

$$\mathrm{Var}\left(\mu_t \mid S^{(t-1)}\right) = a_{t|t-1}/b_{t|t-1}^2 = \omega^{-1}\mathrm{Var}\left(\mu_{t-1} \mid S^{(t-1)}\right).$$

The process $\{\mu_t\}$ is initialized with $a_0 = b_0 = 0$, although that yields an improper distribution until the first time (τ) at which there is a nonzero observation. The likelihood of observations $S_{\tau+1}, \ldots, S_T$, conditional on $S^{(\tau)}$, is obtained as a product of negative binomial probabilities, and may be maximized in order to estimate the 'hyperparameter' ω.

The one-step-ahead forecast function yielded by this approach can be shown to be a weighted mean of observations in which the weights decline exponentially, and given approximately by an exponentially weighted moving average if the sample is large. Harvey shows also that the k-step-ahead forecast function, for $k > 1$, is identical to that for one step ahead.

The dynamic generalized linear model of West and Harrison is an extension both of their own (normal-theory) dynamic linear model and of the usual 'static' generalized linear model. It is an extension of the first in that non-normal distributions (including discrete) are provided for, and of the second in that time-varying parameters are incorporated in the model. It is a more fully Bayesian approach than that of Harvey, and attempts to solve a more general problem. In such a model the scalar observation S_t has an exponential family distribution. The associated canonical parameter has the

natural conjugate prior distribution and is linked to the state vector θ_t (a column vector) by

$$g(\eta_t) = F_t\theta_t,$$

where the monotone function g and the (row) vector F_t are assumed known. Frequently g is the identity function. The state vector evolves according to

$$\theta_t = G_t\theta_{t-1} + \omega_t,$$

where the matrix G_t is known and ω_t has mean zero and known covariance matrix W_t. A useful summary of the recursions involved in obtaining the posterior for η_t, updating the state vector, and obtaining the k-step-ahead forecast distribution for $\{S_t\}$, is presented by West, Harrison and Migon in their section 3.3. They include also four examples of the application of their methods to discrete data.

1.10 Miscellaneous models

An autoregressive model for binary (i.e. zero-one) series that seems to have been used more as a building-block for other models than as a time series model in its own right is that of Kanter (1975). The (stationary) model of order N satisfies:

$$S_t = \begin{cases} S_{t-k} \oplus U_t & \text{with probability } p_k, k = 1, 2, \ldots, N \\ U_t & \text{with probability } p, \end{cases}$$

where $\{U_t\}$ is a sequence of i.i.d. binary random variables, $\oplus$ denotes addition modulo two, and $p + \sum_{k=1}^{N} p_k = 1$. McKenzie (1981) uses Kanter's model, and a similar binary moving average model, to generalize several EARMA processes (for positive continuous-valued series) and the DMA(1) process of Jacobs and Lewis, by replacing the independent binary mixing in the definition of such processes by dependent. He thereby extends the range of

correlations possible in those processes. In the course of so doing he indicates how to generalize Kanter's model to one of general ARMA structure (rather than merely autoregressive).

Another binary sequence sometimes mentioned as a possible time series model is the model of Klotz (1973) for Bernoulli trials with dependence. This is however just a stationary two-state Markov chain parametrized by $p = \delta_2$ and $\lambda = 1 - \gamma_2$, where δ_2 and γ_2 are as in section 1.2. Klotz suggests certain easily computed ad hoc estimators of p and λ, and applies them to rainfall data. Devore (1976) claims that the ad hoc procedure is unnecessary because estimation based on either the full unconditional likelihood or the likelihood conditional on the first observation is very straightforward. He claims in particular that the the unconditional maximum likelihood estimates can be found by solving a quadratic equation. As is pointed out by Bisgaard and Travis (1991), and the authors cited by them, this last claim of Devore is erroneous: the estimates are in fact given by the solution of a cubic equation.

Kedem's book (Kedem, 1980) is mainly concerned with binary series derived from some observed continuous-valued series $\{Z_t\}$ by 'clipping' or 'hard limiting', i.e. by a transformation

$$S_t = \begin{cases} 0 & Z_t < a \\ 1 & Z_t \geq a. \end{cases}$$

Typically the observations Z_t are taken to be generated by a stationary Gaussian autoregressive process, and the object is to estimate parameters by using only the clipped series $\{S_t\}$. Such a procedure may well be attractive in circumstances of very fast data acquisition. By letting F be the step function

$$F(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a, \end{cases}$$

we see that, given $\{Z_t\}$, $\mathrm{P}(S_t = 1) = F(Z_t)$. Apart from the distinction already mentioned in section 1.8, the clipped series is therefore a model of the same general type as those investigated by Keenan. As it is a rather special-purpose model we shall not consider it further here.

Blight (1989) has derived properties of certain discrete-valued series formed by superposing a number of independent renewal processes which all start with an 'event' at time zero and have interevent intervals taking positive integer values only. One such renewal process generates a binary series in obvious fashion: at each time point $t \in \mathrm{N}$ there is either one event or none. If one superposes N independent processes of this kind, not necessarily identical in nature, the result is a series $\{S_t\}$ taking values in $\{0, 1, \ldots, N\}$. If the N processes are independent copies of the same renewal process, the distribution of S_t is binomial. In this case (of identical processes being superposed) Blight derives an ARMA representation for the process $Z_t = S_t - \mathrm{E}(S_t)$ on the assumption that the probability generating function of the interevent intervals is rational or polynomial. For instance, if that generating function is

$$\left(\frac{pz}{1-(1-p)z}\right)^2,$$

then $\{Z_t\}$ satisfies the ARMA representation

$$Z_t - (1-2p)Z_{t-1} = a_t - \beta a_{t-1},$$

where $\beta = (1-p)^{-1}\{(1-p+p^2) - p\sqrt{2-2p+p^2}\}$ and $\{a_t\}$ is an uncorrelated 'noise' sequence. Some generalizations are indicated, e.g. to the case of the superposed processes being nonidentical, and to the case of the p.g.f. being neither rational nor polynomial.

Under the heading of 'Walsh-Fourier analysis' Stoffer (1985, 1987, 1990, 1991) has provided an extensive account of what has been termed a sequency

domain approach to the analysis of time series, especially discrete-valued and categorical series. Although the emphasis in this thesis falls on time domain analysis, we present here a brief description of this alternative approach. The fundamental idea of Walsh-Fourier analysis may be stated as follows. There are series of observations, e.g. discrete-valued, in which the signal can be modelled better by the superposition of square waveforms than sinusoidal. The Walsh functions, like the trigonometric functions employed in Fourier analysis, form a complete orthogonal sequence on [0,1). They are similar in oscillation and many other properties to the trigonometric functions, but unlike the trigonometric functions, which vary smoothly, they assume two values only: $+1$ and -1. Observations displaying sharp discontinuities and a limited number of levels can therefore be represented better by the Walsh functions than by trigonometric. Stoffer discusses the 'Walsh-Fourier' analogues of various standard techniques of Fourier analysis, and in Stoffer (1987) shows how Walsh-Fourier analysis can be used, for instance, to estimate transition probabilities in a 'macro model' which aggregates several independent copies of a Markov chain and superimposes a (discrete-valued) noise term.

1.11 Discussion

It has been stated as recently as 1988 that 'time series models for a sequence of dependent discrete random variables ... are rare' (Al-Osh and Alzaid, 1988). The reader of this chapter might be forgiven, however, for concluding that published applications of many of the models that do exist are rarer. For instance, few applications of the 'marginally specific' models of sections 1.4–1.6 seem to have appeared yet. There are even those who question the usefulness of an approach which is based on marginal distributions rather than on conditional distributions given the history of the process: see the comments of Diggle and Westcott (1985), although it should be noted

that those comments were made in the context of positive-valued series, not discrete-valued. What Diggle and Westcott were suggesting is the use of models which are specified in observation-driven form, rather than models designed to have a given marginal distribution and dependence structure, some of which are rather contrived. But given the success of the Gaussian ARMA models in the analysis of continuous-valued series, it is surely sensible to find out whether any similar approach can be made to work for at least some kinds of discrete-valued series. Difficulties in obtaining the likelihood do seem to be an obstacle to the use of many of the marginally specific models.

The Markov regression models are better-developed and easier to apply: in some cases existing software can be used directly. They are applicable to a variety of discrete-valued time series. The state-space models of section 1.9 present a more or less Bayesian approach to the modelling problem under discussion: more in the case of the dynamic generalized linear model of West and Harrison, less in the case of Harvey's structural time series models. While Harvey believes that there is nothing in his proposed methods to which a classical statistician could object (Harvey and Fernandes, 1989a), there are those who are uneasy about the mix of classical and Bayesian notions involved (Winkler, 1989). The apparent complexity of the structure of the dynamic GLM does present a barrier to its easy application. Furthermore, the methods of West and Harrison are, in common with many other Bayesian proposals, open to the criticism that a large number of quantities are assumed known a priori. Nevertheless the great flexibility of the dynamic GLM makes it potentially a very useful tool, and its application has certainly been taken further than is true of many of the models surveyed in this chapter.

Parameter-driven models, as pointed out by Cox and Snell (1989, p. 101), can be difficult to use as a basis for the analysis of data. Compare, for in-

stance, the parameter-driven models of Zeger (1988) with the observation-driven models of Zeger and Qaqish (1988). Since observation-driven models are not always appropriate, it is worthwhile to try to develop a class of parameter-driven models which are parsimonious, flexible and fairly easy to apply. It will be argued in chapters 2 and 3 that hidden Markov models are such a class of models. The ease with which they can be modified or extended in order to accommodate many different kinds of data is a major advantage. Among the types of time series data for which hidden Markov models can be used are: unbounded counts (i.e. Poisson-like observations), bounded counts (binomial- or multinomial-like observations), multivariate discrete observations, categorical observations, vector observations with some components discrete and some continuous, and discrete observations displaying trend or seasonality or dependence on covariates other than time.

Chapter 2

Hidden Markov models for discrete-valued time series

2.1 Introduction

Probabilistic functions of a Markov chain, also known as hidden Markov models, have long been used in speech processing: see for instance Levinson, Rabiner and Sondhi (1983), Ephraim and Rabiner (1990), or Juang and Rabiner (1991). The term 'hidden Markov model' is apparently due to L.P. Neuwirth (Poritz, 1988). Such models or similar have also been used, although to a lesser extent, in genetics and biochemistry: see Thompson (1983), Churchill (1989) and Guttorp, Newton and Abkowitz (1990). Juang (1985) states in passing that hidden Markov models 'have been found to be extremely useful for stock market behavior', but gives no reference. A similar claim is made by Kemeny, Snell and Knapp (1976, p. 468). Zucchini and Guttorp (1991) apply hidden Markov models to the modelling of the wet-dry sequence at one or several sites — that is, they use them as multivariate binary time series models. MacDonald (1990) proposes hidden Markov models as general-purpose models for discrete-valued time series, and applies them,

inter alia, to the geyser data of Azzalini and Bowman (1990). The work of Albert (1991) applies hidden Markov models to epileptic seizure counts, and that of Leroux and Puterman (1992) applies them to the pattern of movements of a foetal lamb.

In a hidden Markov model an underlying and unobserved sequence of states follows a Markov chain with finite state-space, and the probability distribution of the observation at any time is determined only by the current state of that Markov chain. The main object of this chapter is to develop such models as general-purpose models for discrete-valued time series. First, however, we review relevant aspects of the theory of hidden Markov models as applied in speech processing. We describe and derive in detail the 'Baum-Welch re-estimation algorithm', the algorithm apparently most used to solve the estimation (or 'training') problem in such applications. It will not disturb the continuity greatly if the reader should initially omit section 2.2, as the models and techniques proposed in this thesis are self-contained and do not require any of the results of that section for their derivation or implementation. Nevertheless it is interesting to compare the various different approaches to estimation in hidden Markov models, and this will be done where appropriate.

The models proposed in this thesis, which differ in some respects from those used in speech processing and from those of Albert and those of Leroux and Puterman, first appear in section 2.3. Thereafter their correlation properties are derived, and the key property of these models which makes them feasible as practical statistical models is discussed. This is the property that the likelihood of even a very long sequence of observations can be computed sufficiently fast to enable parameters to be estimated by direct numerical maximization of that likelihood. In section 2.7 this estimation

technique is compared in detail with that of Leroux and Puterman, which (like the Baum-Welch algorithm) is an implementation of the EM algorithm. The marginal, joint and conditional distributions of the models proposed are derived in section 2.6, and the statistical implications of these results are indicated, in particular their relevance to forecasting and the treatment of missing data. Section 2.8 discusses the reversibility or otherwise of the models, and shows that reversibility of the observed process is not equivalent to that of the underlying Markov chain. In section 2.9 some concluding remarks are made, in particular on the way in which the proposed models may be used in practice.

2.2 Some aspects of hidden Markov models in speech processing

One use of hidden Markov models in isolated word recognition (as opposed to continuous speech recognition) may be described briefly as follows. We wish to be able to recognize utterances known to come from a known vocabulary of V words. (The term 'word' is to be interpreted broadly, as meaning the language unit being modelled, not necessarily a word in the usual sense. It could, for instance, be some subword unit.) Each utterance gives rise to an observation sequence (the acoustic signal) $s_1, s_2, \ldots, s_T$, which is regarded as a realization of length T of some random process $\{S_t : t \in \mathbf{N}\}$ of finite state-space. The process $\{S_t\}$ is taken to be generated by two probabilistic mechanisms: firstly, an unobserved (homogeneous) Markov chain $\{C_t\}$ on m states representing the configurations of the vocal tract at successive instants of time, and secondly a set of probability distributions, one for each state of $\{C_t\}$, that produce the observations from a finite set of n possibilities. Such a hidden Markov model can therefore be characterized by the distribution

of C_1 (denoted by δ), the transition probability matrix of the Markov chain (Γ), and the $n \times m$ matrix Π of probabilities defined by

$$\pi_{si} = \mathrm{P}(S_t = s \mid C_t = i).$$

The 'training problem', which must be solved separately for each of the V words in the vocabulary, is that of finding satisfactory estimates of δ, Γ and Π, given an observation sequence known to come from an utterance of a particular word. The 'classification problem' is that of deciding which word in the vocabulary a given observation sequence corresponds to. One way of performing the classification is to compute the probability of the observation sequence for each of the V hidden Markov models derived at the training stage (one for each word), and to choose the word in the vocabulary which maximizes this probability.

If the parameters δ, Γ and Π are to be estimated by maximum likelihood and the classification performed by the above method, the classification and training problems reduce to the evaluation, and maximization with respect to δ, Γ and Π, of the likelihood

$$\mathrm{P}(S_1 = s_1, S_2 = s_2, \ldots, S_T = s_T),$$

which will be denoted by L_T. In this context the likelihood is usually evaluated by the 'forward-backward' algorithm, and maximum likelihood estimates of the parameters computed by the 'Baum-Welch re-estimation algorithm'. (The terminology varies, however: for instance Nádas (1983) uses the term forward-backward algorithm for the estimation algorithm.) The Baum-Welch algorithm was developed by L.E. Baum and his co-workers in a series of papers published between 1966 and 1972: Baum and Petrie (1966), Baum and Eagon (1967), Baum and Sell (1968), Baum, Petrie et al. (1970), and Baum (1972). The name of Welch seems to appear only as joint author

(with Baum) of a paper listed by Baum, Petrie et al. as submitted for publication. The algorithm is in fact an early example of an algorithm of EM type. We shall first derive the forward-backward and Baum-Welch algorithms, and then indicate how the latter algorithm fits into the EM framework.

We therefore consider $\{S_t\}$ and $\{C_t\}$ as described above, and assume explicitly that, given $C^{(T)} = \{C_t : t = 1, 2, \ldots, T\}$, the random variables $S_1, \ldots, S_T$ are mutually independent and the (conditional) distribution of S_t depends only on C_t and is given by

$$\mathrm{P}(S_t = s \mid C_t = i) = \pi_{si}.$$

The independence assumption is not usually mentioned in the speech processing literature, but it (or an equivalent assumption) is implicit in the derivation of the forward-backward and Baum-Welch algorithms. We shall give a rather complete account of these derivations, because some of the results needed seem not to have been proved in the published literature. (Baum, Petrie et al. refer to the apparently unpublished paper by Baum and Welch for certain of the results.) Furthermore, the expository article of Juang and Rabiner (1991) gives only a brief account (on p. 256) of the relation between the Baum-Welch and EM algorithms, and it therefore seems useful to discuss that relation more fully here.

We begin by stating four properties that are needed. We shall often use a somewhat abbreviated notation, in which for instance the event that $S_t = s_t$ is denoted by S_t. Firstly, for $t = 1, 2, \ldots, T$:

$$\mathrm{P}(S_1, \ldots, S_T \mid C_t) = \mathrm{P}(S_1, \ldots, S_t \mid C_t)\, \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t). \tag{2.1}$$

(In the case $t = T$ we shall use the convention that $\mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t) = 1$.)

Secondly, for $t = 1, 2, \ldots, T-1$:

$$\mathrm{P}(S_1, \ldots, S_T \mid C_t, C_{t+1}) = \mathrm{P}(S_1, \ldots, S_t \mid C_t)\ \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_{t+1}). \quad (2.2)$$

Thirdly, for $1 \leq t \leq l \leq T$:

$$\mathrm{P}(S_l, \ldots, S_T \mid C_t, \ldots, C_l) = \mathrm{P}(S_l, \ldots, S_T \mid C_l). \quad (2.3)$$

Finally, for $t = 1, 2, \ldots, T$:

$$\mathrm{P}(S_t, \ldots, S_T \mid C_t) = \mathrm{P}(S_t \mid C_t)\ \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t). \quad (2.4)$$

These properties are all rather intuitive, and we defer the proofs to Appendix A.

The forward-backward and Baum-Welch algorithms involve the computation of the 'forward probabilities' $\alpha_t(i)$ and 'backward probabilities' $\beta_t(i)$. These are defined as follows for all states i of the Markov chain, and all t from 1 to T:

$$\alpha_t(i) = \mathrm{P}(S_1 = s_1, \ldots, S_t = s_t, C_t = i), \text{ and}$$

$$\beta_t(i) = \mathrm{P}(S_{t+1} = s_{t+1}, \ldots, S_T = s_T \mid C_t = i).$$

(The convention noted above implies that $\beta_T(i) = 1$ for all i.) From these definitions and property (2.1) we have, for $t = 1, 2, \ldots, T$:

$$\begin{aligned} \alpha_t(i)\beta_t(i) &= \mathrm{P}(C_t = i)\mathrm{P}(S_1, \ldots, S_t \mid C_t = i)\mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t = i) \\ &= \mathrm{P}(C_t = i)\mathrm{P}(S_1, \ldots, S_T \mid C_t = i) \\ &= \mathrm{P}(S_1, \ldots, S_T, C_t = i), \end{aligned} \quad (2.5)$$

and

$$\sum_{i=1}^{m} \alpha_t(i)\beta_t(i) = \mathrm{P}(S_1, \ldots, S_T) = L_T. \quad (2.6)$$

Hence, if we can evaluate the vectors α_t and β_t for all t, we have available T different ways of computing the likelihood. For instance, setting $t = T$

yields $L_T = \sum_{i=1}^{m} \alpha_T(i)$, the formula usually quoted in the speech processing literature.

In order to find α_t and β_t we note that $\beta_T(i) = 1$ and

$$\alpha_1(i) = \mathrm{P}(C_1 = i)\mathrm{P}(S_1 = s_1 \mid C_1 = i) = \delta_i \pi_{s_1 i},$$

and we use these values to start the two recursions derived below, which are valid for $1 \leq t \leq T-1$. Firstly, by using property (2.2), we have

$$\begin{aligned}
\alpha_{t+1}(j) &= \sum_{i=1}^{m} \mathrm{P}(S_1, \ldots, S_{t+1}, C_t = i, C_{t+1} = j) \\
&= \sum \mathrm{P}(C_t = i, C_{t+1} = j)\mathrm{P}(S_1, \ldots, S_{t+1} \mid C_t = i, C_{t+1} = j) \\
&= \sum \mathrm{P}(C_t = i)\gamma_{ij}\mathrm{P}(S_1, \ldots, S_t \mid C_t = i)\mathrm{P}(S_{t+1} \mid C_{t+1} = j) \\
&= \sum \mathrm{P}(S_1, \ldots, S_t, C_t = i)\gamma_{ij}\pi_{s_{t+1}j} \\
&= \left(\sum_{i=1}^{m} \alpha_t(i)\gamma_{ij} \right) \pi_{s_{t+1}j}.
\end{aligned}$$

Secondly, by using property (2.3) (with $l = t+1$) and property (2.4), we have

$$\begin{aligned}
\beta_t(i) &= \sum_{j=1}^{m} \mathrm{P}(S_{t+1}, \ldots, S_T, C_t = i, C_{t+1} = j)/\mathrm{P}(C_t = i) \\
&= \sum \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t = i, C_{t+1} = j)\mathrm{P}(C_t = i, C_{t+1} = j)/\mathrm{P}(C_t = i) \\
&= \sum \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_{t+1} = j)\gamma_{ij} \\
&= \sum \mathrm{P}(S_{t+1} \mid C_{t+1} = j)\mathrm{P}(S_{t+2}, \ldots, S_T \mid C_{t+1} = j)\gamma_{ij} \\
&= \sum_{j=1}^{m} \pi_{s_{t+1}j}\beta_{t+1}(j)\gamma_{ij}.
\end{aligned}$$

As Levinson et al. point out, the above recursions can be stated more succinctly in matrix notation. A matrix expression for the likelihood will be derived in section 2.5, and its uses described in sections 2.6 and 2.7.

We now discuss the rationale of the three 're-estimation formulae' which constitute the Baum-Welch algorithm for improving estimates of δ, Π and Γ. We define in passing some of the notation used by Levinson et al. (e_i, d_{jk} and c_{ij}), as this will be useful in due course. By equation (2.5), the probability that the initial state is i, given the observations, is just

$$\begin{aligned} \mathrm{P}(C_1{=}i \mid S_1,\ldots,S_T) &= \mathrm{P}(S_1,\ldots,S_T,C_1{=}i)/L_T \\ &= \alpha_1(i)\beta_1(i)/L_T. \end{aligned} \tag{2.7}$$

This suggests that this last quantity, computed on the basis of the current estimates of the parameters, may provide an improved estimate of δ_i. (In Levinson's notation the 're-estimate' (2.7) is e_i/L_T, or equivalently $e_i/\sum_{i=1}^m e_i$.)

To motivate an estimate for π_{kj}, we note that, given the observations, the expected number of occurrences of state j is

$$\sum_{t=1}^{T} \mathrm{P}(C_t{=}j \mid S_1,\ldots,S_T) = \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)/L_T.$$

The expected number of such occurrences for which $S_t = k$ is

$$\begin{aligned} \sum_{t=1}^{T} \mathrm{P}(S_t{=}k, C_t{=}j \mid S_1,\ldots,S_T) &= \sum_{\{t:s_t=k\}} \mathrm{P}(C_t{=}j \mid S_1,\ldots,S_T) \\ &= \sum_{\{t:s_t=k\}} \alpha_t(j)\beta_t(j)/L_T. \end{aligned}$$

(In Levinson's notation this is d_{jk}/L_T.) The ratio

$$\frac{\sum_{\{t:s_t=k\}} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} = \frac{d_{jk}}{\sum_{k=1}^{n} d_{jk}} \tag{2.8}$$

is therefore the expected proportion of the occurrences of state j for which the corresponding observation is k. This expression, evaluated at the current parameter estimates, provides a new estimate of π_{kj}.

In similar fashion we see that the expected number of transitions out of state i, given the observations, is

$$\sum_{t=1}^{T-1} \mathrm{P}(C_t=i \mid S_1,\ldots,S_T) = \sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)/L_T, \tag{2.9}$$

and the expected number from i to j is

$$\begin{aligned} &\sum_{t=1}^{T-1} \mathrm{P}(C_t=i, C_{t+1}=j \mid S_1,\ldots,S_T) \qquad (2.10)\\ = \; & L_T^{-1} \sum_{t=1}^{T-1} \mathrm{P}(S_1,\ldots,S_T \mid C_t=i, C_{t+1}=j)\mathrm{P}(C_t=i)\gamma_{ij}. \end{aligned}$$

But properties (2.2) and (2.4), and the definitions of α_t and β_{t+1}, imply that

$$\begin{aligned} & \mathrm{P}(S_1,\ldots,S_T \mid C_t=i, C_{t+1}=j)\\ = \; & \mathrm{P}(S_1,\ldots,S_t \mid C_t=i)\mathrm{P}(S_{t+1},\ldots,S_T \mid C_{t+1}=j)\\ = \; & \mathrm{P}(S_1,\ldots,S_t \mid C_t=i)\mathrm{P}(S_{t+1} \mid C_{t+1}=j)\mathrm{P}(S_{t+2},\ldots,S_T \mid C_{t+1}=j)\\ = \; & (\alpha_t(i)/\mathrm{P}(C_t=i))\,\pi_{s_{t+1}j}\beta_{t+1}(j). \end{aligned}$$

Hence, finally, the expected number of transitions from i to j is

$$\gamma_{ij} \sum_{t=1}^{T-1} \alpha_t(i)\pi_{s_{t+1}j}\beta_{t+1}(j)/L_T.$$

(In Levinson's notation this is c_{ij}/L_T.) The resulting new estimate of γ_{ij} is the following ratio, evaluated at the current parameter estimates:

$$\frac{\gamma_{ij}\sum_{t=1}^{T-1} \alpha_t(i)\pi_{s_{t+1}j}\beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} = \frac{c_{ij}}{\sum_{j=1}^{m} c_{ij}}. \tag{2.11}$$

It turns out that the re-estimates (2.7), (2.8) and (2.11) strictly increase the value of L_T, the likelihood function, except at critical points of L_T: we now describe in outline how this may be established.

The concavity of the log function implies that

$$\log\left(L_T(\bar{\lambda})/L_T(\lambda)\right) \geq \left(Q(\lambda,\bar{\lambda}) - Q(\lambda,\lambda)\right)/L_T(\lambda)$$

for a certain function Q. Here $L_T(\lambda)$ denotes the likelihood of the observations $s_1, \ldots, s_T$ evaluated at parameter values $\lambda = (\delta, \Pi, \Gamma)$, and similarly $L_T(\bar{\lambda})$ for $\bar{\lambda} = (\bar{\delta}, \bar{\Pi}, \bar{\Gamma})$. Hence replacing λ by any $\bar{\lambda}$ such that $Q(\lambda, \bar{\lambda}) > Q(\lambda, \lambda)$ will increase the likelihood. Levinson et al. show that $Q(\lambda, \bar{\lambda})$ is maximized (as a function of $\bar{\lambda}$) by setting $\bar{\lambda}$ equal to $\hat{\lambda} = \left(\hat{\delta}, \hat{\Pi}, \hat{\Gamma}\right)$, where:

$$\hat{\delta}_i = e_i / \textstyle\sum_{i=1}^{m} e_i$$
$$\hat{\pi}_{kj} = d_{jk} / \textstyle\sum_{k=1}^{n} d_{jk}$$
$$\hat{\gamma}_{ij} = c_{ij} / \textstyle\sum_{j=1}^{m} c_{ij}.$$

These are precisely the Baum-Welch re-estimates given in equations (2.7), (2.8) and (2.11). Furthermore, as Baum, Petrie et al. show under mild assumptions (see p. 166 of their paper), the strict inequality $L_T(\hat{\lambda}) > L_T(\lambda)$ holds except when λ is a critical point of the likelihood, in which case $\hat{\lambda} = \lambda$. The Baum-Welch algorithm therefore guarantees an improvement in the likelihood except at a critical point of the likelihood.

The function Q referred to above is given by

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{m} e_i \log \bar{\delta}_i + \sum_{j=1}^{m} \sum_{k=1}^{n} d_{jk} \log \bar{\pi}_{kj} + \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} \log \bar{\gamma}_{ij}$$

for all λ and $\bar{\lambda}$. This is the function maximized by taking $\bar{\lambda}$ to be $\hat{\lambda}$, the Baum-Welch re-estimates. Examination of this expression reveals that the Baum-Welch algorithm, because it proceeds by maximizing $Q(\lambda, \bar{\lambda})$ as a function of $\bar{\lambda}$, is an example of the EM algorithm (as described on p. 130 of Little and Rubin (1987)). In the present context the 'missing data' are the states $i_1, \ldots, i_T$ occupied by the Markov chain, and the 'complete data' are $s_1, \ldots, s_T, i_1, \ldots, i_T$. $Q(\lambda, \bar{\lambda})/L_T(\lambda)$ is just the complete-data log-likelihood, evaluated on the basis of the current parameter estimates λ, with those functions of the missing data which appear in it replaced by their conditional expectations given the observations. These conditional expectations

are: $e_i/L_T(\lambda)$, the expectation of the number of times (0 or 1) that $C_1 = i$; $d_{jk}/L_T(\lambda)$, the expected number of times the observation is k when the state is j; and $c_{ij}/L_T(\lambda)$, the expected number of transitions from state i to state j.

Because of underflow and other problems, it is not possible to implement the forward-backward and Baum-Welch algorithms exactly as described above. Levinson et al. indicate how these problems may be overcome in practice, e.g. the use of scaling to prevent underflow in the computation of the forward and backward probabilities. We shall discuss scaling in some detail in section 2.5.

2.3 Hidden Markov time series models: definition and notation

In this section we introduce the time series models which will be studied in detail in the rest of the chapter. The notation is essentially that of the preceding section, but for ease of reference we define it in full here.

Let $\{C_t : t \in \mathbf{N}\}$ be an irreducible homogeneous Markov chain on the state-space $\{1, 2, \ldots, m\}$, with transition probability matrix Γ. That is, $\Gamma = (\gamma_{ij})$, where for all states i and j and times t:

$$\gamma_{ij} = \mathrm{P}(C_t = j \mid C_{t-1} = i).$$

By the irreducibility of $\{C_t\}$, there exists a unique, strictly positive, stationary distribution, which we shall denote by the vector $\delta = (\delta_1 \ \delta_2 \ \ldots \ \delta_m)$. Suppose throughout that $\{C_t\}$ is stationary, so that δ is for all t the distribution of C_t. (In this respect, the stationarity of the Markov chain, the models we consider here differ from the speech-processing models described in section 2.2, and from those of Albert (1991) and those of Leroux and Puterman

(1992). Another respect in which our models differ from those used in speech processing is that the state-space of the observations is not here assumed to be finite in general.)

Now let the nonnegative integer-valued random process $\{S_t : t \in \mathbf{N}\}$ be such that, conditional on $C^{(T)} = \{C_t : t = 1, \ldots, T\}$, the random variables $\{S_t : t = 1, \ldots, T\}$ are mutually independent and, if $C_t = i$, S_t takes the value s with probability ${}_t\pi_{si}$. That is, for $t = 1, \ldots, T$, the distribution of S_t conditional on $C^{(T)}$ is given by

$$\mathrm{P}(S_t = s \mid C_t = i) = {}_t\pi_{si}.$$

We shall refer to the probabilities ${}_t\pi_{si}$ as the 'state-dependent probabilities'. When these probabilities ${}_t\pi_{si}$ do not depend on t, the subscript t will be omitted. The two cases we shall discuss in detail are: (i) the conditional distribution of S_t is Poisson; and (ii) the conditional distribution of S_t is binomial.

In case (i), let the conditional mean of S_t be

$$\mu(t) = \sum_{i=1}^{m} \lambda_i Z_i(t),$$

where the random variable $Z_i(t)$ is the indicator of the event $\{C_t = i\}$. Then if $C_t = i$, S_t has a Poisson distribution with mean λ_i, and the state-dependent probabilities are given for all nonnegative integers s by:

$$\pi_{si} = e^{-\lambda_i} \lambda_i^s / s!.$$

In case (ii), let the conditional binomial distribution of S_t have parameters n_t and $p(t)$, where n_t is a known positive integer and

$$p(t) = \sum_{i=1}^{m} p_i Z_i(t).$$

Hence, for $s = 0, 1, \ldots, n_t$:

$$ {}_t\pi_{si} = \binom{n_t}{s} p_i^s (1 - p_i)^{n_t - s}. $$

We shall refer to the models $\{S_t\}$ thus defined as Poisson- and binomial-hidden Markov models. In each case there are m^2 parameters: m parameters λ_i or p_i, and $m^2 - m$ transition probabilities γ_{ij}, e.g. the off-diagonal elements of Γ, to specify the 'hidden Markov chain' $\{C_t\}$. We note, however, that the off-diagonal elements must satisfy, in addition to the obvious nonnegativity constraints, the m constraints $\sum_{j \neq i} \gamma_{ij} \leq 1$, one for each i. The only constraints on λ_i and p_i are $\lambda_i \geq 0$ and $0 \leq p_i \leq 1$.

The special case $m = 2$ is of particular interest in practice: if $m = 2$ we write the transition probability matrix of $\{C_t\}$ as

$$ \Gamma = \begin{pmatrix} 1 - \gamma_1 & \gamma_1 \\ \gamma_2 & 1 - \gamma_2 \end{pmatrix}, $$

and it follows that

$$ \delta = \frac{1}{\gamma_1 + \gamma_2} \begin{pmatrix} \gamma_2 & \gamma_1 \end{pmatrix}. $$

(Note that γ_1 and γ_2 are strictly positive by the irreducibility assumption, and that the two constraints $\sum_{j \neq i} \gamma_{ij} \leq 1$ reduce to $\gamma_i \leq 1$, $i = 1, 2$.) The following expression for Γ^k, obtained by diagonalizing Γ, will be useful in deriving the properties of $\{S_t\}$ when $m = 2$:

$$ \Gamma^k = \begin{pmatrix} \delta_1 & \delta_2 \\ \delta_1 & \delta_2 \end{pmatrix} + w^k \begin{pmatrix} \delta_2 & -\delta_2 \\ -\delta_1 & \delta_1 \end{pmatrix}, \tag{2.12} $$

where $w = 1 - \gamma_1 - \gamma_2$.

We propose to show in this chapter and the next that the models $\{S_t\}$ defined above (and certain modifications and generalizations thereof) have

properties which make them suitable as models for a wide range of discrete-valued time series. Although we have chosen to concentrate on the Poisson or binomial as the conditional distribution of the observations $\{S_t\}$, it will be clear that other discrete distributions may similarly be used. In fact a continuous distribution could be used if continuous-valued hidden Markov models were required: McInnes and Jack (1988, p. 32) report that such continuous-valued models have been applied more successfully in speech recognition technology than discrete-valued models of the type described in the previous section.

2.4 Correlation properties

One of the main characteristics of interest relating to a time series model is its autocorrelation function. In this section we derive inter alia expressions for the autocorrelations of the models defined in section 2.3. As a preliminary we state two results which will generally be of use in deriving moment properties for a hidden Markov model $\{S_t\}$. Firstly, provided the relevant expectations exist,

$$\mathrm{E}\,(f(S_t)) = \sum_{i=1}^{m} \mathrm{E}\,(f(S_t) \mid C_t{=}i)\,\delta_i. \tag{2.13}$$

This is proved by conditioning on C_t and noting that $\delta_i = \mathrm{P}(C_t{=}i)$. Secondly, provided again that the relevant expectations exist, we have for $k \in \mathbf{N}$ that

$$\mathrm{E}\,(f(S_t, S_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(f(S_t, S_{t+k}) \mid C_t{=}i, C_{t+k}{=}j)\delta_i\gamma_{ij}(k), \tag{2.14}$$

where $\gamma_{ij}(k) = (\Gamma^k)_{ij}$. To prove this, we condition on $C^{(t+k)}$ and exploit the fact that the conditional expectation of $f(S_t, S_{t+k})$, given $C^{(t+k)}$, is the conditional expectation given only C_t and C_{t+k}. Summing $\mathrm{P}(C_1, \ldots, C_{t+k})$ over the states at all times other than t and $t+k$ gives $\mathrm{P}(C_t, C_{t+k})$, and we

then get

$$\mathrm{E}(f(S_t, S_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(f(S_t, S_{t+k}) \mid C_t = i, C_{t+k} = j)\mathrm{P}(C_t = i, C_{t+k} = j).$$

The result (2.14) follows, since $\mathrm{P}(C_t = i, C_{t+k} = j) = \delta_i \gamma_{ij}(k)$.

2.4.1 The autocorrelation function of a Poisson-hidden Markov model

Let $\{S_t\}$ be a Poisson-hidden Markov model, as defined in section 2.3. By equation (2.13) the mean is given by

$$\mathrm{E}(S_t) = \sum_{i=1}^{m} \mathrm{E}(S_t \mid C_t = i)\delta_i = \sum_{i=1}^{m} \lambda_i \delta_i = \delta\lambda',$$

where λ is defined as $(\lambda_1, \ldots, \lambda_m)$. To derive the variance we need $\mathrm{E}(S_t^2)$:

$$\mathrm{E}(S_t^2) = \sum_{i=1}^{m} \mathrm{E}(S_t^2 \mid C_t = i)\delta_i = \sum_{i=1}^{m} (\lambda_i^2 + \lambda_i)\delta_i.$$

It then follows that

$$\begin{aligned} \mathrm{Var}(S_t) &= \sum(\lambda_i^2 + \lambda_i)\delta_i - (\sum \lambda_i \delta_i)^2 \\ &= \lambda D \lambda' + \delta\lambda' - (\delta\lambda')^2, \end{aligned}$$

with D denoting $\mathrm{diag}(\delta)$. Alternatively, defining Λ as $\mathrm{diag}(\lambda)$, we may write this result as

$$\mathrm{Var}(S_t) = \delta\Lambda\lambda' + \delta\lambda' - (\delta\lambda')^2.$$

To find the covariance we note that, for $k \in \mathbf{N}$ but not $k = 0$,

$$\mathrm{E}(S_t S_{t+k} \mid C_t = i, C_{t+k} = j) = \lambda_i \lambda_j.$$

Hence by equation (2.14)

$$\mathrm{E}(S_t S_{t+k}) = \sum_{i,j=1}^{m} \lambda_i \lambda_j \delta_i \gamma_{ij}(k) = \delta\Lambda\Gamma^k\lambda'.$$

The covariance is therefore

$$\text{Cov}(S_t, S_{t+k}) = \delta\Lambda\Gamma^k\lambda' - (\delta\lambda')^2$$

and the ACF

$$\begin{aligned}\rho_k &= \text{Corr}(S_t, S_{t+k}) \\ &= \frac{\text{Cov}(S_t, S_{t+k})}{\text{Var}(S_t)} \\ &= \frac{\delta\Lambda\Gamma^k\lambda' - (\delta\lambda')^2}{\delta\Lambda\lambda' + \delta\lambda' - (\delta\lambda')^2}.\end{aligned}$$

This expression for ρ_k is valid for all $k \in \mathbf{N}$. It is interesting to note that ρ_k depends on k only through Γ^k.

An alternative proof of the above results for the mean, variance and ACF proceeds via the conditional mean (and variance) $\mu(t) = \sum_{i=1}^{m} \lambda_i Z_i(t)$. The main steps in the proof are:

$$\begin{aligned}\text{E}(S_t) &= \text{E}(\mu(t)) \\ \text{E}(S_t^2) &= \text{E}(\mu(t)^2 + \mu(t)) \\ \text{E}(S_t S_{t+k}) &= \text{E}(\mu(t)\mu(t+k)), \text{ and} \\ \text{Cov}(S_t, S_{t+k}) &= \text{Cov}(\mu(t), \mu(t+k)).\end{aligned}$$

As before, the result for the covariance is valid for $k \in \mathbf{N}$, but not $k = 0$, relying as it does on the conditional independence of S_t and S_{t+k}.

In the special case $m = 2$, i.e. if the hidden Markov chain has only two states, we make use of expression (2.12) for Γ^k to conclude that, for $k \in \mathbf{N}$,

$$\begin{aligned}\rho_k &= \frac{\delta_1\delta_2(\lambda_2 - \lambda_1)^2}{\delta_1\delta_2(\lambda_2 - \lambda_1)^2 + \delta\lambda'} w^k \\ &= Aw^k,\end{aligned}$$

where $w = 1 - \gamma_1 - \gamma_2$. This expression for ρ_k could also have been arrived at by applying equation (10) on p. 196 of Cox and Lewis (1966). By taking

$S_1, S_2, \ldots$ to be the interevent intervals in a point process on the line, one generates a (non-orderly) semi-Markov process with two types of interval. Equation (10) gives the correlation between intervals in such a process, i.e. the correlation between S_t and S_{t+k}. (Although the result of Cox and Lewis apparently refers to the case of continuous-valued intervals, its derivation applies equally to the discrete case.) If $\lambda_1 \neq \lambda_2$, the autocorrelation can be written as

$$\rho_k = \left(1 + \frac{\delta\lambda'}{(\lambda_2 - \lambda_1)^2 \delta_1 \delta_2}\right)^{-1} w^k.$$

Since if $m = 2$ the autocorrelation function for the Markov chain $\{C_t\}$ is w^k (see section 1.2), the above two formulas for ρ_k display clearly the effect of the extra level of randomness present in a hidden Markov model: the autocorrelations are reduced by the factor A, which lies between 0 and 1. If $\lambda_1 = \lambda_2$, A attains the bound of 0, and S_t and S_{t+k} are uncorrelated. This is to be expected: if, conditional on $\{C_t\}$, S_t and S_{t+k} are independent and have distributions unaffected by the current state of $\{C_t\}$, relaxing the conditioning on $\{C_t\}$ will not induce any correlation between S_t and S_{t+k}. The upper bound of 1 may be approached, for example, by fixing λ_1 and letting λ_2 approach infinity. These observations, and the fact that w can be arbitrarily close to 1 (or -1), suggest that a wide range of correlations can be attained by at least the two-state models under discussion. In fact, given ϵ and η lying strictly between 0 and 1, the four parameters $\gamma_1, \gamma_2, \lambda_1$ and λ_2 can be chosen in such a way that the autocorrelation function reduces to $(1 - \eta)(1 - \epsilon)^k$: one possibility is to take $\gamma_1 = \gamma_2 = \epsilon/2$ and

$$\lambda_2 = \lambda_1 + (\nu - 1)^{-1}\{1 + (4\lambda_1(\nu - 1) + 1)^{1/2}\},$$

where $\nu = (1 - \eta)^{-1}$. (Note that $\nu > 1$.) Similarly one can choose the four parameters in such a way that the autocorrelation function reduces to

$(1-\eta)(-1+\epsilon)^k$. Hence any autocorrelation function of the form

$$Aw^k \quad (0 < A < 1,\ -1 < w < 1)$$

is possible in the case $m = 2$.

For general m we recall from section 1.2 that 1 is a simple eigenvalue of the transition probability matrix Γ, and Γ can always be written in Jordan canonical form as $\Gamma = U\Omega U^{-1}$, where U, U^{-1} and Ω are of the following forms: $U = (\mathbf{1}'\ R)$, $U^{-1} = \begin{pmatrix} \delta \\ W \end{pmatrix}$ and $\Omega = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0}' & \Psi \end{pmatrix}$. Hence

$$\Gamma^k = U\Omega^k U^{-1} = \mathbf{1}'\delta + R\Psi^k W$$

and, for $k \in \mathbf{N}$,

$$\begin{aligned} \mathrm{Cov}(S_t, S_{t+k}) &= \delta\Lambda(\mathbf{1}'\delta + R\Psi^k W)\lambda' - (\delta\lambda')^2 \\ &= (\delta\Lambda R)\Psi^k(W\lambda'). \end{aligned} \tag{2.15}$$

The covariance, and therefore the ACF, are thus seen to involve powers up to the kth of the eigenvalues of Γ. If all the quantities λ_i happen to be equal, i.e. if the conditional mean of S_t is unaffected by the value of C_t, the covariance and correlation are zero. This may be demonstrated as follows. Since $\begin{pmatrix} \delta \\ W \end{pmatrix} (\mathbf{1}'\ R) = U^{-1}U = I$, it follows that $W\mathbf{1}' = \mathbf{0}'$. Hence, if λ is a multiple of $\mathbf{1}$, $W\lambda' = \mathbf{0}'$ and the expression (2.15) reduces to zero.

If Γ is in fact diagonalizable, we have $\Omega = \mathrm{diag}(1, \omega_2, \ldots, \omega_m)$, the columns of U are right eigenvectors of Γ and the rows of U^{-1} left eigenvectors. In this case the covariance of S_t and S_{t+k} is more simply expressed as

$$\begin{aligned} \mathrm{Cov}(S_t, S_{t+k}) &= (\delta\Lambda U)\Omega^k(U^{-1}\lambda') - (\delta\lambda')^2 \\ &= c\Omega^k d' - c_1 d_1 \\ &= \sum_{i=2}^{m} c_i \omega_i^k d_i, \end{aligned}$$

where we have defined $c = \delta\Lambda U$ and $d' = U^{-1}\lambda'$. (Note that $c_1 = \delta\Lambda\mathbf{1}' = \delta\lambda'$ and $d_1 = \delta\lambda'$.) This expression for the covariance is valid for all $k \in \mathbf{N}$, but not for $k = 0$. If Γ is diagonalizable, therefore, the ACF is a linear combination of the kth powers of $\omega_2, \ldots, \omega_m$. These quantities ω_i, which are not necessarily distinct, are the eigenvalues of Γ other than 1. In modulus they may equal 1, but assuming aperiodicity of $\{C_t\}$ would rule out that possibility and give us the strict inequality $|\omega_i| < 1$ for all $i \geq 2$.

We illustrate the case of Γ nondiagonalizable by the following example.

Example Let the Markov chain $\{C_t\}$ have transition probability matrix

$$\Gamma = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

This matrix is not diagonalizable, but a Jordan canonical form, as described above, is

$$\begin{aligned} \Gamma &= \begin{pmatrix} 1 & -1 & -9 \\ 1 & -1 & 15 \\ 1 & 3 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1/3 & 1 \\ 0 & 0 & -1/3 \end{pmatrix} \frac{1}{96} \begin{pmatrix} 45 & 27 & 24 \\ -15 & -9 & 24 \\ -4 & 4 & 0 \end{pmatrix} \\ &= U\Omega U^{-1}. \end{aligned}$$

Note that the first row of U^{-1} is the stationary distribution, δ. In the notation used above

$$R = \begin{pmatrix} -1 & -9 \\ -1 & 15 \\ 3 & 0 \end{pmatrix}, \; W = \tfrac{1}{96} \begin{pmatrix} -15 & -9 & 24 \\ -4 & 4 & 0 \end{pmatrix} \text{ and } \Psi = \begin{pmatrix} -1/3 & 1 \\ 0 & -1/3 \end{pmatrix}.$$

From equation (2.15) the covariance of S_t and S_{t+k} is $(\delta\Lambda R)\Psi^k(W\lambda')$. Since $\Psi = -\frac{1}{3}I + N$, where N is the nilpotent matrix $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, it follows that

$\Psi^k = \begin{pmatrix} (-1/3)^k & k(-1/3)^{k-1} \\ 0 & (-1/3)^k \end{pmatrix}$. For $k \in \mathbf{N}$, $\rho_k = \mathrm{Corr}(S_t, S_{t+k})$ is therefore the following linear combination of $(-1/3)^k$ and $k(-1/3)^{k-1}$:

$$\frac{(-\frac{1}{3})^k\{3(-5\lambda_1-3\lambda_2+8\lambda_3)^2+180(\lambda_2-\lambda_1)^2\}+k(-\frac{1}{3})^{k-1}\{4(-5\lambda_1-3\lambda_2+8\lambda_3)(\lambda_2-\lambda_1)\}}{32\{15(\lambda_1^2+\lambda_1)+9(\lambda_2^2+\lambda_2)+8(\lambda_3^2+\lambda_3)\}-\{15\lambda_1+9\lambda_2+8\lambda_3\}^2}.$$

2.4.2 The autocorrelation function of a binomial-hidden Markov model

Let $\{S_t\}$ be a binomial-hidden Markov model. The conditional distribution of S_t is binomial with parameters n_t and $p(t) = \sum_{i=1}^m p_i Z_i(t)$. The derivation of the mean, variance and covariance function is very similar to that in the Poisson case, and will be described fairly briefly. We shall assume throughout this subsection that Γ is diagonalizable, since that is the case of most practical interest and the modifications necessary if Γ is nondiagonalizable are quite analogous to those already described in subsection 2.4.1. We use the notation $p = (p_1\ p_2\ \ldots\ p_m)$, $P = \mathrm{diag}(p)$, and $D = \mathrm{diag}(\delta)$ as in subsection 2.4.1.

From equation (2.13) we have:

$$\mathrm{E}(S_t) = \sum_{i=1}^m (n_t p_i)\delta_i = n_t \delta p'$$

and

$$\begin{aligned} \mathrm{E}(S_t^2) &= \sum_{i=1}^m \left(n_t p_i(1-p_i) + n_t^2 p_i^2\right)\delta_i \\ &= n_t\delta p' + n_t(n_t-1)pDp' \\ &= n_t\delta p' + n_t(n_t-1)\delta Pp'. \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{Var}(S_t) &= n_t(n_t-1)\delta Pp' + n_t\delta p' - n_t^2(\delta p')^2 \\ &= n_t^2\left(\delta Pp' - (\delta p')^2\right) + n_t(\delta p' - \delta Pp'). \end{aligned}$$

From equation (2.14) we have, for $k \in \mathbf{N}$,

$$\mathrm{E}(S_t S_{t+k}) = \sum_{i,j=1}^{m} (n_t p_i)(n_{t+k} p_j)\delta_i \gamma_{ij}(k) = n_t n_{t+k} \delta P \Gamma^k p'.$$

Hence the result for the covariance is

$$\mathrm{Cov}(S_t, S_{t+k}) = n_t n_{t+k} \left(\delta P \Gamma^k p' - (\delta p')^2\right).$$

Alternatively we could have proceeded via the following steps:

$$\begin{aligned}
\mathrm{E}(S_t) &= n_t \mathrm{E}(p(t)) \\
\mathrm{E}(S_t^2) &= \mathrm{E}(n_t p(t)(1 - p(t)) + n_t^2 p(t)^2) \\
\mathrm{E}(S_t S_{t+k}) &= n_t n_{t+k} \mathrm{E}(p(t)p(t+k)), \text{ and} \\
\mathrm{Cov}(S_t, S_{t+k}) &= n_t n_{t+k} \mathrm{Cov}(p(t), p(t+k))
\end{aligned}$$

and arrived at the same conclusions.

The autocorrelation is therefore given by:

$$\mathrm{Corr}(S_t, S_{t+k}) = \frac{n_t n_{t+k} \left(\delta P \Gamma^k p' - (\delta p')^2\right)}{(\mathrm{Var}\ S_t\ \mathrm{Var}\ S_{t+k})^{1/2}}.$$

Provided that $\delta P p' - (\delta p')^2 = \mathrm{Var}\ p(t)$ is nonzero, we can define

$$\alpha = \frac{\delta p' - \delta P p'}{\delta P p' - (\delta p')^2}$$

and write the autocorrelation as

$$(1 + \alpha/n_t)^{-1/2}(1 + \alpha/n_{t+k})^{-1/2} \frac{\delta P \Gamma^k p' - (\delta p')^2}{\delta P p' - (\delta p')^2}. \tag{2.16}$$

Since it can be shown that $\delta p' \geq \delta P p'$, it follows that α is nonnegative, and the factors $(1 + \alpha/n_t)^{-1/2}$ and $(1 + \alpha/n_{t+k})^{-1/2}$ are at most 1. If n_t is constant with respect to t, the expression (2.16) shows that the autocorrelation depends on k only through Γ^k (as in the Poisson case).

It is of interest here also to find a simple expression for the autocorrelation in the case $m = 2$, i.e. if the Markov chain has only two states. In that case, with w denoting $1 - \gamma_1 - \gamma_2$ as usual,

$$\delta P p' - (\delta p')^2 = \delta_1 \delta_2 (p_2 - p_1)^2,$$

$$\delta P \Gamma^k p' - (\delta p')^2 = \delta_1 \delta_2 (p_2 - p_1)^2 w^k,$$

and

$$\delta p' - \delta P p' = \delta_1 p_1 (1 - p_1) + \delta_2 p_2 (1 - p_2).$$

Hence the autocorrelation is just $(1 + \alpha/n_t)^{-1/2}(1 + \alpha/n_{t+k})^{-1/2} w^k$, where α is given by

$$\frac{\delta_1 p_1 (1 - p_1) + \delta_2 p_2 (1 - p_2)}{\delta_1 \delta_2 (p_2 - p_1)^2}.$$

If n_t is constant with respect to t, we have $\rho_k = (1 + \alpha/n)^{-1} w^k$, which could also have been deduced from equation (10) on p. 196 of Cox and Lewis (1966). Again we see the correlation-reducing effect of the extra level of randomness imposed on top of the Markov chain $\{C_t\}$. But if $p_1 = 0$ and $p_2 = 1$ (and n_t is general) $\alpha = 0$ and the autocorrelation is exactly w^k: the hidden Markov model collapses to a Markov chain. The other extreme is that $p_1 = p_2$: in that case the autocorrelation is zero. If $n_t = n$ for all t we can, as in the case of a Poisson-HM model, achieve an autocorrelation of $(1 - \eta)(1 - \epsilon)^k$ or $(1 - \eta)(-1 + \epsilon)^k$ for given η and ϵ lying strictly between 0 and 1. To do so we can, for instance, let γ_1 and γ_2 both equal $\epsilon/2$ or $1 - \epsilon/2$, and take $p_1 = 0$ and $p_2 = (1 + n(\nu - 1)/2)^{-1}$, where ν (> 1) denotes $(1 - \eta)^{-1}$. The reduction factor $(1 + \alpha/n)^{-1}$ is then $1 - \eta$, as required, and $w = 1 - \epsilon$ or $-1 + \epsilon$. Hence any ACF of the form

$$A w^k \quad (0 < A \leq 1, -1 < w < 1)$$

is possible in the case of n_t constant and $m = 2$.

For general m and Γ diagonalizable we can express the covariance of S_t and S_{t+k} in terms of the kth powers of the eigenvalues ω_i (other than 1) of Γ. With $c = \delta PU$, $d' = U^{-1}p'$, and U as in subsection 2.4.1, we have for $k \in \mathbf{N}$

$$\mathrm{Cov}(S_t, S_{t+k}) = n_t n_{t+k} \sum_{i=2}^{m} c_i \omega_i^k d_i,$$

and a corresponding expression for the autocorrelation can be written down.

The case $n_t = 1$ (and m general) merits some attention, as it provides models for binary time series and the results derived above simplify to the following:

$$\mathrm{E}(S_t) = \delta p'$$

$$\mathrm{Var}(S_t) = \delta p' - (\delta p')^2$$

$$\mathrm{Corr}(S_t, S_{t+k}) = \frac{\delta P \Gamma^k p' - (\delta p')^2}{\delta p' - (\delta p')^2}.$$

2.4.3 The partial autocorrelation function

As in the case of a stationary Markov chain (see section 1.2), or for that matter any second-order stationary process, it is possible to deduce the partial autocorrelations ϕ_{kk} of a (stationary) binomial- or Poisson-hidden Markov model from the autocorrelations ρ_k by means of $\phi_{kk} = |P_k^*|/|P_k|$. The case of Markov chains warns us, however, that we should not expect any cut-off property except perhaps in hidden Markov models based on a two-state Markov chain. Even in that case, since ρ_k is of the form Aw^k for all $k \geq 1$, we find that $|P_2^*| = \rho_2 - \rho_1^2 = Aw^2 - (Aw)^2$: apart from some degenerate cases, ϕ_{22} is nonzero. It seems therefore that there is no cut-off property which might be useful for model identification purposes.

2.5 Evaluation of the likelihood function

Consider a sequence of observations $s_1, s_2, \ldots, s_T$ assumed to be generated by a hidden Markov model $\{S_t\}$ of either Poisson or binomial type. For several reasons it is important that we should be able to evaluate routinely the likelihood function

$$L_T = \mathrm{P}(S_1 = s_1, S_2 = s_2, \ldots, S_T = s_T).$$

Firstly, it provides the finite-dimensional distributions of the process $\{S_t\}$, hence all the marginal, joint and conditional probabilities associated with the random variables S_1, S_2, etc. Secondly, by maximizing L_T with respect to Γ and λ (or Γ and p, whichever is appropriate) we can estimate the m^2 parameters of the model. In parameter-driven processes, of which hidden Markov models are examples, maximum likelihood estimation is often not possible: see Azzalini (1982), for instance. It is therefore a very pleasant property of the hidden Markov models we are discussing that the likelihood (more precisely, its logarithm) can be evaluated sufficiently fast to permit direct numerical maximization. We first derive an expression for the likelihood in terms of a multiple sum, and then show how this may be rewritten in matrix notation in a way which suggests an efficient algorithm for computing L_T.

Given $C^{(T)}$, the joint probability of $S_1, \ldots, S_T$ is, by the assumed conditional independence, the product of Poisson or binomial probabilities ${}_t\pi_{si}$. More specifically, if we condition on the event $\{C_1 = i_1, \ldots, C_T = i_T\}$, the probability that $S_t = s_t$ for $1 \leq t \leq T$ is ${}_1\pi_{s_1 i_1}\, {}_2\pi_{s_2 i_2} \cdots {}_T\pi_{s_T i_T}$. To relax the conditioning we multiply by

$$\mathrm{P}(C_1 = i_1, \ldots, C_T = i_T) = \delta_{i_1} \gamma_{i_1 i_2} \gamma_{i_2 i_3} \cdots \gamma_{i_{T-1} i_T}$$

and sum for all indices i_t over $\{1, 2, \ldots, m\}$. The result is that

$$L_T = \sum_{i_1=1}^{m} \cdots \sum_{i_T=1}^{m} ({}_1\pi_{s_1 i_1} \cdots {}_T\pi_{s_T i_T})(\delta_{i_1} \gamma_{i_1 i_2} \gamma_{i_2 i_3} \cdots \gamma_{i_{T-1} i_T}). \tag{2.17}$$

As it stands, this expression is of little or no computational use, because it has m^T terms and cannot be evaluated except for very small T. The following rearrangement, however, enables us to write L_T in a more useful form:

$$\begin{aligned} L_T &= \sum \cdots \sum \delta_{i_1} {}_1\pi_{s_1 i_1} \gamma_{i_1 i_2} {}_2\pi_{s_2 i_2} \gamma_{i_2 i_3} \cdots \gamma_{i_{T-1} i_T} {}_T\pi_{s_T i_T} \\ &= \delta \, {}_1\lambda(s_1) \Gamma \, {}_2\lambda(s_2) \Gamma \cdots \Gamma \, {}_T\lambda(s_T) \mathbf{1}', \end{aligned} \tag{2.18}$$

where the matrices ${}_t\lambda(s)$ are defined by

$${}_t\lambda(s) = \operatorname{diag}({}_t\pi_{s1}, \; {}_t\pi_{s2}, \; \ldots, \; {}_t\pi_{sm}).$$

The matrix expression (2.18) for the likelihood, or a very similar one, appears for instance in Levinson et al. (1983, p. 1040) and Zucchini and Guttorp (1991). (As noted earlier in this thesis, the subscript t before a symbol is necessary only in the binomial case, and then only if n_t is not constant with respect to t.) The likelihood can now be written as

$$L_T = a \left(\prod_{t=2}^{T} B_t \right) \mathbf{1}'$$

if we define a and B_t by $a = \delta \, {}_1\lambda(s_1)$ (i.e. $a_j = \delta_j \, {}_1\pi_{s_1 j}$), and $B_t = \Gamma \, {}_t\lambda(s_t) = (\gamma_{ij} \, {}_t\pi_{s_t j})$. To evaluate L_T we can (in principle) just postmultiply a successively by B_2, B_3, etc. and add the elements of the resulting vector: 'in principle' because there is a numerical complication, which will be discussed in the next paragraph. The computational effort involved in such an algorithm is linear in T (as opposed to worse than exponential in the case of formula (2.17)), and quadratic in m. It is essentially the same as the case $t = T$ of the 'forward-backward' algorithm $L_T = \sum_i \alpha_t(i)\beta_t(i)$ described in section 2.2. The case $t = 1$ would correspond to beginning with $\mathbf{1}'$, and

premultiplying it successively by $B_T, \ldots, B_2$ and a. Other choices of t would correspond to other ways yet of 'bracketing' the matrix product $\prod_{t=2}^{T} B_t$.

The numerical complication referred to is that the computation of L_T as described may suffer from underflow even for relatively small values of T: the elements of the vector $u = a \prod_{t=2}^{r} B_t$ held at a particular stage of the algorithm may be too small to be distinguishable from zero, even if the quantities δ_i and π_{si} are all of moderate size. Since the likelihood is additive in form, it is not possible merely to work with logarithms. What can be done, however, and was done in this work, is to scale the vector u at each stage so that the average element is 1, i.e. to divide u by $m^{-1} \sum_{i=1}^{m} u_i$, and accumulate the logarithms of these scale factors. Once the scaled likelihood has been computed thus, the sum of the logs of the scale factors is added to the log of the scaled likelihood. This procedure will avoid underflow in many cases and will yield $\log L_T$. Clearly many variations of this technique are possible: the scale factor could be chosen instead to be the largest element of the vector u, or the sum of the elements. Levinson et al. describe the use of this last possibility in scaling the forward and backward probabilities $\alpha_t(i)$ and $\beta_t(i)$ and computing the log of the likelihood.

From the above discussion it will be apparent that we disagree with Albert (1991), who states on p. 1372 and elsewhere that the evaluation of the likelihood (conditional on the first state occupied by the Markov chain) is computationally infeasible, even for an observation sequence of moderate length and the two-state Markov chain he uses in his models. In order to estimate the parameters he therefore seeks alternatives to direct numerical maximization, and implements a variation of EM involving an approximation at the E-step. In section 4.5 we shall describe some fairly complex models of hidden Markov type which are fitted, by direct numerical maximization of

the unconditional likelihood, to a sequence of more than 35000 observations. (As the evaluation of the conditional likelihood requires less computation than does that of the unconditional likelihood, Albert's claim, if true, would apply a fortiori to the unconditional likelihood we use in this work.)

2.6 Marginal, joint and conditional distributions

Given the values of the parameters Γ and λ (or p) of a hidden Markov model $\{S_t\}$, we are able to find the likelihood

$$L_T = \mathrm{P}(S_1 = s_1, S_2 = s_2, \ldots, S_T = s_T) = \delta\, {}_1\lambda(s_1)\, \Gamma\, {}_2\lambda(s_2)\, \Gamma \cdots {}_T\lambda(s_T)\, \mathbf{1}' \tag{2.19}$$

of an observation sequence $s_1, s_2, \ldots, s_T$. It is largely a routine matter, therefore, to find various distributions of interest associated with the random variables $\{S_t\}$. One of the purposes of this section is to present some of these distributions and discuss questions of statistical interest arising from them. The other purpose is to derive certain probabilities associated with the Markov chain $\{C_t\}$, conditional on observations $s_1, \ldots, s_T$.

We note firstly that some generalizations or modifications of equation (2.19) are easily proved. By considering time points $t, t+1, \ldots, T$ (rather than $1, 2, \ldots, T$) one arrives at

$$\mathrm{P}(S_t = s_t, S_{t+1} = s_{t+1}, \ldots, S_T = s_T) = \delta\, {}_t\lambda(s_t)\, \Gamma\, {}_{t+1}\lambda(s_{t+1})\, \Gamma \cdots {}_T\lambda(s_T)\, \mathbf{1}'. \tag{2.20}$$

Another kind of modification is exemplified by the following two probabilities:

$$\mathrm{P}(S_1 = s, S_3 = u, S_7 = v) \;=\; \sum_{i,j,k} {}_1\pi_{si}\, {}_3\pi_{uj}\, {}_7\pi_{vk} \mathrm{P}(C_1 = i, C_3 = j, C_7 = k)$$

$$
\begin{aligned}
&= \sum_{i,j,k} {}_1\pi_{si}\,{}_3\pi_{uj}\,{}_7\pi_{vk}\delta_i\gamma_{ij}(2)\gamma_{jk}(4) \\
&= \delta\,{}_1\lambda(s)\,\Gamma^2\,{}_3\lambda(u)\,\Gamma^4\,{}_7\lambda(v)\,\mathbf{1}'; \\
\mathrm{P}(S_t=u, S_{t+k}=v) &= \sum_{i,j} {}_t\pi_{ui}\,{}_{t+k}\pi_{vj}\delta_i\gamma_{ij}(k) \\
&= \delta\,{}_t\lambda(u)\,\Gamma^k\,{}_{t+k}\lambda(v)\,\mathbf{1}'. \qquad (2.21)
\end{aligned}
$$

This points to an advantage of hidden Markov models as practical statistical tools: the ease with the often-awkward issue of missing data may be handled. Suppose for instance that only one observation, s_l, is missing from an observation sequence of length T. The likelihood of the observations $s_1, \ldots, s_{l-1}, s_{l+1}, \ldots, s_T$ is

$$
\begin{aligned}
&\mathrm{P}(S_1=s_1, \ldots, S_{l-1}=s_{l-1}, S_{l+1}=s_{l+1}, \ldots, S_T=s_T) \\
= \;& \delta\,{}_1\lambda(s_1)\,\Gamma \cdots \Gamma\,{}_{l-1}\lambda(s_{l-1})\,\Gamma^2\,{}_{l+1}\lambda(s_{l+1})\,\Gamma \cdots \Gamma\,{}_T\lambda(s_T)\mathbf{1}'. \qquad (2.22)
\end{aligned}
$$

The only difference between the expression (2.22) and the standard expression (2.19) for the likelihood of the full observation sequence is that the diagonal matrix ${}_l\lambda(s_l)$ has been replaced by the identity matrix. In evaluating (2.22) one therefore proceeds as usual except that the probabilities ${}_l\pi_{s_l i}$ are all taken to be one. More generally, if several observations are missing, one replaces all the corresponding matrices ${}_t\lambda(s_t)$ by the identity.

To establish the marginal distribution of S_t we use the case $T = t$ of equation (2.20):

$$
\mathrm{P}(S_t=s) = \delta\,{}_t\lambda(s)\,\mathbf{1}' = \sum_{i=1}^{m} \delta_i\,{}_t\pi_{si}.
$$

Hence S_t has a compound Poisson or compound binomial distribution, the compounding distribution being a discrete one, essentially the stationary distribution of the Markov chain. Bivariate distributions are available from

$$
\mathrm{P}(S_t=u, S_{t+1}=v) = \delta\,{}_t\lambda(u)\,\Gamma\,{}_{t+1}\lambda(v)\,\mathbf{1}' = \sum_{i,j=1}^{m} \delta_i\,{}_t\pi_{ui}\gamma_{ij}\,{}_{t+1}\pi_{vj}
$$

or, more generally, from the expression (2.21). In the case $m=2$ of a Poisson-HM model we therefore have, for instance,

$$\begin{aligned} \mathrm{P}(S_t=u, S_{t+1}=v) &= \delta_1(1-\gamma_1)\pi_{u1}\pi_{v1} + \delta_1\gamma_1\pi_{u1}\pi_{v2} + \\ &\qquad \delta_2\gamma_2\pi_{u2}\pi_{v1} + \delta_2(1-\gamma_2)\pi_{u2}\pi_{v2}, \end{aligned}$$

where $\Gamma = \begin{pmatrix} 1-\gamma_1 & \gamma_1 \\ \gamma_2 & 1-\gamma_2 \end{pmatrix}$, $\delta = \frac{1}{\gamma_1+\gamma_2}(\gamma_2 \ \gamma_1)$, and $\pi_{si} = e^{-\lambda_i}\lambda_i^s/s!$ for all nonnegative integers s and $i = 1, 2$. It is clear that trivariate and higher-order joint distributions may similarly be obtained.

Since joint distributions are available, conditional distributions follow easily. One conditional distribution of particular statistical interest is the one-step-ahead forecast distribution, i.e. the distribution of S_{T+1}, given $S_1, \ldots, S_T$. This is given by a ratio of likelihood values as follows:

$$\begin{aligned} \mathrm{P}(S_{T+1}=s_{T+1} \mid S_1=s_1, \ldots, S_T=s_T) &= L_{T+1}/L_T \\ &= \frac{\delta_1\lambda(s_1)\Gamma \cdots {}_{T+1}\lambda(s_{T+1})\,\mathbf{1}'}{\delta_1\lambda(s_1)\Gamma \cdots {}_{T}\lambda(s_T)\,\mathbf{1}'}. \end{aligned}$$

More generally, the k-step-ahead forecast distribution is given by

$$\begin{aligned} &\mathrm{P}(S_{T+k}=s_{T+k} \mid S_1=s_1, \ldots, S_T=s_T) \\ &= \mathrm{P}(S_1=s_1, \ldots, S_T=s_T, S_{T+k}=s_{T+k})/L_T \\ &= \delta\,\lambda(s_1)\Gamma \cdots {}_{T}\lambda(s_T)\Gamma^k {}_{T+k}\lambda(s_{T+k})\,\mathbf{1}'/L_T. \end{aligned}$$

A question of interest which may be answered by evaluating conditional probabilities is whether hidden Markov models are in general themselves Markov processes. While it may seem obvious that the answer is no, for the sake of completeness a simple nonpathological counterexample to the Markov

property is presented here.

Example Consider the binomial-HM model with $\Gamma = \frac{1}{4}\begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix}$, $n_t = 1$ for all t, and $p = \begin{pmatrix} \frac{1}{2} & 1 \end{pmatrix}$. It follows that $\delta = \frac{1}{3}\begin{pmatrix} 1 & 2 \end{pmatrix}$ and $\lambda(1) = \text{diag}(p)$. The probabilities we need are:

$$\begin{aligned} \mathrm{P}(S_2=1) &= \delta\lambda(1)\mathbf{1}' \\ &= 5/6; \\ \mathrm{P}(S_1=S_2=1) &= \mathrm{P}(S_2=S_3=1) \\ &= \delta\lambda(1)\Gamma\lambda(1)\mathbf{1}' \\ &= 17/24; \\ \mathrm{P}(S_1=S_2=S_3=1) &= \delta\lambda(1)\Gamma\lambda(1)\Gamma\lambda(1)\mathbf{1}' \\ &= 29/48. \end{aligned}$$

Hence $\mathrm{P}(S_3 = 1 \mid S_2 = 1) = 17/20$ and $\mathrm{P}(S_3 = 1 \mid S_2 = 1, S_1 = 1) = 29/34$, which contradicts the Markov property. □

In subsection 2.4.1, which discussed the ACF of Poisson-HM models, it was proved that, if all the elements of λ are equal, the random variables S_t and S_{t+k} are uncorrelated. This suggests that they may be independent. Using the matrix expression for the likelihood, we are in fact able to show, for Poisson- and binomial-HM models, that the random variables $S_1, \ldots, S_T$ are in general mutually independent if λ_i (or p_i) is constant. For notational convenience we demonstrate this only for Poisson- or stationary binomial-HM models, but the proof for the binomial case with nonconstant n_t is entirely similar.

Suppose, therefore, that π_{si} is, for each s, constant with respect to i.

Then for all s,

$$\lambda(s) = \mathrm{diag}(\pi_{s1} \ldots \pi_{sm}) = \pi_{s1} I_m.$$

Hence

$$\begin{aligned} \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T) &= \delta\lambda(s_1)\Gamma\lambda(s_2)\Gamma \cdots \Gamma\lambda(s_T)\mathbf{1}' \\ &= \pi_{s_1 1}\pi_{s_2 1} \cdots \pi_{s_T 1}\delta\Gamma^{T-1}\mathbf{1}' \\ &= \pi_{s_1 1}\pi_{s_2 1} \cdots \pi_{s_T 1}, \end{aligned}$$

since $\Gamma\mathbf{1}' = \mathbf{1}'$ and $\delta\mathbf{1}' = 1$. By the multiplicative form of their joint probability, $S_1, \ldots, S_T$ are mutually independent and

$$\mathrm{P}(S_t = s) = \pi_{s1} = \begin{cases} e^{-\lambda}\lambda^s/s! \\ \binom{n}{s} p^s (1-p)^{n-s}. \end{cases}$$

Not unexpectedly, therefore, $S_1, \ldots, S_T$ are in this case mutually independent Poisson or binomial random variables. Even if n_t is not constant, $S_1, \ldots, S_T$ are mutually independent, although not identically distributed. In that case

$$\mathrm{P}(S_t = s) = \binom{n_t}{s} p^s (1-p)^{n_t - s}.$$

Conditional probabilities of the form $\mathrm{P}(C_t = i \mid S_1 = s_1, \ldots, S_T = s_T)$ may now also be derived: these results are a slight generalization of the corresponding ones of Zucchini and Guttorp (1991). We exclude for the moment the binomial case with nonconstant n_t. Consider first the case $t > T$, that of 'state prediction'. We denote by $A_{\bullet i}$ the ith column of a matrix A, i.e. a column vector, and by $A_{i\bullet}$ the ith row. Since

$$\begin{aligned} & \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, C_t = i) \\ = & \sum_{i_1} \cdots \sum_{i_T} (\pi_{s_1 i_1} \cdots \pi_{s_T i_T}) \left(\delta_{i_1}\gamma_{i_1 i_2} \cdots \gamma_{i_{T-1} i_T}\gamma_{i_T i}(t-T)\right) \\ = & \ \delta\lambda(s_1)\Gamma\lambda(s_2) \cdots \Gamma\lambda(s_T)(\Gamma^{t-T})_{\bullet i}, \end{aligned}$$

we have, for $t > T$,

$$\mathrm{P}(C_t = i \mid S_1 = s_1, \ldots, S_T = s_T) = \delta\lambda(s_1)\Gamma\lambda(s_2)\cdots\Gamma\lambda(s_T)(\Gamma^{t-T})_{\bullet i}/L_T.$$

The case $t = T$, that of 'filtering', is proved similarly:

$$\begin{aligned} & \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, C_T = i) \\ & \sum_{i_1} \cdots \sum_{i_{T-1}} (\pi_{s_1 i_1} \cdots \pi_{s_{T-1} i_{T-1}} \pi_{s_T i})(\delta_{i_1} \gamma_{i_1 i_2} \cdots \gamma_{i_{T-2} i_{T-1}} \gamma_{i_{T-1} i}) \\ = \; & \delta\lambda(s_1)\Gamma\lambda(s_2)\cdots\Gamma\lambda(s_{T-1})\Gamma_{\bullet i}\pi_{s_T i}. \end{aligned}$$

Hence

$$\mathrm{P}(C_T = i \mid S_1 = s_1, \ldots, S_T = s_T) = \delta\lambda(s_1)\Gamma\lambda(s_2)\cdots\Gamma\lambda(s_{T-1})\Gamma_{\bullet i}\pi_{s_T i}/L_T.$$

The case $1 \lesseqgtr t < T$, that of 'smoothing', is conveniently split into $1 < t < T$ and $t = 1$. For $1 < t < T$,

$$\begin{aligned} & \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, C_t = i) \\ = \; & \sum_{i_1} \cdots \sum_{i_{t-1}} \sum_{i_{t+1}} \cdots \sum_{i_T} (\pi_{s_1 i_1} \cdots \pi_{s_{t-1} i_{t-1}} \pi_{s_t i} \pi_{s_{t+1} i_{t+1}} \cdots \pi_{s_T i_T}) \times \\ & \qquad\qquad (\delta_{i_1} \gamma_{i_1 i_2} \cdots \gamma_{i_{t-1} i} \gamma_{i i_{t+1}} \cdots \gamma_{i_{T-1} i_T}) \\ = \; & \delta\lambda(s_1)\Gamma\cdots\lambda(s_{t-1})\Gamma_{\bullet i}\pi_{s_t i}\Gamma_{i\bullet}\lambda(s_{t+1})\Gamma\cdots\Gamma\lambda(s_T)\mathbf{1}'. \end{aligned}$$

This last quantity, divided by L_T, therefore gives $\mathrm{P}(C_t = i \mid S_1 = s_1, \ldots, S_T = s_T)$ for the case $1 < t < T$. Finally, for $t = 1$, we have

$$\begin{aligned} & \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, C_1 = i) \\ = \; & \sum_{i_2} \cdots \sum_{i_T} (\pi_{s_1 i} \pi_{s_2 i_2} \cdots \pi_{s_T i_T})(\delta_i \gamma_{i i_2} \cdots \gamma_{i_{T-1} i_T}) \\ = \; & \delta_i \pi_{s_1 i} \Gamma_{i\bullet} \lambda(s_2)\Gamma\lambda(s_3)\cdots\Gamma\lambda(s_T)\mathbf{1}', \end{aligned}$$

and division by L_T gives $\mathrm{P}(C_1 = i \mid S_1 = s_1, \ldots, S_T = s_T)$.

In all four cases above, all that needs to be done to include binomial-hidden Markov models with nonconstant n_t is to insert a subscript, indicating the time, before each appearance of the symbols π and λ.

For $1 \leq t \leq T$, an alternative approach to the computation of the probabilities $\mathrm{P}(C_t = i \mid S_1 = s_1, \ldots, S_T = s_T)$ is provided by the forward and backward probabilities, $\alpha_t(i)$ and $\beta_t(i)$, of section 2.2. Equation (2.5) of that section tells us that, for $1 \leq t \leq T$,

$$\alpha_t(i)\beta_t(i) = \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, C_t = i),$$

and equation (2.6) that $\sum_{i=1}^{m} \alpha_t(i)\beta_t(i) = L_T$, for $1 \leq t \leq T$. If, therefore, we successively compute $\alpha_1(i), \alpha_2(i), \ldots, \alpha_t(i)$ and $\beta_T(i), \beta_{T-1}(i), \ldots, \beta_t(i)$ for each i, we can find the conditional distribution of C_t, given $S_1, \ldots, S_T$ as $\alpha_t(i)\beta_t(i)/L_T$. While it is true that the formulation of section 2.2 does not allow for the variation of the probabilities π_{si} with time, that is a modification easily introduced.

Finally we record here the joint distribution of $C_1, C_2, \ldots, C_T$ given the observations $s_1, s_2, \ldots, s_T$:

$$\begin{aligned} &\mathrm{P}(C_1 = i_1, \ldots, C_T = i_T \mid S_1 = s_1, \ldots, S_T = s_T) \\ = \; &({}_1\pi_{s_1 i_1} \cdots {}_T\pi_{s_T i_T})(\delta_{i_1}\gamma_{i_1 i_2} \cdots \gamma_{i_{T-1} i_T})/L_T. \end{aligned}$$

In principle the corresponding conditional distribution for C_t only could be found from this expression by summation, but that is practicable only for very small T.

2.7 Parameter estimation

Since the logarithm of the likelihood function can be evaluated routinely, even for very long sequences of observations, it is (as already remarked) fea-

sible to perform parameter estimation in hidden Markov models by direct numerical maximization of the log-likelihood.

We consider first the case of a Poisson-HM model in which the Markov chain has only two states, i.e. $m=2$. The four parameters are, in the usual notation, $\gamma_1, \gamma_2, \lambda_1$ and λ_2. The only constraints on these parameters are $0 < \gamma_i \leq 1$ and $\lambda_i \geq 0$. For practical purposes these may be treated as $0 < \gamma_i < 1$ and $\lambda_i > 0$. The log-likelihood may then be reparametrized as the appropriate function of logit $\gamma_i = \log(\frac{\gamma_i}{1-\gamma_i})$ and $\log \lambda_i$, for $i = 1, 2$. An algorithm for unconstrained numerical maximization can be applied to obtain maximum likelihood estimates of logit γ_i and $\log \lambda_i$ $(i = 1, 2)$ — or equivalently estimates of γ_1, γ_2, λ_1 and λ_2. A derivative-free algorithm such as the simplex algorithm of Nelder and Mead (Press et al., 1986, p. 289) is convenient for this purpose, although numerical differentiation of the likelihood would make possible the use of an algorithm requiring derivatives.

The case of a binomial-HM model with $m=2$ can be handled similarly. The four parameters are γ_1, γ_2, p_1 and p_2. The constraints on these parameters are $0 < \gamma_i \leq 1$ and $0 \leq p_i \leq 1$, but for practical purposes we can treat all four as lying strictly between 0 and 1, and apply the logistic transform in order to use an unconstrained maximization algorithm.

For $m > 2$ the generalized upper bound constraints already mentioned in section 2.3 must also be taken into account. That is, in maximizing the log-likelihood with respect to the $m^2 - m$ independent transition probabilities γ_{ij}, $i \neq j$, and the parameters λ_i or p_i, we must satisfy the m additional constraints $\sum_{j \neq i} \gamma_{ij} \leq 1$. This considerably alters the nature of the optimization problem: it is in this case necessary to maximize a (nonlinear) objective function subject to linear constraints other than the usual simple lower and

upper bounds of 0 and 1 — preferably without supplying any derivatives of the objective. The program NPSOL (Gill et al., 1986) and the NAG version thereof, E04UCF (Numerical Algorithms Group, 1990), are designed to handle such problems (inter alia) and do not demand that values of the derivatives be supplied. The method used is a sequential quadratic programming algorithm. Although one can try to ensure that a global optimum is reached, by trying many sets of starting values of the parameters, there is no guarantee that this will succeed.

In general it should be noted that the distribution of the observations is invariant under permutation of the states of the Markov chain, and this implies nonuniqueness of the maximum likelihood estimators. This is not in practice a problem, and one can if necessary order the states, e.g. in increasing order of λ_i or p_i. This can be done by adding the relevant constraints, e.g. $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$, to the optimization problem.

At this stage it is useful to compare the estimation problem described above, and its solution, with the estimation problem solved by the Baum-Welch algorithm. The two problems are by no means identical. Firstly, the speech-processing models of section 2.2 do not assume that the Markov chain is stationary, but here we do assume just that. Our probabilities δ_i are therefore not initial probabilities requiring estimation, but are the stationary probabilities completely determined by the transition probabilities γ_{ij}. Secondly, the speech-processing models involve an $n \times m$ matrix Π of probabilities, $(n-1)m$ of which are independently determined. We assume instead that the distribution of the observation for a given state (i) is dependent on one parameter only (λ_i or p_i), not $n-1$, and we allow the state-space of the observations to be infinite. As it stands, therefore, the Baum-Welch algorithm is not applicable to the problem under discussion.

Another possible approach is contained in recent work of Leroux and Puterman (1992), which deals inter alia with hidden Markov models with Poisson conditional distribution. Their models are intermediate between those discussed here and the speech-processing models. In their models the distribution of an observation given the current state of the Markov chain depends, as it does here, on only one parameter (λ_j in their notation), but the initial probabilities are not asssumed to be the stationary probabilities. Leroux and Puterman maximize the likelihood with respect to the $m^2 - m$ independent transition probabilities, the m parameters λ_j, and the initial probabilities (but still take the number of parameters estimated to be m^2, e.g. in model selection). As they point out, that maximization can be accomplished by solving the m separate lower-dimensional maximization problems defined by starting from a fixed initial state: that is, one can take the initial probability of each of the states of the Markov chain in turn to be 1, and then choose as the initial state that one which produces the largest maximized log-likelihood. Although this property has been noted in the speech-processing literature (Levinson et al., p. 1055), it does not seem to have been utilized explicitly except by Leroux and Puterman.

Leroux and Puterman use the EM algorithm to find the maximizing values of the transition probabilities and of the parameters λ_j. The 'missing data', i.e. the sequence of states $i_1, \ldots, i_T$ followed by the Markov chain, are represented by the indicator random variables defined as follows: $v_{jk}(t) = 1$ if $i_{t-1} = j$ and $i_t = k$, and $u_j(t) = 1$ if $i_t = j$. The complete-data log-likelihood is given by

$$\sum_{t=2}^{T} \log \gamma_{i_{t-1} i_t} + \sum_{t=1}^{T} \log \pi_{s_t i_t} = \sum_{j=1}^{m} \sum_{k=1}^{m} \log \gamma_{jk} \sum_{t=2}^{T} v_{jk}(t) + \sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log \pi_{s_t j}.$$

This expression can be seen to consist of two parts: firstly the log-likelihood

of the Markov chain, conditioned on the initial state (i_1), and secondly the log-likelihood of T independent observations. The first of these depends only on the parameters γ_{jk}, i.e. the transition probabilities, and the jth term of the second part (as on the right-hand side above) depends only on the single parameter λ_j.

The E-step replaces $v_{jk}(t)$ and $u_j(t)$ in the complete-data log-likelihood by their conditional expectations given the observations (and current parameter estimates):

$$\hat{v}_{jk}(t) = \mathrm{P}(C_{t-1}=j, C_t=k \mid S_1, \ldots, S_T)$$

and

$$\hat{u}_j(t) = \mathrm{P}(C_t=j \mid S_1, \ldots, S_T).$$

The forward and backward probabilities, as defined in section 2.2, are needed to compute these conditional probabilities. Leroux and Puterman use the standard recursions (with scaling) to find the forward and backward probabilities, and derive the above conditional probabilities from them in the same way as described in section 2.2. The scaling method they use is to divide the forward probability $\alpha_t(i)$ by 10^p, where p is such that $10^{-p}\sum_i \alpha_t(i)$ lies between 0.1 and 1. The backward probabilities are scaled similarly.

The M-step separately maximizes the two parts of the complete-data log-likelihood. The Markov chain part is straightforward, since it is the standard problem of conditional maximum likelihood in a Markov chain, apart from the replacement of the missing data by their conditional expectations. The solution is otherwise as in section 1.2. The other part of the complete-data log-likelihood is maximized by setting λ_j equal to that value which maximizes

$$\sum_{t=1}^{T} \hat{u}_j(t) \log \pi_{s_t j}. \tag{2.23}$$

In the Poisson case this implies that

$$\lambda_j = \sum_{t=1}^{T} \hat{u}_j(t) s_t / \sum_{t=1}^{T} \hat{u}_j(t).$$

Leroux and Puterman do not discuss models with binomial conditional distribution, i.e. the case

$${}_t\pi_{sj} = \binom{n_t}{s} p_j^s (1-p_j)^{n_t - s},$$

but in that case the value of p_j which maximizes the expression (2.23) can be shown to be given by

$$\frac{p_j}{1-p_j} = \frac{\sum_{t=1}^{T} \hat{u}_j(t) s_t}{\sum_{t=1}^{T} \hat{u}_j(t)(n_t - s_t)}.$$

Although neither the approach of Leroux and Puterman nor the Baum-Welch algorithm is directly applicable here, simply because the respective problems do not coincide, the EM algorithm can also be used for our models. The complete-data log-likelihood is in this case

$$\begin{aligned} &\log \delta_{i_1} + \sum_{t=2}^{T} \log \gamma_{i_{t-1} i_t} + \sum_{t=1}^{T} \log \pi_{s_t i_t} \\ &= \sum_{j=1}^{m} u_j(1) \log \delta_j + \sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{t=2}^{T} v_{jk}(t) \right) \log \gamma_{jk} + \sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log \pi_{s_t j}, \end{aligned}$$

with δ the stationary distribution implied by Γ. The simplest way of describing the method is to say that the estimation procedure of Leroux and Puterman is used except that the estimation of the transition probabilities γ_{jk} is based on the unconditional likelihood of a stationary Markov chain rather than on the likelihood of a (not necessarily stationary) Markov chain conditioned on the initial state. This does have the consequence that the neat explicit expression Leroux and Puterman can use to estimate the transition probabilities (their equation (6)) is replaced by an optimization problem of the following form, in the $m^2 - m$ off-diagonal transition probabilities γ_{jk}, $j \neq k$: subject to $\sum_{k \neq j} \gamma_{jk} \leq 1$ $(j = 1, 2, \ldots, m)$,

and with δ denoting the stationary distribution implied by Γ, maximize $\sum_{j=1}^{m} a_j \log \delta_j + \sum_{j=1}^{m} \sum_{k=1}^{m} b_{jk} \log \gamma_{jk}$. The numerical solution of such a problem may for instance be performed by NPSOL or E04UCF, with starting values of γ_{jk} supplied by the previous iteration of EM. It seems likely that such an EM algorithm, requiring the solution of a constrained nonlinear optimization problem at each M-step, would be slow compared to the direct numerical maximization technique described earlier in this section. It may be possible, however, to use a method for accelerating the EM algorithm, e.g. that of Meilijson (1989).

It is interesting, although not directly relevant, to note that Campillo and Le Gland (1989) have compared the use of the EM algorithm with direct numerical maximization, in the case of a partially observed diffusion process. One of their conclusions, as might perhaps be expected, is that the EM algorithm can be very slow in some circumstances.

The properties of the maximum likelihood estimators used in this thesis require investigation. In the absence of exact or asymptotic distributional results, however, the parametric bootstrap (Efron, 1982, p. 29) may be used to provide an estimate of the covariance matrix of the estimators. Albert (1991) uses this method to compute standard errors for his estimators, and we shall illustrate the technique in the examples presented in sections 4.2 and 4.7. The parametric bootstrap may also be used to provide standard errors of forecasts: this point will be illustrated in section 4.2.

2.8 Reversibility

A random process is said to be reversible if its finite-dimensional distributions are invariant under reversal of time. More specifically, $\{X(t)\}$ is reversible if

the random vector

$$(X(t_1), X(t_2), \ldots, X(t_n))$$

has the same distribution as

$$(X(\tau - t_1), X(\tau - t_2), \ldots, X(\tau - t_n))$$

for all positive integers n and all (appropriate) $\tau, t_1, \ldots, t_n$. In the case of a stationary irreducible Markov chain with transition probability matrix Γ and stationary distribution δ, it is necessary and sufficient for reversibility that the 'detailed balance conditions'

$$\delta_i \gamma_{ij} = \delta_j \gamma_{ji}$$

be satisfied for all states i and j (Kelly, 1979, p. 5). Equivalently, if the states are ordered in some way, it is necessary and sufficient that the detailed balance conditions be satisfied for all states i and j such that $i < j$. These conditions are trivially satisfied by all two-state stationary irreducible Markov chains, which are thereby reversible. The Markov chain of the example in subsection 2.4.1 is not reversible, however, because $\delta_1 \gamma_{12} = (15/32)(1/3) = 5/32$ and $\delta_2 \gamma_{21} = (9/32)(2/3) = 6/32$.

For some applications one may wish to use reversible time series models, and for others irreversible. The classic example of a time series displaying irreversibility is (deseasonalized) streamflow. Since Gaussian processes are characterized by their first and second moments, the stationary normal-theory time series models are all reversible. We show that, for $\{S_t\}$ a Poisson-HM or stationary binomial-HM model, reversibility of $\{C_t\}$ implies that of $\{S_t\}$, but not conversely: an irreversible $\{C_t\}$ may be associated with an $\{S_t\}$ either reversible or irreversible.

Let $\{C_t\}$, then, be reversible and let $\{S_t\}$ be as specified above. It will suffice to show that, for all T and $s_1, s_2, \ldots, s_T$:

$$\mathrm{P}(S_1 = s_1, S_2 = s_2, \ldots, S_T = s_T) = \mathrm{P}(S_T = s_1, S_{T-1} = s_2, \ldots, S_1 = s_T).$$

One way to prove this is to use the matrix expression for the likelihood, i.e. to show that

$$\delta\lambda(s_1)\Gamma\lambda(s_2)\cdots\Gamma\lambda(s_T)\mathbf{1}' = \delta\lambda(s_T)\Gamma\lambda(s_{T-1})\cdots\Gamma\lambda(s_1)\mathbf{1}'. \tag{2.24}$$

This is accomplished by writing Γ as AD, where as before D is defined as $\mathrm{diag}(\delta)$, and the matrix A by $a_{ij} = \gamma_{ij}/\delta_j$. Note that, by the detailed balance conditions, A is symmetric. Note also that D and $\lambda(s)$, being diagonal matrices, commute under multiplication and are symmetric. The left-hand side of equation (2.24), being a scalar, equals its transpose, viz.

$$\begin{aligned} & \mathbf{1}\lambda(s_T)DA\lambda(s_{T-1})DA\cdots\lambda(s_2)DA\lambda(s_1)(\mathbf{1}D)' \\ = \; & \mathbf{1}D\lambda(s_T)AD\lambda(s_{T-1})\cdots\lambda(s_2)AD\lambda(s_1)\mathbf{1}' \\ = \; & \delta\lambda(s_T)\Gamma\lambda(s_{T-1})\cdots\lambda(s_2)\Gamma\lambda(s_1)\mathbf{1}'. \end{aligned}$$

This completes the proof. It is however interesting to note that another, perhaps more obvious, method of proof establishes this result without recourse to the Markov property of $\{C_t\}$: one simply conditions on $C^{(T)}$ and exploits the reversibility of $\{C_t\}$. The details are as follows:

$$\begin{aligned} & \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T) \\ = \; & \sum_{i_1} \cdots \sum_{i_T} (\pi_{s_1 i_1}\pi_{s_2 i_2}\cdots\pi_{s_T i_T})\mathrm{P}(C_1 = i_1, \ldots, C_T = i_T) \\ = \; & \sum_{i_1,\ldots,i_T} (\pi_{s_1 i_1}\cdots\pi_{s_T i_T})\mathrm{P}(C_T = i_1, \ldots, C_1 = i_T) \\ = \; & \sum_{i_1,\ldots,i_T} \mathrm{P}(S_T = s_1 \mid C_T = i_1)\cdots\mathrm{P}(S_1 = s_T \mid C_1 = i_T)\mathrm{P}(C_T = i_1, \ldots, C_1 = i_T) \\ = \; & \mathrm{P}(S_T = s_1, \ldots, S_1 = s_T). \end{aligned}$$

(The second-last equality holds because $\mathrm{P}(S_t = s \mid C_t = i) = \pi_{si}$ for all t.)

To see that $\{C_t\}$ irreversible does not imply $\{S_t\}$ irreversible, let $\{C_t\}$ be irreversible and let $\{S_t\}$ be a Poisson-HM or stationary binomial-HM model based on it, with all the parameters λ_i or p_i equal. As demonstrated in section 2.6, $\{S_t\}$ is then just a sequence of independent and identically distributed random variables, and thereby reversible. At the end of this section we provide an example of a stationary irreversible hidden Markov model $\{S_t\}$, which is necessarily associated with an irreversible $\{C_t\}$.

One possible advantage of using a reversible rather than a general Markov chain in a hidden Markov model is parsimony. To specify a general chain on m states takes $m^2 - m$ parameters: to specify a reversible one takes $m - 1 + \binom{m}{2} = \frac{1}{2}(m-1)(m+2)$. This is because one needs to specify $m-1$ of the elements of δ, and γ_{ij} for $i < j$. The remaining elements of Γ are then available from the detailed balance conditions and the row sum constraints on Γ. Hence there is a saving of

$$m^2 - m - (1/2)(m-1)(m+2) = (1/2)(m-1)(m-2)$$

parameters in choosing $\{C_t\}$ to be reversible. As expected, there is no saving in the two-state case. One disadvantage of this approach, however, is that, if one seeks to maximize the likelihood with respect to the $(1/2)(m-1)(m+2)$ parameters, there are now nonlinear constraints of the form

$$\sum_{j=i+1}^{m} (\gamma_{ij} + \gamma_{ij}\delta_i/\delta_j) \leq 1$$

to be satisfied.

Given any stationary process $\{S_t\}$, a means of detecting irreversibility in some cases is to compare directional moments like $\mathrm{E}(S_t S_{t+k}^2)$ and $\mathrm{E}(S_t^2 S_{t+k})$.

These will be equal if $\{S_t\}$ is reversible, otherwise possibly not. For that reason we include here a derivation of these quantities for both Poisson-HM and binomial-HM models. Although for the latter only the case of n_t constant is relevant to reversibility, the expressions for the directional moments are as easily derived for general n_t, and we do so. An economical way of proving all these results is to use equation (2.14) of section 2.4:

$$\mathrm{E}(f(S_t, S_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}\left(f(S_t, S_{t+k}) \mid C_t = i, C_{t+k} = j\right) \delta_i \gamma_{ij}(k).$$

The results for the Poisson case, valid for $k \in \mathbf{N}$, are then

$$\mathrm{E}(S_t S_{t+k}^2) = \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) \lambda_i (\lambda_j + \lambda_j^2)$$

and

$$\mathrm{E}(S_t^2 S_{t+k}) = \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) (\lambda_i + \lambda_i^2) \lambda_j.$$

For the binomial case they are

$$\mathrm{E}(S_t S_{t+k}^2) = \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) (n_t p_i) \left(n_{t+k} p_j (1 - p_j) + n_{t+k}^2 p_j^2\right)$$

and

$$\mathrm{E}(S_t^2 S_{t+k}) = \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) \left(n_t p_i (1 - p_i) + n_t^2 p_i^2\right) (n_{t+k} p_j).$$

With these expressions available for directional moments, we can now provide the example of a stationary irreversible hidden Markov model which is referred to above.

Example Let $\{C_t\}$ again be the (irreversible) Markov chain with transition probability matrix

$$\Gamma = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

and stationary distribution $\delta = \frac{1}{32}(15\ 9\ 8)$. Let $\{S_t\}$ be the binomial-HM model with $n_t = 2$ for all t and $p = (0\ \frac{1}{2}\ 1)$. Then irreversibility is shown by comparing $\mathrm{E}(S_t S_{t+1}^2)$ with $\mathrm{E}(S_t^2 S_{t+1})$:

$$\begin{aligned}
\mathrm{E}(S_t S_{t+1}^2) &= \sum_{i,j=1}^{3} \delta_i \gamma_{ij}\ (2p_i)\left(2p_j(1-p_j) + 4p_j^2\right) \\
&= 4 \sum_{i,j=2}^{3} \delta_i \gamma_{ij} p_i p_j (1 + p_j) \\
&= 3/4; \\
\mathrm{E}(S_t^2 S_{t+1}) &= \sum_{i,j=1}^{3} \delta_i \gamma_{ij}\ \left(2p_i(1-p_i) + 4p_i^2\right)(2p_j) \\
&= 4 \sum_{i,j=2}^{3} \delta_i \gamma_{ij} p_i (1 + p_i) p_j \\
&= 25/32.
\end{aligned}$$

□

2.9 Discussion

This chapter has introduced a class of hidden Markov models for time series of bounded or unbounded counts, that is for both 'binomial-like' and 'Poisson-like' observations. Clearly these models can accommodate a wider range of correlation structures than can Markov chains, or for that matter many of the other competing models described in Chapter 1. Negative correlation seems to be as easily modelled as positive, and the number of parameters is not excessive if m, the number of states of the Markov chain, is small. Furthermore, the relative ease with which parameters may be estimated by maximum likelihood contrasts quite sharply with the apparent difficulties of estimation in e.g. some of the discrete-valued ARMA processes of Chapter 1. Model selection by minimizing Akaike's information criterion or the Bayesian information criterion (Schwarz, 1978) is straightforward to

perform in the case of hidden Markov models, and the use of these criteria will be illustrated in Chapter 4.

Such models are similar, but not identical, to models used in speech processing and models recently studied by Leroux and Puterman, but the EM algorithm is more easily applied to these other classes of models than to the ones introduced here. Direct numerical maximization works well here, however, and EM is available as an alternative. An advantage of using models in which the Markov chain is assumed stationary is that the autocorrelation function is available as a means of model identification (provided, of course, that n_t is constant if the model is one of binomial type). As will be seen in Chapter 3, however, the basic models of this chapter can be extended in various ways to cater for nonstationarity in the observations, without any transformation of the data being necessary, and without much modification of the estimation technique.

In some applications of hidden Markov models the states of the Markov chain may have, or may turn out to have, a useful substantive interpretation. That is, they may have a definite interpretation in terms of the subject-matter of the application concerned. For instance, the 'climate state' of the multi-site precipitation models of Zucchini and Guttorp (1991) may correspond to a meteorologically defined climate state affecting all the sites. Even if the models are not substantive ones, however, they may be very useful as empirical ones. To illustrate this point we may consider the Gaussian ARMA models for continuous-valued time series: they are most often used as empirical models, i.e. without any close link to subject-matter considerations, yet they are no less useful for that. (We use the terms 'empirical model' and 'substantive model' in the sense used by Cox (1990).)

An important aspect of hidden Markov models which will now be discussed in detail is the ease with which the basic models of this chapter can be modified to accommodate a wide variety of different kinds of observation.

Chapter 3

Hidden Markov time series models: extensions and modifications

3.1 Introduction

One of the striking features of the hidden Markov models introduced in the previous chapter is the variety of ways in which they can be modified or generalized in order to provide a broader class of potentially useful models. One possible modification, the use of a reversible Markov chain for $\{C_t\}$, has already been described in some detail in section 2.8. That modification resulted in some reduction in the number of parameters, provided that m, the number of states of $\{C_t\}$, exceeded two.

More generally, any a priori restriction on the nature of the Markov chain, or equivalently on its transition probability matrix Γ, may result in a useful saving of parameters. For instance, if there were some reason to suppose that two states i and j satisfied $\gamma_{ij} = \gamma_{ji}$, that would save one parameter. If

it were supposed that $\gamma_{ij} = \gamma_{ji}$ for all i and j, that would save $\frac{1}{2}m(m-1)$ parameters. To take an extreme example, consider a Markov chain with all off-diagonal elements of Γ equal, which requires only one parameter to define it: a hidden Markov model based on such a Markov chain has only $m+1$ parameters. In all of these cases the use of the model would need to be justified by the a priori reasonableness of the restrictions placed on the structure of Γ. The evaluation of the likelihood could, however, proceed exactly as before, and the maximization thereof would present no new features apart from appropriate modifications of the constraints on the parameters.

The rest of this chapter will therefore discuss other extensions and modifications. Most of these extensions fall into one of two broad classes: modifications of the 'parameter process' $\{C_t\}$, and modifications of the state-dependent probabilities ${}_t\pi_{si}$ which, together with $\{C_t\}$, make up the model. As the purpose is the exploration of statistically useful models and their properties, we shall not always seek the greatest generality possible if that would not contribute to this purpose.

3.2 Models based on a second-order Markov chain

One potentially useful extension of hidden Markov models involves the replacement of the underlying (first-order) Markov chain by a stationary second-order Markov chain $\{C_t\}$ on state space $M = \{1, 2, \ldots, m\}$. We suppose that $\{C_t\}$ has transition probabilities

$$p(i,j,k) = \mathrm{P}(C_t = k \mid C_{t-1} = j, C_{t-2} = i)$$

and stationary probabilities $u(j,k) = \mathrm{P}(C_{t-1}=j, C_t=k)$, satisfying

$$u(j,k) = \sum_{i=1}^{m} u(i,j)p(i,j,k)$$

and

$$\sum_{j=1}^{m}\sum_{k=1}^{m} u(j,k) = 1.$$

The process $\{C_t\}$ is not a Markov chain, but if we make the definition $X_t = (C_{t-1}, C_t)$, $\{X_t\}$ is indeed a Markov chain, on state space M^2. Given the three-dimensional array of probabilities $p(i,j,k)$, the matrix U of stationary probabilities $u(j,k)$ can be determined by finding the stationary distribution vector of $\{X_t\}$, i.e. the normalized left eigenvector corresponding to eigenvalue 1, of the transition probability matrix of $\{X_t\}$.

A model $\{S_t\}$ based on a general second-order Markov chain $\{C_t\}$ on m states may be grossly overparametrized for statistical purposes. Using instead the Pegram or Raftery submodel (see section 1.3) can result in a considerable reduction in the number of parameters, as will be evident from Table 3.1. In all cases of that table the 'parameter process' $\{C_t\}$ has m states.

For any model involving some second-order Markov chain as parameter process, the case $m = 2$ is the most interesting, as it may provide a practical alternative to the use of an ordinary hidden Markov model with $m > 2$ states. That is, if a model based on a two-state first-order Markov chain is found to be inadequate, one can consider the use of a model based on a two-state second-order Markov chain instead of one based on (e.g.) a three-state first-order Markov chain. This last model would require a total of nine parameters, while the one based on the second-order chain would require either six or five, depending on whether a general second-order Markov chain or the Pegram-Raftery submodel is used as $\{C_t\}$. (Recall that for $m = 2$ the

Table 3.1: A comparison of the numbers of parameters needed to specify various models of hidden Markov type.

$\{C_t\}$	no. of parameters to specify $\{C_t\}$	total no. to specify $\{S_t\}$
Markov chain	$m(m-1)$	m^2
general second-order Markov chain	$m^2(m-1)$	m^3-m^2+m
Pegram model for second-order MC	$m+1$	$2m+1$
Raftery model for second-order MC	$m(m-1)+1$	m^2+1

Pegram and Raftery models are equivalent.) We therefore consider the case $m=2$ in some detail here.

If a general two-state second-order Markov chain is used as $\{C_t\}$, the associated Markov chain $\{X_t\}$ has transition probability matrix of the form:

$$\begin{pmatrix} 1-a & a & 0 & 0 \\ 0 & 0 & c & 1-c \\ 1-d & d & 0 & 0 \\ 0 & 0 & b & 1-b \end{pmatrix}.$$

(The states are, in order: (1,1), (1,2), (2,1), (2,2).) The four parameters of $\{C_t\}$, viz. a, b, c and d, are bounded by 0 and 1, but otherwise unconstrained.

If the Pegram-Raftery submodel is used for $\{C_t\}$, we require that

$$c = l_1 b + l_2(1-a) \tag{3.1}$$

and

$$d = l_1 a + l_2(1-b), \tag{3.2}$$

where $l_1 + l_2 = 1$ but l_1 and l_2 are not in general assumed to be nonnegative. The three parameters of $\{C_t\}$ can then be taken to be a, b and l_1, subject to c and d (as defined by equations (3.1) and (3.2)) being bounded by 0 and 1.

The stationary distribution of $\{X_t\}$ is proportional to

$$\begin{pmatrix} b(1-d) & ab & ab & a(1-c) \end{pmatrix},$$

hence the matrix of stationary probabilities for $\{C_t\}$ is

$$U = \frac{1}{b(1-d) + 2ab + a(1-c)} \begin{pmatrix} b(1-d) & ab \\ ab & a(1-c) \end{pmatrix}.$$

Consider now the general problem of evaluating the likelihood of T consecutive observations from a model $\{S_t\}$ of hidden Markov type, but with the parameter process $\{C_t\}$ a second-order Markov chain on state space M, with transition probabilities $p(i,j,k)$ and stationary probabilities $u(j,k)$. For $t \geq 2$, define

$$\nu_t(i,j) = \mathrm{P}(S_1 = s_1, \ldots, S_t = s_t, C_{t-1} = i, C_t = j).$$

For instance, $\nu_2(i,j) = {}_1\pi_{s_1 i}\, {}_2\pi_{s_2 j}\, u(i,j)$. It follows from this definition that

$$\begin{aligned} & \nu_t(i_{t-1}, i_t) \\ &= \sum_{i_1=1}^{m} \cdots \sum_{i_{t-2}=1}^{m} \mathrm{P}(S_1 = s_1, \ldots, S_t = s_t, C_1 = i_1, \ldots, C_t = i_t) \\ &= \sum_{i_1, \ldots, i_{t-2}} {}_1\pi_{s_1 i_1} \cdots {}_t\pi_{s_t i_t} u(i_1, i_2) p(i_1, i_2, i_3) p(i_2, i_3, i_4) \cdots p(i_{t-2}, i_{t-1}, i_t). \end{aligned}$$

This includes the case $t=2$, provided we interpret an empty product as 1, and note that for $t=2$ there is no summation.

Hence for $t \geq 3$

$$\begin{aligned}
& \nu_t(i_{t-1}, i_t) \\
= \; & {}_t\pi_{s_t i_t} \sum_{i_{t-2}} p(i_{t-2}, i_{t-1}, i_t) \times \\
& \left(\sum_{i_1, \ldots, i_{t-3}} {}_1\pi_{s_1 i_1} \cdots {}_{t-1}\pi_{s_{t-1} i_{t-1}} u(i_1, i_2) p(i_1, i_2, i_3) \cdots p(i_{t-3}, i_{t-2}, i_{t-1}) \right) \\
= \; & {}_t\pi_{s_t i_t} \sum_{i_{t-2}} p(i_{t-2}, i_{t-1}, i_t)\, \nu_{t-1}(i_{t-2}, i_{t-1}).
\end{aligned}$$

Since we know $\nu_2(i,j)$ for all i and j, this recursion enables us to arrive at $\nu_T(i,j)$ for all i and j and so at

$$\sum_i \sum_j \nu_T(i,j) = \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T), \tag{3.3}$$

which is the required likelihood. The computational effort involved in this algorithm is linear in T and cubic in m. As in the case of models based on a first-order Markov chain, scaling is necessary in practice to avoid underflow: at each stage an m by m matrix of probabilities $\nu_t(i,j)$ is held, and it is this matrix that must be scaled. Once the scaled likelihood has been computed by equation (3.3), its logarithm is then adjusted by the sum of the logs of the scale factors (as before). It will be noted that the quantities $\nu_t(i,j)$ are just an extension of the forward probabilities of section 2.2.

With the likelihood at our disposal, we are able to find the one-step-ahead forecast distribution exactly as before:

$$\mathrm{P}(S_{T+1} = s_{T+1} \mid S_1 = s_1, \ldots, S_T = s_T) = \frac{\mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, S_{T+1} = s_{T+1})}{\mathrm{P}(S_1 = s_1, \ldots, S_T = s_T)}.$$

The general k-step-ahead forecast distribution is slightly more awkward. If m^{k-1} is sufficiently small, we can evaluate

$$\begin{aligned} &\sum_{s_{T+1}} \cdots \sum_{s_{T+k-1}} \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, S_{T+1} = s_{T+1}, \ldots, S_{T+k} = s_{T+k}) \\ = \; &\mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, S_{T+k} = s_{T+k}). \end{aligned} \tag{3.4}$$

That is, we merely sum over the $k-1$ indices $s_{T+1}, \ldots, s_{T+k-1}$. Division by the likelihood of the first T observations then yields the required conditional probability, $\mathrm{P}(S_{T+k} = s_{T+k} \mid S_1 = s_1, \ldots, S_T = s_T)$.

It is interesting to note, however, that the recursive algorithm for finding ν_T and hence the likelihood function can be modified to provide an algorithm linear in k for computing the joint probability $\mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, S_{T+k} = s)$. We proceed as follows: for integers $t \geq T+1$ and all states i and j, define

$$\phi_t(i,j) = \mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, S_t = s, C_{t-1} = i, C_t = j).$$

It then follows that, for $t \geq T+2$,

$$\begin{aligned} \phi_t(i_{t-1}, i_t) &= \sum_{i_1} \cdots \sum_{i_{t-2}} {}_1\pi_{s_1 i_1} \cdots {}_T\pi_{s_T i_T} \, {}_t\pi_{s i_t} \\ &\qquad \times u(i_1, i_2) p(i_1, i_2, i_3) \cdots p(i_{t-2}, i_{t-1}, i_t) \\ &= \frac{{}_t\pi_{s i_t}}{{}_{t-1}\pi_{s i_{t-1}}} \sum_{i_{t-2}} p(i_{t-2}, i_{t-1}, i_t) \sum_{i_1, \ldots, i_{t-3}} {}_1\pi_{s_1 i_1} \cdots {}_T\pi_{s_T i_T} \, {}_{t-1}\pi_{s i_{t-1}} \\ &\qquad \times u(i_1, i_2) p(i_1, i_2, i_3) \cdots p(i_{t-3}, i_{t-2}, i_{t-1}) \\ &= \frac{{}_t\pi_{s i_t}}{{}_{t-1}\pi_{s i_{t-1}}} \sum_{i_{t-2}} p(i_{t-2}, i_{t-1}, i_t) \phi_{t-1}(i_{t-2}, i_{t-1}). \end{aligned}$$

The recursion is started by noting that, for all i and j,

$$\phi_{T+1}(i,j) = \nu_{T+1}(i,j)$$

where ν_{T+1} is calculated on the basis of the observation sequence $s_1, s_2, \ldots, s_T, s$. The joint probability we seek is then given by

$$\mathrm{P}(S_1 = s_1, \ldots, S_T = s_T, S_{T+k} = s) = \sum_i \sum_j \phi_{T+k}(i,j). \tag{3.5}$$

Bivariate and marginal distributions for $\{S_t\}$ can also be written down:

$$\mathrm{P}(S_t=u, S_{t+1}=v) = \sum_i \sum_j u(i,j)\, {}_t\pi_{ui\ t+1}\pi_{vj},$$

and

$$\mathrm{P}(S_t=u) = \sum_i \delta_i\, {}_t\pi_{ui},$$

where $\delta_i = \mathrm{P}(C_t=i) = \sum_j u(i,j) = \sum_j u(j,i)$. One can use either equation (3.4) or equation (3.5) to find $\mathrm{P}(S_1=u, S_{1+k}=v)$ and thereby the kth-order autocovariance and autocorrelation of a stationary $\{S_t\}$ based on a second-order Markov chain. The nonstationary case, i.e. models involving a binomial distribution with nonconstant n_t, can be handled similarly, although equations (3.4) and (3.5) do not apply exactly as they stand, and would need to be generalized slightly to be relevant.

3.3 Multinomial-hidden Markov models

Here we consider the multinomial extension of binomial-hidden Markov models. Essentially all this involves is the presence of q mutually exclusive categories (rather than merely two), into one of which each trial falls. As in the binomial case, there are n_t trials at time t. One set of data for which such a model is useful is the data discussed in section 4.8, relating to homicides and suicides in Cape Town during the period from 1986 to 1991. There n_t represents the total number of deaths due to homicide or suicide in week t, and the $q=5$ categories are: firearm homicide, non-firearm homicide, firearm suicide, non-firearm suicide and 'legal intervention homicide'.

More formally, let $\{C_t\}$ be the usual stationary first-order Markov chain on m states, and suppose that, conditional on $C^{(T)}$, the T random vectors

$$\underline{S}_t = (S_{t1}, S_{t2}, \ldots, S_{tq}) \quad (t = 1, 2, \ldots, T)$$

have independent multinomial distributions. We suppose in particular that, if $C_t = i$, $(S_{t1}, S_{t2}, \ldots, S_{tq})$ has the multinomial distribution with parameters n_t (which is known) and $p_{i1}, p_{i2}, \ldots, p_{iq} = 1 - \sum_{j=1}^{q-1} p_{ij}$. There are therefore $m^2 - m + (q-1)m = m^2 + m(q-2)$ parameters. In addition to the usual constraints on the transition probabilities γ_{ij}, and the obvious nonnegativity requirements $p_{ij} \geq 0$, there are the m constraints $\sum_{j=1}^{q-1} p_{ij} \leq 1$, one for each state i of the Markov chain, on the $(q-1)m$ independently determined multinomial probabilities.

One way in which one might view such processes is that they provide models for time series of discrete compositional data. As most of the models for compositional data are based on continuous distributions (see e.g. Aitchison (1986) or Grunwald (1987)), this may be a useful perspective. The case $n_t = 1$, it will be noted, provides a model for a single categorical time series: at each time point there is one observation, which falls into one of the q categories. In subsection 3.3.3 we shall deal specifically with models of this kind.

3.3.1 The likelihood

The computation of the likelihood of observations $\underline{s}_1, \ldots, \underline{s}_T$ from a general multinomial-HM model differs little from the case of a binomial-HM model: the only difference is that the binomial probabilities

$${}_t\pi_{s_t i} = \binom{n_t}{s_t} p_i^{s_t} (1 - p_i)^{n_t - s_t}$$

are replaced by the multinomial probabilities

$$ {}_t\pi_{\underline{s}_t i} = \mathrm{P}(\underline{S}_t = \underline{s}_t \mid C_t = i) = \binom{n_t}{s_{t1}, s_{t2}, \ldots, s_{tq}} p_{i1}^{s_{t1}} p_{i2}^{s_{t2}} \cdots p_{iq}^{s_{tq}}. $$

Otherwise the computation proceeds as before. The likelihood is therefore given by

$$ \sum_{i_1,\ldots,i_T=1}^{m} \left(\delta_{i_1} \gamma_{i_1 i_2} \cdots \gamma_{i_{T-1} i_T}\right) \left({}_1\pi_{\underline{s}_1 i_1} \cdots {}_T\pi_{\underline{s}_T i_T}\right) = \delta\, {}_1\lambda(\underline{s}_1)\Gamma \cdots \Gamma\, {}_T\lambda(\underline{s}_T)\mathbf{1}', $$

where

$$ {}_t\lambda(\underline{s}) = \mathrm{diag}\left({}_t\pi_{\underline{s}1}, \ldots, {}_t\pi_{\underline{s}m}\right). $$

In maximizing the likelihood in order to estimate parameters, one must observe all the constraints noted above.

3.3.2 Marginal properties and cross-correlations

Since $\{S_{tj} : t \in \mathbf{N}\}$ is, for each j, a binomial-hidden Markov model, the mean, variance, autocorrelation and distributional properties of $\{S_{tj} : t \in \mathbf{N}\}$ are exactly as derived in the preceding chapter. For instance, the mean and variance are given by

$$ \mathrm{E}(S_{tj}) = n_t \sum_{i=1}^{m} \delta_i p_{ij} = n_t \delta p'_{(j)} $$

and

$$ \begin{aligned} \mathrm{Var}(S_{tj}) &= n_t(n_t - 1) \sum_i \delta_i p_{ij}^2 + n_t \sum_i \delta_i p_{ij} - (n_t \sum_i \delta_i p_{ij})^2 \\ &= n_t(n_t - 1)\delta P_{(j)} p'_{(j)} + n_t \delta p'_{(j)} - n_t^2 (\delta p'_{(j)})^2 \\ &= n_t^2 \left(\delta P_{(j)} p'_{(j)} - (\delta p'_{(j)})^2\right) + n_t(\delta p'_{(j)} - \delta P_{(j)} p'_{(j)}), \end{aligned} \qquad (3.6) $$

where we define $p_{(j)} = (p_{1j}\ p_{2j}\ \ldots\ p_{mj})$ and $P_{(j)} = \mathrm{diag}(p_{(j)})$. In order to determine the cross-correlations $\mathrm{Corr}(S_{t1}, S_{t+k\,2})$ we therefore need in addition only $\mathrm{E}(S_{t1} S_{t+k\,2})$. (There is no loss of generality in considering only

categories 1 and 2, as the categories can if necessary be renumbered.) We deal with the cases $k = 0$ and $k > 0$ separately.

Firstly,

$$\mathrm{E}(S_{t1}S_{t2}) = \sum_{i=1}^{m} \delta_i \mathrm{E}(S_{t1}S_{t2} \mid C_t{=}i) = \sum_i \delta_i n_t(n_t - 1)p_{i1}p_{i2},$$

since the conditional joint distribution of S_{t1} and S_{t2} is multinomial (more precisely, trinomial). Hence

$$\mathrm{E}(S_{t1}S_{t2}) = n_t(n_t - 1)\delta P_{(1)}p'_{(2)}$$

and

$$\mathrm{Cov}(S_{t1}, S_{t2}) = n_t(n_t - 1)\delta P_{(1)}p'_{(2)} - n_t^2(\delta p'_{(1)})(\delta p'_{(2)}),$$

which yields the correlation of S_{t1} and S_{t2} on division by the appropriate expressions for the standard deviations of S_{t1} and S_{t2}: see equation (3.6) above.

Secondly, note that for $k \in \mathbf{N}$

$$\begin{aligned}
\mathrm{E}(S_{t1}S_{t+k\,2}) &= \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) \mathrm{E}(S_{t1}S_{t+k\,2} \mid C_t{=}i, C_{t+k}{=}j) \\
&= \sum_{i,j} \delta_i \gamma_{ij}(k)(n_t p_{i1})(n_{t+k} p_{j2}) \\
&= n_t n_{t+k} \delta P_{(1)} \Gamma^k p'_{(2)},
\end{aligned}$$

and

$$\mathrm{Cov}(S_{t1}, S_{t+k\,2}) = n_t n_{t+k} \left(\delta P_{(1)} \Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)}) \right).$$

Division by the standard deviations of S_{t1} and $S_{t+k\,2}$ then gives us the correlation of S_{t1} and $S_{t+k\,2}$, which is:

$$(1{+}\alpha_1/n_t)^{-1/2}(1{+}\alpha_2/n_{t+k})^{-1/2} \frac{\delta P_{(1)}\Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)})}{\left(\delta P_{(1)}p'_{(1)} - (\delta p'_{(1)})^2\right)^{1/2} \left(\delta P_{(2)}p'_{(2)} - (\delta p'_{(2)})^2\right)^{1/2}},$$

where

$$\alpha_j = \frac{\delta p'_{(j)} - \delta P_{(j)} p'_{(j)}}{\delta P_{(j)} p'_{(j)} - (\delta p'_{(j)})^2} .$$

In the case $m = 2$, i.e. if the underlying Markov chain has two states, we have:

$$\begin{aligned}
\delta P_{(j)} p'_{(j)} - (\delta p'_{(j)})^2 &= \delta_1 \delta_2 (p_{2j} - p_{1j})^2 \\
\delta p'_{(j)} - \delta P_{(j)} p'_{(j)} &= \delta_1 p_{1j}(1 - p_{1j}) + \delta_2 p_{2j}(1 - p_{2j}) \\
\delta P_{(1)} p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)}) &= \delta_1 \delta_2 (p_{21} - p_{11})(p_{22} - p_{12}) \\
\delta P_{(1)} p'_{(2)} &= \delta_1 p_{11} p_{12} + \delta_2 p_{21} p_{22} \\
\delta P_{(1)} \Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)}) &= \delta_1 \delta_2 (p_{21} - p_{11})(p_{22} - p_{12}) w^k .
\end{aligned}$$

(As usual, w denotes $1 - \gamma_1 - \gamma_2$.) Hence we have for $m = 2$ the following results for the variances and covariances, and the cross-correlation of order k:

$$\begin{aligned}
\mathrm{Var}(S_{tj}) &= n_t^2 \left(\delta P_{(j)} p'_{(j)} - (\delta p'_{(j)})^2 \right) + n_t (\delta p'_{(j)} - \delta P_{(j)} p'_{(j)}) \\
&= n_t^2 \delta_1 \delta_2 (p_{2j} - p_{1j})^2 + n_t \{ \delta_1 p_{1j}(1 - p_{1j}) + \delta_2 p_{2j}(1 - p_{2j}) \}; \\
\mathrm{Cov}(S_{t1}, S_{t2}) &= n_t^2 \left(\delta P_{(1)} p'_{(2)} - \delta p'_{(1)} \delta p'_{(2)} \right) - n_t \delta P_{(1)} p'_{(2)} \\
&= n_t^2 \delta_1 \delta_2 (p_{21} - p_{11})(p_{22} - p_{12}) - n_t (\delta_1 p_{11} p_{12} + \delta_2 p_{21} p_{22});
\end{aligned}$$

and, for all $k \in \mathbf{N}$,

$$\begin{aligned}
\mathrm{Cov}(S_{t1}, S_{t+k\,2}) &= n_t n_{t+k} \left(\delta P_{(1)} \Gamma^k p'_{(2)} - \delta p'_{(1)} \delta p'_{(2)} \right) \\
&= n_t n_{t+k} \delta_1 \delta_2 (p_{21} - p_{11})(p_{22} - p_{12}) w^k
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Corr}(S_{t1}, S_{t+k\,2}) \\
&= n_t n_{t+k} \delta_1 \delta_2 (p_{21} - p_{11})(p_{22} - p_{12}) w^k / \left(\mathrm{Var}(S_{t1}) \mathrm{Var}(S_{t+k\,2}) \right)^{1/2} \\
&= (1 + \alpha_1 / n_t)^{-1/2} (1 + \alpha_2 / n_{t+k})^{-1/2} \mathrm{sgn}((p_{21} - p_{11})(p_{22} - p_{12})) w^k ,
\end{aligned}$$

where for $j = 1, 2$

$$\alpha_j = \frac{\delta_1 p_{1j}(1 - p_{1j}) + \delta_2 p_{2j}(1 - p_{2j})}{\delta_1 \delta_2 (p_{2j} - p_{1j})^2}.$$

From the above it is clear that, if n_t is constant with respect to t (and $m = 2$), the cross-correlation of order $k \in \mathbf{N}$ depends on k only through w^k, and therefore falls off geometrically with increasing k.

3.3.3 A model for categorical time series

We now consider the case $n_t = 1$ (and m general), the model for categorical time series. This case is of considerable practical interest, as models for categorical time series do seem rare. The only approaches that seem to be available for general categorical series are the observation-driven models of Kaufmann (1987) and Fahrmeir and Kaufmann (1987) (see section 1.7), and the sequency domain approach of Stoffer, which was described in section 1.10. The models we discuss here, being parameter-driven, may be applicable to data for which observation-driven models are inappropriate.

From the general results established above we have here:

$$\begin{aligned} \mathrm{E}(S_{tj}) &= \delta p'_{(j)} \\ \mathrm{Var}(S_{tj}) &= \delta p'_{(j)} - (\delta p'_{(j)})^2 \\ \mathrm{Cov}(S_{t1}, S_{t2}) &= -(\delta p'_{(1)})(\delta p'_{(2)}), \end{aligned}$$

and, for all $k \in \mathbf{N}$,

$$\mathrm{Cov}(S_{t1}, S_{t+k\,2}) = \delta P_{(1)} \Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)}) \ .$$

The corresponding expressions for the cross-correlations follow in obvious

fashion:

$$\mathrm{Corr}(S_{t1}, S_{t2}) = \frac{-(\delta p'_{(1)})(\delta p'_{(2)})}{\left(\delta p'_{(1)} - (\delta p'_{(1)})^2\right)^{1/2}\left(\delta p'_{(2)} - (\delta p'_{(2)})^2\right)^{1/2}}$$

and, for $k \in \mathbf{N}$,

$$\mathrm{Corr}(S_{t1}, S_{t+k\,2}) = \frac{\delta P_{(1)}\Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)})}{\left(\delta p'_{(1)} - (\delta p'_{(1)})^2\right)^{1/2}\left(\delta p'_{(2)} - (\delta p'_{(2)})^2\right)^{1/2}} .$$

It is convenient to repeat here the definitions of some of the notation being used throughout section 3.3, and to note how it specializes in the present case. Given $C_t = i$, the probability that S_{tj}, the jth component of $\underline{S}_t$, equals 1 (and $S_{tl} = 0$ for all $l \neq j$) is p_{ij}. Hence $\sum_{j=1}^{q} p_{ij} = 1$ for each i from 1 to m. The vector $p_{(j)}$ is defined as $(p_{1j}, \ldots, p_{mj})$, and the matrix $P_{(j)}$ as $\mathrm{diag}(p_{(j)})$. Because $\sum_{j=1}^{q} p_{ij} = 1$, we have $\sum_{j=1}^{q} p_{(j)} = \mathbf{1}$ and $\sum_{j=1}^{q} P_{(j)} = I_m$.

The state-dependent probabilities ${}_t\pi_{\underline{s}i}$ and the matrix expression for the likelihood simplify considerably here. If the jth component of $\underline{s}$ is 1 (and the others are therefore 0) we have

$${}_t\pi_{\underline{s}i} = p_{ij} ,$$

and the subscript t is clearly unnecessary. It follows that

$$\begin{aligned} \lambda(\underline{s}) &= \mathrm{diag}\left(\pi_{\underline{s}1}, \ldots, \pi_{\underline{s}m}\right) \\ &= \mathrm{diag}\left(p_{1j}, \ldots, p_{mj}\right) \\ &= P_{(j)} , \end{aligned}$$

where again $\underline{s}$ is the vector with jth component 1 and the others 0. Hence the likelihood of observing categories $j_1, \ldots, j_T$ at times $1, 2, \ldots, T$ is given by

$$\delta P_{(j_1)}\Gamma P_{(j_2)}\Gamma \cdots P_{(j_T)}\mathbf{1}' .$$

This implies, for instance, that the probability of observing category j at time t, given category l is observed at time $t-1$, is

$$\frac{\delta P_{(l)} \Gamma P_{(j)} \mathbf{1}'}{\delta P_{(l)} \mathbf{1}'} .$$

3.4 Multivariate models

Consider now the following multivariate extension of hidden Markov models.

Let $\{C_t\}$ be the usual first-order Markov chain on m states, and suppose that, conditional on $C^{(T)}$, the Tq random variables $\{S_{tj} : t = 1, \ldots, T,\ j = 1, \ldots, q\}$ are mutually independent. That is, we consider the q time series $\{S_{tj} : t = 1, \ldots, T\}$, and assume that there is conditional independence across time as well as the usual conditional independence along time. An example of such a model is the multisite precipitation model of Zucchini and Guttorp (1991): in the application they describe, five binary time series represent the presence or absence of rain at each of five sites linked by a common 'climate process' $\{C_t\}$. In this section we shall discuss inter alia the properties of models slightly more general than theirs, involving binomial distributions rather than merely Bernoulli, and those of similar models involving Poisson distributions.

A more general model yet could be obtained by relaxing the assumption of conditional independence across time, and in fact such a modification is suggested by Zucchini and Guttorp in the context of rainfall stations situated in close proximity: in that case the assumption of conditional independence between stations may be unrealistic. The multinomial-hidden Markov model of section 3.3 is an example of a multivariate model in which conditional independence across time is not assumed. A further example of this type

would be a model in which the conditional distribution of the random vector $\underline{S}_t = (S_{t1}, S_{t2}, \ldots, S_{tq})$ is a multivariate Poisson with parameters determined by the current state of the underlying Markov chain.

3.4.1 The likelihood function for multivariate models

We begin by giving the matrix expression for the likelihood of observations $\underline{s}_1, \ldots, \underline{s}_T$ from a general multivariate-HM model, i.e. one in which conditional independence across time is not assumed. This has effectively already been established in section 3.3. With the definitions

$$ {}_t\pi_{\underline{s}i} = \mathrm{P}(\underline{S}_t = \underline{s} \mid C_t = i) $$

and

$$ {}_t\lambda(\underline{s}) = \mathrm{diag}({}_t\pi_{\underline{s}1}, \ldots, {}_t\pi_{\underline{s}m}), $$

we have as the likelihood

$$ \delta\, {}_1\lambda(\underline{s}_1)\Gamma\, {}_2\lambda(\underline{s}_2) \cdots \Gamma\, {}_T\lambda(\underline{s}_T)\mathbf{1}'. $$

In the case of conditional independence across time, the state-dependent probabilities ${}_t\pi_{\underline{s}i}$ are given by a product:

$$ {}_t\pi_{\underline{s}_t i} = \prod_{j=1}^{q} \mathrm{P}(S_{tj} = s_{tj} \mid C_t = i). $$

The multisite precipitation model of Zucchini and Guttorp is a case in point. There the random variables S_{tj} are binary, and if p_{ij} denotes

$$ \mathrm{P}(S_{tj} = 1 \mid C_t = i) = 1 - \mathrm{P}(S_{tj} = 0 \mid C_t = i), $$

then

$$ {}_t\pi_{\underline{s}_t i} = \prod_{j=1}^{q} p_{ij}^{s_{tj}} (1 - p_{ij})^{1 - s_{tj}}. $$

3.4.2 Cross-correlations of models assuming conditional independence across time

For each j, $\{S_{tj} : t \in \mathbf{N}\}$ is a univariate hidden Markov model. We shall therefore say little about its marginal properties, i.e. the mean, variance, autocorrelation and distributional properties of $\{S_{tj} : t \in \mathbf{N}\}$ for a specific j. We consider in this subsection the cross-correlation structure of models with conditional independence across time (as well as along time) and either a Poisson or a binomial conditional distribution for each S_{tj}.

In the Poisson case, let S_{tj} have mean λ_{ij} if $C_t = i$. Define $\lambda_{(j)} = (\lambda_{1j}\ \lambda_{2j}\ \ldots\ \lambda_{mj})$ and $\Lambda_{(j)} = \mathrm{diag}(\lambda_{(j)})$. We then have for all $k \in \mathbf{N}$:

$$\begin{aligned}
\mathrm{E}(S_{t1}S_{t+k\,2}) &= \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) \mathrm{E}(S_{t1}S_{t+k\,2} \mid C_t{=}i, C_{t+k}{=}j) \\
&= \sum_{i,j} \delta_i \lambda_{i1} \gamma_{ij}(k) \lambda_{j2} \\
&= \delta \Lambda_{(1)} \Gamma^k \lambda'_{(2)}; \\
\mathrm{Cov}(S_{t1}, S_{t+k\,2}) &= \delta \Lambda_{(1)} \Gamma^k \lambda'_{(2)} - (\delta \lambda'_{(1)})(\delta \lambda'_{(2)}).
\end{aligned}$$

Because of the assumed independence across time the conclusions drawn above are valid also for $k{=}0$:

$$\mathrm{E}(S_{t1}S_{t2}) = \sum_{i=1}^{m} \delta_i \mathrm{E}(S_{t1}S_{t2} \mid C_t{=}i) = \sum_{i=1}^{m} \delta_i \lambda_{i1} \lambda_{i2} = \delta \Lambda_{(1)} \lambda'_{(2)},$$

and

$$\mathrm{Cov}(S_{t1}, S_{t2}) = \delta \Lambda_{(1)} \lambda'_{(2)} - (\delta \lambda'_{(1)})(\delta \lambda'_{(2)}).$$

Using the expression for the variance derived in subsection 2.4.1, that is:

$$\mathrm{Var}(S_{tj}) = \lambda_{(j)} D \lambda'_{(j)} + \delta \lambda'_{(j)} - (\delta \lambda'_{(j)})^2 = \delta \Lambda_{(j)} \lambda'_{(j)} + \delta \lambda'_{(j)} - (\delta \lambda'_{(j)})^2\ ,$$

we can therefore write down the correlation of S_{t1} and $S_{t+k\,2}$ for all nonnegative integers k.

The special case $m = 2$ of this model (i.e. of the Poisson-based model with independence across time) yields the following result for the cross-covariances, valid for all nonnegative integers k:

$$\mathrm{Cov}(S_{t1}, S_{t+k\,2}) = \delta_1\delta_2(\lambda_{21} - \lambda_{11})(\lambda_{22} - \lambda_{12})w^k \ ,$$

where $w = 1 - \gamma_1 - \gamma_2$. From this and the corresponding result for the variances (see subsection 2.4.1):

$$\mathrm{Var}(S_{tj}) = \delta_1\delta_2(\lambda_{2j} - \lambda_{1j})^2 + \delta\lambda'_{(j)},$$

we can write down an expression for the cross-correlation $\mathrm{Corr}(S_{t1}, S_{t+k\,2})$, valid for all nonnegative integers k.

The binomial version of the above model requires that the conditional distribution of S_{tj} $(t = 1, \ldots, T;\ j = 1, \ldots, q)$ be binomial with parameters n_{tj} (known) and p_{ij} if $C_t = i$. The results are very similar to the Poisson case. For nonnegative integers k

$$\mathrm{Cov}(S_{t1}, S_{t+k\,2}) = n_{t1}n_{t+k\,2}\left(\delta P_{(1)}\Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)})\right).$$

Since we have from subsection 2.4.2 that

$$\mathrm{Var}(S_{tj}) = n_{tj}^2\left(\delta P_{(j)}p'_{(j)} - (\delta p'_{(j)})^2\right) + n_{tj}(\delta p'_{(j)} - \delta P_{(j)}p'_{(j)}),$$

the cross-correlation is available in general. Two special cases of such binomial-based models may be of interest: $m = 2$, and $n_{tj} = 1$ for all t and j. (The latter case corresponds to the multisite precipitation model.)

For $m = 2$

$$\begin{aligned}\mathrm{Cov}(S_{t1}, S_{t+k\,2}) &= n_{t1}n_{t+k\,2}\,\delta P_{(1)}\begin{pmatrix}\delta_2 & -\delta_2\\ -\delta_1 & \delta_1\end{pmatrix}p'_{(2)}w^k\\ &= n_{t1}n_{t+k\,2}\,\delta_1\delta_2(p_{21} - p_{11})(p_{22} - p_{12})w^k\end{aligned}$$

— i.e. essentially the same conclusion as holds only for *positive* integers k in the case of multinomial-HM models. With the corresponding result for the variances (see subsection 2.4.2):

$$\mathrm{Var}(S_{tj}) = n_{tj}^2 \delta_1 \delta_2 (p_{2j} - p_{1j})^2 + n_{tj} \left\{ \delta_1 p_{1j}(1 - p_{1j}) + \delta_2 p_{2j}(1 - p_{2j}) \right\},$$

this yields the cross-correlations $\mathrm{Corr}(S_{t1}, S_{t+k\,2})$ for nonnegative integer k.

If $n_{tj} = 1$ for all t and j (and m is general), then we have the results:

$$\begin{aligned}
\mathrm{Cov}(S_{t1}, S_{t+k\,2}) &= \delta P_{(1)} \Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)}); \\
\mathrm{Var}(S_{tj}) &= \delta p'_{(j)} - (\delta p'_{(j)})^2 \text{ ; and} \\
\mathrm{Corr}(S_{t1}, S_{t+k\,2}) &= \frac{\delta P_{(1)} \Gamma^k p'_{(2)} - (\delta p'_{(1)})(\delta p'_{(2)})}{\left(\delta p'_{(1)} - (\delta p'_{(1)})^2\right)^{1/2} \left(\delta p'_{(2)} - (\delta p'_{(2)})^2\right)^{1/2}} .
\end{aligned}$$

3.4.3 Cross-correlations of models not assuming conditional independence across time

We have already discussed, in section 3.3, the cross-correlations of one class of models which does not assume conditional independence across time, the multinomial-HM models. We now indicate, as a further example, how the cross-correlation function can be obtained for a bivariate hidden Markov model in which the conditional distribution is a multivariate Poisson. (Because of the independence along time, the only cross-correlation that actually differs from those obtained for the Poisson-based models discussed in subsection 3.4.2 is the cross-correlation at lag zero.) Suppose then that, if $C_t = i$, S_{t1} and S_{t2} have the bivariate distribution with joint generating function

$$\exp(\lambda_{i1}(u - 1) + \lambda_{i2}(v - 1) + a_i(u - 1)(v - 1)) ,$$

where λ_{i1}, λ_{i2} and a_i are all positive. This bivariate g.f. is of the form implied by the multivariate Poisson distribution of Teicher (1954), and in

turn implies means λ_{i1} and λ_{i2}, and covariance a_i. (Clearly we must require that $a_i \leq (\lambda_{i1}\lambda_{i2})^{1/2}$.) Then for all $k \in \mathbf{N}$ we have as before that

$$\mathrm{E}(S_{t1}S_{t+k\,2}) = \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) \lambda_{i1} \lambda_{j2} = \delta \Lambda_{(1)} \Gamma^k \lambda'_{(2)} \; .$$

The corresponding result for lag zero is

$$\mathrm{E}(S_{t1}S_{t2}) = \sum_{i=1}^{m} \delta_i (a_i + \lambda_{i1}\lambda_{i2}) \; .$$

Since expressions are available from subsection 2.4.1 for the mean and variance of S_{tj}, the cross-covariances and -correlations can therefore be computed.

3.4.4 Multivariate models with time lags

Suppose we consider models in which there is the usual Markov chain $\{C_t\}$, and a vector of observations of which some depend on C_{t-1} and some on C_t. To illustrate the nature of such models we consider in particular one in which there are only two observations at each time point. We assume that, conditional on $C_0, C_1, \ldots, C_T$, the random variables $S_1, \ldots, S_T$ and $U_1, \ldots, U_T$ are mutually independent and the (conditional) distributions of S_t and U_t are given by:

$$\begin{aligned} \mathrm{P}(S_t = s \mid C_{t-1} = i) &= \pi_{si} \\ \mathrm{P}(U_t = u \mid C_t = j) &= \sigma_{uj} \; . \end{aligned} \tag{3.7}$$

The likelihood of T consecutive observations $(S_1, U_1), \ldots, (S_T, U_T)$ is then seen to be

$$\begin{aligned} & \sum_{i_0, i_1, \ldots, i_T = 1}^{m} \left(\delta_{i_0} \gamma_{i_0 i_1} \cdots \gamma_{i_{T-1} i_T}\right) \left(\pi_{s_1 i_0} \cdots \pi_{s_T i_{T-1}}\right) \left(\sigma_{u_1 i_1} \cdots \sigma_{u_T i_T}\right) \\ = \; & \delta A_1 A_2 \cdots A_T \mathbf{1}', \end{aligned}$$

where A_t is defined for $i, j = 1, 2, \ldots, m$ by

$$(A_t)_{ij} = \gamma_{ij} \pi_{s_t i} \sigma_{u_t j} \; .$$

(In fact the form of A_t suggests that it is not difficult to generalize this result to a model in which the distributions of S_t and U_t depend on both C_{t-1} and C_t. If we define:

$$\pi_{s;ij} = \mathrm{P}(S_t = s \mid C_{t-1} = i, C_t = j)$$
$$\sigma_{u;ij} = \mathrm{P}(U_t = u \mid C_{t-1} = i, C_t = j)$$
$$(B_t)_{ij} = \gamma_{ij} \pi_{s_t;ij} \sigma_{u_t;ij} \ ,$$

then the likelihood can be written as $\delta B_1 B_2 \cdots B_T \mathbf{1}'$. We shall however not pursue this generalization, as a more efficient way of obtaining properties of such a process would to be redefine the parameter process to be $\{X_t\}$, where $X_t = (C_{t-1}, C_t)$, and treat the model as a standard hidden Markov model of the type described in Chapter 2.)

For the models defined by the equations (3.7), there is nothing new that needs to be said about the marginal properties of $\{U_t\}$, or for that matter $\{S_t\}$, although in the latter case it is notationally convenient to define $R_{t-1} = S_t$ and then consider the standard hidden Markov model $\{R_t\}$. In obtaining the cross-correlations of $\{S_t\}$ and $\{U_t\}$, this 'renaming' is also useful, since one can then apply the results of subsection 3.4.2 to $\{R_t\}$ and $\{U_t\}$.

3.4.5 Multivariate models in which some variables are discrete and the others continuous

Suppose there is the usual hidden Markov chain $\{C_t\}$ and that it provides the parameters for the joint distribution of a discrete-valued time series (e.g. the wet-dry sequence at some site) and an associated continuous series (e.g. another climatic variable like humidity). Assume that there is the usual conditional independence along time, but not necessarily across time. Conditional on $\{C_t\}$, let the joint distribution of S_t (discrete) and U_t (continuous) be given by the joint probability mass/density function $f_i(s, u)$. If there is

indeed conditional independence across time, then

$$f_i(s,u) = \pi_{si} g_i(u) \ , \tag{3.8}$$

where $\pi_{\cdot i}$ and g_i are (in general) the marginal probability mass and density functions deducible from f_i. What follows, however, refers to the more general model unless otherwise indicated.

The likelihood is given by:

$$\sum_{i_1,\ldots,i_T=1}^{m} \left(\delta_{i_1}\gamma_{i_1 i_2}\cdots\gamma_{i_{T-1}i_T}\right)\prod_{t=1}^{T} f_{i_t}(s_t,u_t) = \delta\lambda(s_1,u_1)\Gamma\lambda(s_2,u_2)\cdots\Gamma\lambda(s_T,u_T)\mathbf{1}' \ ,$$

where $\lambda(s,u) = \mathrm{diag}\,(f_1(s,u),\ldots,f_m(s,u))$. The marginal properties of $\{S_t\}$, the discrete variable, involve nothing essentially new. The marginal properties of $\{U_t\}$ merely require the use of integration rather than summation in various places. For instance

$$\begin{aligned} \mathrm{E}(U_t^k) &= \mathrm{E}\left(\mathrm{E}(U_t^k \mid C^{(T)})\right) \\ &= \sum_{i=1}^{m} \delta_i \int_{-\infty}^{\infty} u^k g_i(u)\, du \ . \end{aligned}$$

The cross-correlations can therefore be computed once we have found $\mathrm{E}(S_t U_{t+k})$ and $\mathrm{E}(U_t S_{t+k})$ for nonnegative integers k. For $k \in \mathbf{N}$ we have:

$$\begin{aligned} \mathrm{E}(S_t U_{t+k}) &= \sum_{i,j=1}^{m} \delta_i \gamma_{ij}(k) \mathrm{E}(S_t \mid C_t = i)\mathrm{E}(U_{t+k} \mid C_{t+k} = j) \\ &= \sum_{i,j} \delta_i \gamma_{ij}(k) \lambda_i \mu_j \\ &= \delta\Lambda\Gamma^k\mu' \ , \end{aligned}$$

where λ_i and μ_i are the conditional means of S_t and U_t given $C_t = i$, $\lambda = (\lambda_1,\ldots,\lambda_m)$, $\mu = (\mu_1,\ldots,\mu_m)$ and $\Lambda = \mathrm{diag}(\lambda)$. Similarly, for $k \in \mathbf{N}$, we have

$$\mathrm{E}(U_t S_{t+k}) = \delta M \Gamma^k \lambda' \ ,$$

where M denotes diag(μ). The cross-correlation at lag zero requires

$$\mathrm{E}(S_t U_t) = \sum_{i=1}^{m} \delta_i \mathrm{E}(S_t U_t \mid C_t = i) \ ,$$

with this last conditional expectation being given by

$$\sum_s \int_{-\infty}^{\infty} su f_i(s,u)\, du \ ,$$

which equals $\lambda_i \mu_i$ if assumption (3.8) holds.

It is clear that this class of models can fairly routinely be extended, if necessary, to cater for (i) several discrete components and several continuous; (ii) the joint distribution of S_t and U_t depending on t as well as on the state currently occupied by the Markov chain.

3.5 Models with state-dependent probabilities depending on covariates

Hidden Markov models can be modified to allow for the influence of covariates by postulating dependence of the state-dependent probabilities ${}_t\pi_{si}$ on those covariates. This opens the way for such models to incorporate time trend and seasonality, for instance. We take $\{C_t\}$ to be the usual Markov chain, and we suppose, in the case of Poisson-HM models, that the conditional mean ${}_t\lambda_i$ depends on the (row) vector x_t of q covariates, for instance as follows:

$$\log \ {}_t\lambda_i = \beta_i x_t' \ .$$

(We continue here to use the convention that the subscript t before a symbol indicates time-dependence of the quantity concerned.) In the case of binomial-HM models, the corresponding assumption is that

$$\text{logit} \ {}_t p_i = \beta_i x_t' \ .$$

To be more specific, the elements of x_t could include a constant, time (t), sinusoidal components expressing seasonality (e.g. $\cos(2\pi t/r)$ and $\sin(2\pi t/r)$ for some positive integer r), and any other covariates thought relevant. A binomial-HM model with

$$\text{logit } {}_tp_i = \beta_{i1} + \beta_{i2}t + \beta_{i3}\cos(2\pi t/r) + \beta_{i4}\sin(2\pi t/r) + \beta_{i5}y_t + \beta_{i6}z_t$$

allows for a (logistic-linear) time trend, r-period seasonality and the influence of covariates y_t and z_t, in the state-dependent 'success probabilities' ${}_tp_i$. Further sine-cosine pairs can be included if necessary. Similar models for the log of the conditional mean ${}_t\lambda_i$ are possible in the Poisson-HM case. Clearly link functions other than the canonical ones used here could be considered too.

The expression $\delta\, {}_1\lambda(s_1)\Gamma\, {}_2\lambda(s_2)\cdots\Gamma\, {}_T\lambda(s_T)\mathbf{1}'$ for the likelihood of T consecutive observations $s_1,\ldots,s_T$ remains valid for these models involving covariates: what changes is the precise definition of ${}_t\pi_{si}$, and hence of

$${}_t\lambda(s) = \text{diag}({}_t\pi_{s1},\ldots,{}_t\pi_{sm})\ .$$

Expressions for moments, including autocorrelations, are found by the usual methods, although here the autocorrelations are less useful than they are in the case of stationary models. For the Poisson-HM models with covariates we have

$$\begin{aligned}
\text{E}(S_t) &= \textstyle\sum_{i=1}^m \delta_i\, {}_t\lambda_i\ ; \\
\text{E}(S_t^2) &= \textstyle\sum_i \delta_i\,({}_t\lambda_i + ({}_t\lambda_i)^2)\ \text{ ; and} \\
\text{E}(S_tS_{t+k}) &= \textstyle\sum_{i,j=1}^m \delta_i\gamma_{ij}(k)\, {}_t\lambda_i\, {}_{t+k}\lambda_j\ ,
\end{aligned}$$

where $k \in \mathbf{N}$ and in all cases ${}_t\lambda_i = \exp(\beta_i x_t')$. For the binomial-HM model with covariates:

$$\begin{aligned}
\text{E}(S_t) &= n_t \textstyle\sum_i \delta_i\, {}_tp_i\ ; \\
\text{E}(S_t^2) &= \textstyle\sum_i (n_t\, {}_tp_i(1-{}_tp_i) + n_t^2\, {}_tp_i^2)\ \text{ ; and} \\
\text{E}(S_tS_{t+k}) &= n_tn_{t+k}\textstyle\sum_{i,j}\delta_i\gamma_{ij}(k)\, {}_tp_i\, {}_{t+k}p_j\ ,
\end{aligned}$$

where $k \in \mathbf{N}$ and ${}_tp_i = \exp(\beta_i x_t')/(1+\exp(\beta_i x_t'))$. Expressions for $\mathrm{Var}(S_t)$, $\mathrm{Cov}(S_t, S_{t+k})$ and $\mathrm{Corr}(S_t, S_{t+k})$ then follow for the two kinds of model.

3.6 Models in which the Markov chain is homogeneous but not assumed stationary

If the Markov chain $\{C_t\}$ underlying a hidden Markov model is homogeneous but not necessarily stationary, the model is of the type discussed by Leroux and Puterman (1992) and described in section 2.7. It will be recalled from that section that the likelihood can in that case be maximized by taking the Markov chain to start (with probability 1) at a particular state, and applying the EM algorithm to estimate the transition probabilities γ_{ij} and the remaining parameters, which in the application of Leroux and Puterman are the means λ_j of the Poisson conditional distributions. Since models of this kind have been discussed in some detail by those authors, we shall not dwell on the topic here.

3.7 Models in which the Markov chain is nonhomogeneous

A further way in which one can seek to accommodate time trend and seasonality in hidden Markov models is to drop the assumption that the Markov chain is homogeneous, and assume instead that the transition probabilities depend on a vector of covariates x_t. We indicate here one possible way of incorporating the covariates into the transition probabilities.

Consider a model based on a two-state Markov chain $\{C_t\}$ with

$$\begin{aligned} \mathrm{P}(C_t=2 \mid C_{t-1}=1) &= {}_t\gamma_1, \\ \mathrm{P}(C_t=1 \mid C_{t-1}=2) &= {}_t\gamma_2 \end{aligned}$$

and, for $i=1,2$

$$\text{logit } {}_t\gamma_i = \beta_{(i)} x_t' \, .$$

For example, a model incorporating 12-period seasonality is that with

$$\text{logit } {}_t\gamma_i = \beta_{(i)1} + \beta_{(i)2}\cos(2\pi t/12) + \beta_{(i)3}\sin(2\pi t/12) \, .$$

In general the above assumption on logit ${}_t\gamma_i$ implies that the transition probability matrix, for transitions between times $t-1$ and t, is given by

$${}_t\Gamma = \begin{pmatrix} \frac{1}{1+\exp(\beta_{(1)}x_t')} & \frac{\exp(\beta_{(1)}x_t')}{1+\exp(\beta_{(1)}x_t')} \\ \frac{\exp(\beta_{(2)}x_t')}{1+\exp(\beta_{(2)}x_t')} & \frac{1}{1+\exp(\beta_{(2)}x_t')} \end{pmatrix} \, .$$

Extension of this model to the case $m > 2$ presents difficulties, but they appear not to be insuperable. Aitchison (1986, Chapter 6) presents several one-to-one transformations (e.g. the generalized, or 'additive', logistic) from $\mathbf{R}^n$ to the n-dimensional unit simplex. By applying such a transform to $n=m-1$ appropriate linear functions $\beta x_t'$ of the covariates we can model the $m-1$ off-diagonal transition probabilities in each row in a fashion consistent with the row sum constraint. Since this has to be done separately for each of the m rows of the transition probability matrix, the number of parameters used may be large unless some restrictions are imposed on the coefficients appearing in the linear functions.

Clearly the incorporation of covariates into the Markov chain is more difficult than incorporating them into the state-dependent probabilities. The main reason why it might be worthwhile is that the resulting Markov chain may have a useful substantive interpretation: e.g. as a meaningful 'climate

process' which is itself complicated but determines rainfall probabilities at several sites in fairly simple fashion.

One important difference between the class of models proposed here and other hidden Markov models (and a consequence of the nonhomogeneity) is that we cannot assume there is a stationary distribution for the Markov chain. We assume instead that there is some initial distribution δ at time $t = 1$. This has implications for the way in which parameters may be estimated. We now have to estimate three sets of parameters: the initial probabilities δ, the parameters appearing in the transition probabilities, and the parameters determining the state-dependent probabilities ${}_t\pi_{si}$. The unconditional likelihood is a convex linear combination of likelihood values conditioned on a particular initial state, and may for instance be maximized (as in the work of Leroux and Puterman) by choosing as the initial state that one which produces the largest maximized conditional likelihood. The EM algorithm may be used to estimate the other parameters. At the M-step the two parts of the complete-data log-likelihood can be maximized separately. The part involving the state-dependent probabilities may be maximized exactly as in the work of Leroux and Puterman, hence by closed-form expressions in the case of Poisson or binomial conditional distributions. The maximization of the part involving the transition probabilities is more complicated, and will need numerical solution except possibly in very special cases.

3.8 Models combining the binomial distribution with a Poisson-hidden Markov model

So far, when discussing binomial-HM models, we have always taken n_t, the number of trials at time t, to be a known constant. However n_t could itself be an observation at time t on some nonnegative integer-valued random process $\{N_t\}$, for example a Poisson-HM model. One could then take the process $\{S_t\}$ to be such that, conditional on $\{N_t\}$, $\{S_t\}$ is a binomial-HM model with the numbers of trials n_t supplied by $\{N_t\}$, and driven by the same Markov chain as drives $\{N_t\}$ or even by another independent one. This includes as a special case the possibility that, given $\{N_t\}$, the observations S_t are independent binomial random variables.

To motivate this class of models, we consider one fairly simple example thereof. Suppose that $\{N_t\}$ is a Poisson-HM model based on the two-state Markov chain $\{C_t\}$, with

$$\log {}_t\lambda_i = a_i + bt + c\cos(2\pi t/12) + d\sin(2\pi t/12) ,$$

for $i = 1, 2$ and $t = 1, \ldots, T$. Suppose further that, given $N^{(T)}$, the random variables S_t $(t = 1, \ldots, T)$ are independent binomial $(N_t, {}_tp)$, with

$$\text{logit } {}_tp = \alpha + \beta t . \tag{3.9}$$

The observations N_t and S_t could represent respectively the total number of births and the number of deliveries by a particular method, at a hospital in month t. Such a model for $\{N_t\}$ and $\{S_t\}$ would have a total of nine parameters, and would allow for time trend both in the number of births and in the proportion by the particular method of interest, for an annual

cycle in the number of births, and for two underlying states which influence the mean number of births occurring at the hospital in a month. (Transport difficulties caused by weather or other factors could sometimes influence the number of births occurring at the hospital, and would be allowed for by the two states.) The two parameters α and β can be estimated, independently of the other parameters, by ordinary logistic regression, and the remaining seven parameters by numerical maximization of the Poisson-HM likelihood derived from the observations $N_1, \ldots, N_T$.

A modification of the above model which could perhaps prove useful in the application described is to add to the expression for logit ${}_tp$ in equation (3.9) a term involving n_t or some function thereof. This could be used to accommodate a 'busy period' effect: certain methods of delivery might be preferred by the obstetricians during particularly busy months. Other variations can similarly be incorporated into either the model $\{N_t\}$ or the model $\{S_t\}$.

The most general model of this kind which we shall consider here is the following. The process $\{N_t\}$ is a Poisson-HM model with underlying stationary Markov chain $\{C_t\}$. Given $C^{(T)}$ and $N^{(T)}$, the random variables $S_1, \ldots, S_T$ are independent binomial with parameters N_t and ${}_tp_i$, where $C_t = i$. (The same Markov chain is taken to drive the two hidden Markov models involved.) The likelihood of the observations $N_1, \ldots, N_T$ and $S_1, \ldots, S_T$ is given by

$$\sum_{i_1,\ldots,i_T=1}^{m} \left(\delta_{i_1}\gamma_{i_1 i_2}\cdots\gamma_{i_{T-1}i_T}\right)\left({}_1\pi_{n_1 i_1}\cdots{}_T\pi_{n_T i_T}\right)\left({}_1\nu_{s_1;n_1 i_1}\cdots{}_T\nu_{s_T;n_T i_T}\right) ,$$

where we define

$${}_t\pi_{ni} = \mathrm{P}(N_t = n \mid C_t = i) \quad \text{(the appropriate Poisson probability)}$$

and

$$ {}_t\nu_{s;ni} = \mathrm{P}(S_t = s \mid N_t = n, C_t = i) = \binom{n}{s} {}_tp_i^s(1 - {}_tp_i)^{n-s} . $$

If we make the further definition that D_t is the diagonal matrix with ith diagonal element ${}_t\pi_{n_t i}\, {}_t\nu_{s_t;n_t i}$, we can therefore write the likelihood as

$$ \delta D_1 \Gamma D_2 \Gamma \cdots D_T \mathbf{1}' , $$

and use this expression for parameter estimation.

3.9 Discussion

Although many variations on a hidden Markov theme have been presented here, it will be clear that the list is not exhaustive. One could, for instance, consider models allowing for some, preferably simple, conditional dependence along time. Which further variations are worth pursuing will be determined by applications. As a general class of models for discrete-valued series, the hidden Markov models are certainly very flexible and able to accommodate the characteristics of many types of data.

The device of using an underlying Markov chain to introduce dependence between variables that are otherwise independent results in a parsimony of parametrization that is not easily achieved by competing models. Compare for instance second-order Markov chains on three states with binomial-hidden Markov models with $n_t = 2$ for all t. (In all these models the observations take one of three values.) The general second-order Markov chain has eighteen parameters and the corresponding Raftery model has seven. The hidden Markov model has m^2 parameters if its underlying Markov chain has m states. Hence even a hidden Markov model with $m = 4$ has fewer parameters than a general second-order Markov chain, and a hidden Markov model with

$m=2$ has fewer than the Raftery model.

Furthermore, the relative ease with which the likelihood may be evaluated and maximized with respect to the parameters greatly adds to the usefulness of hidden Markov models as practical statistical tools. The particular class of hidden Markov models we have chosen to concentrate on in this thesis, consisting of those based on a stationary Markov chain, has the added advantage that the theoretical autocorrelation function can be found easily and compared with the sample autocorrelations for model identification purposes. This use of the ACF in the context of hidden Markov models seems not to have been explored previously. While it is true that the EM algorithm is not as easily applied to hidden Markov models of this class as to those of Leroux and Puterman, the availability of sophisticated optimization software and the conceptual simplicity of direct numerical maximization mean that this is not really a disadvantage.

We now present in Chapter 4 a number of illustrative examples of applications of the models introduced in this thesis.

Chapter 4

Examples of applications

4.1 Introduction

The purpose of this chapter is to describe the application of some of the models introduced in Chapters 2 and 3, to data sets from a variety of subjects. In the application of any new methodology it is helpful to have available both examples of its use which appear successful and examples which appear less so. We therefore report here not only those applications in which hidden Markov models have turned out to be useful models, but also some in which the nature of such models seems to make them inappropriate, or at least less appropriate than some competing model. It is hoped that this will be a more useful contribution to knowledge than the presentation of a few carefully selected 'successful' applications would be.

4.2 The durations of successive eruptions of the 'Old Faithful' geyser

Azzalini and Bowman (1990) have presented an interesting analysis of data on eruptions of the 'Old Faithful' geyser in the Yellowstone National Park in

the U.S.A. The data, which appear in full in the abovementioned reference, consist of two series of length 299, collected continuously from 1 August to 15 August 1985. The first series is of the durations, d_t, of successive eruptions, and the second is of the waiting times, w_t, preceding those eruptions (defined as the differences between the starting times of the relevant eruptions). It is true of both series that most of the observations can be described as either long or short, with very few observations intermediate in length, and with relatively low variation within the low and high groups. It is therefore very natural to treat these series as binary time series: Azzalini and Bowman do so by discretizing the two series at 3 and 68 minutes respectively, denoting short by 0 and long by 1. (There is, in respect of the durations series, the complication that some of the durations were observed only as short, medium or long, and the medium durations have to be treated as either short or long. This will be discussed further in due course.) It emerges that $\{D_t\}$ and $\{W_t\}$, the discretized versions of the series $\{d_t\}$ and $\{w_t\}$, are very similar — almost identical, in fact — and Azzalini and Bowman therefore concentrate on the series $\{D_t\}$ as representing most of the information relevant to the state of the system. We shall do the same here.

On examination of the series $\{D_t\}$ one notices that 0 is always followed by 1, and 1 by either 0 or 1. If we treat the two eruptions of medium duration as longs, i.e. if we represent them by 1, we can summarize the data as follows. There are:

105 zeroes;

194 ones;

no transitions from 0 to 0;

104 transitions from 0 to 1;

105 transitions from 1 to 0;

89 transitions from 1 to 1;

69 transitions from 0,1 to 0;

35 transitions from 0,1 to 1;

104 transitions from 1,0 to 1;

35 transitions from 1,1 to 0; and

54 transitions from 1,1 to 1.

In brief, what Azzalini and Bowman first did was to fit a (first-order) Markov chain model. This model seemed quite plausible from a geophysical point of view, but did not match the sample ACF at all well. They then fitted a second-order Markov chain model, which matched the ACF much better, but they did not attempt a geophysical interpretation for this second model. As there appears to be a slight discrepancy in their work, it is necessary to describe in more detail what Azzalini and Bowman did, before proceeding to fit hidden Markov models and compare them with corrected versions of their models.

With the convention that a medium is treated as a long, and using the estimator of the ACF described on pp. 32–33 of Box and Jenkins (1976), Azzalini and Bowman estimated the ACF and PACF of $\{D_t\}$ as appears in Table 4.1. Since the ACF is not even very approximately of the form α^k, a Markov chain is not a satisfactory model. (Azzalini and Bowman note 'moreover' that the PACF should be close to zero after lag 1 if a Markov chain is to be regarded as an adequate model. However, since $\rho_k = \alpha^k$ for all positive integers k implies that $\phi_{rr} = 0$ for $r > 1$ (see section 1.2), the high value of $\hat{\phi}_{22}$ is not really additional evidence against the adequacy of a Markov chain, merely a restatement of the evidence provided by the ACF.) Azzalini and Bowman therefore fitted a second-order Markov chain, which turned out not to be consistent with a first-order model. They mention also that they fitted a third-order model, which did produce estimates consistent with a second-order model.

Table 4.1: Geyser data. Sample ACF and PACF of the series $\{D_t\}$.

k	1	2	3	4	5	6	7	8
$\hat{\rho}_k$	−.538	.478	−.346	.318	−.256	.208	−.161	.136
$\hat{\phi}_{kk}$	−.538	.266	−.021	.075	−.021	−.009	.010	.006

An estimate of the transition probability matrix of the first-order Markov chain, based on maximizing the likelihood conditional on the first observation, is

$$\begin{pmatrix} 0 & 1 \\ \frac{105}{194} & \frac{89}{194} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ .5412 & .4588 \end{pmatrix}. \tag{4.1}$$

Although it is not central to this discussion, it is worth noting that unconditional maximum likelihood is very easy in this case. Because there are no transitions from 0 to 0, an explicit formula applies: see Bisgaard and Travis (1991). The result is that the t.p.m. is estimated as

$$\begin{pmatrix} 0 & 1 \\ .5404 & .4596 \end{pmatrix}. \tag{4.2}$$

The serves to confirm as reasonable the expectation that, for a series of length 299, conditional maximum likelihood differs very little from unconditional.

The discrepancy referred to above is that Azzalini and Bowman report as their estimated t.p.m. $\begin{pmatrix} 0 & 1 \\ .557 & .443 \end{pmatrix}$. This appears to arise as $\begin{pmatrix} 0 & 1 \\ \frac{107}{192} & \frac{85}{192} \end{pmatrix}$, whch is in turn based on the convention that a medium is treated as a *short* (and on conditional maximum likelihood). Azzalini and Bowman therefore seem to have used one convention for estimating the ACF, and the other for estimating the t.p.m. This impression is strengthened by examination of the estimated t.p.m. of the second-order model. The three states used to express

the second-order model as a first-order Markov chain are, in order: (0,1), (1,0), (1,1). (The sequence (0,0) does not occur.) With the convention that a medium is treated as a long (to which we shall adhere throughout), the t.p.m. is

$$\begin{pmatrix} 0 & \frac{69}{104} & \frac{35}{104} \\ 1 & 0 & 0 \\ 0 & \frac{35}{89} & \frac{54}{89} \end{pmatrix} = \begin{pmatrix} 0 & .6635 & .3365 \\ 1 & 0 & 0 \\ 0 & .3933 & .6067 \end{pmatrix}, \tag{4.3}$$

yet Azzalini and Bowman report that the t.p.m. is

$$\begin{pmatrix} 0 & .689 & .311 \\ 1 & 0 & 0 \\ 0 & .388 & .612 \end{pmatrix},$$

which matrix presumably arises as

$$\begin{pmatrix} 0 & \frac{73}{106} & \frac{33}{106} \\ 1 & 0 & 0 \\ 0 & \frac{33}{85} & \frac{52}{85} \end{pmatrix}.$$

This last matrix is also based on the convention that a medium is treated as a short.

It is therefore necessary to correct the theoretical ACF quoted by Azzalini and Bowman for the second-order model before using it in any way. We compute the ACF of model (4.3), which has stationary distribution $\frac{1}{297}(104\ 104\ 89)$, by

$$\begin{aligned} \rho_k &= \frac{\mathrm{E}(D_t D_{t+k}) - \mathrm{E}(D_t)\mathrm{E}(D_{t+k})}{\mathrm{Var}(D_t)} \\ &= \frac{297^2 \mathrm{P}(D_t = D_{t+k} = 1) - 193^2}{193 \times 104}. \end{aligned}$$

The resulting figures for $\{\rho_k\}$ are given in Table 4.2, and match the sample ACF $\{\hat{\rho}_k\}$ well, better (as one would expect) than do Azzalini and Bowman's

own figures for $\{\rho_k\}$.

Finally we now proceed to consider hidden Markov models for the series $\{D_t\}$, and to compare the various models discussed by means of Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) of Schwarz (1978).

Binomial-hidden Markov models with $n_t = 1$ and $m = 2$, 3 or 4 were fitted to the series $\{D_t\}$. (The notation used here is as in section 2.3.) We describe the two-state model in some detail. This model has log-likelihood -127.31, $\Gamma = \begin{pmatrix} .000 & 1.000 \\ .827 & .173 \end{pmatrix}$, and $p = (.225\ 1.000)$. That is, there are two (unobserved) states, state 1 always being followed by state 2, and state 2 by state 1 with probability .827. In state 1, a long has probability .225, in state 2 it has probability 1. A convenient interpretation of this model is that it is a fairly simple stationary two-state Markov chain with some 'noise' in the first state: if the probability .225 were instead zero, the model would be exactly a Markov chain. A long has unconditional probability $\delta\lambda(1)\mathbf{1}' = .649$ (cf. 193/297=.650 for the second-order Markov chain). A long is followed by a short with probability $\delta\lambda(1)\Gamma\lambda(0)\mathbf{1}'/\delta\lambda(1)\mathbf{1}' = .541$ (cf. 105/194=.541). It is also easily established that in such a model a short is always followed by a long. In the notation of subsection 2.4.2, the ACF is given for all $k \in \mathbf{N}$ by $\rho_k = (1+\alpha)^{-1}w^k$, where $w = 1 - \gamma_1 - \gamma_2 = -.827$ and $\alpha = .529$. Hence $\rho_k = .654 \times (-.827)^k$. In Table 4.2 the resulting figures are compared with the sample ACF and with the theoretical ACF of the (corrected) second-order Markov chain model, i.e. model (4.3). It seems reasonable to conclude that the hidden Markov model fits well in this respect, not quite as well as the second-order Markov chain model as regards the first three autocorrelations, but better for longer lags.

Table 4.2: Geyser data. Autocorrelations of two models fitted compared with the sample autocorrelation function.

k	1	2	3	4	5	6	7	8
sample ACF, $\hat{\rho}_k$	−.538	.478	−.346	.318	−.256	.208	−.161	.136
ρ_k for model (4.3)	−.539	.482	−.335	.262	−.194	.147	−.110	.083
ρ_k for HM model	−.541	.447	−.370	.306	−.253	.209	−.173	.143

The parametric bootstrap, with a sample size of 100, was used to estimate the covariance matrix of the maximum likelihood estimators of the four parameters γ_1, γ_2, p_1 and p_2. That is, 100 series of length 299 were generated from the two-state model described above, and a model fitted in the usual way to each of these series. The random number generator used was that of Wichmann and Hill (1982). The sample mean vector for the four parameters is (1.000 .819 .215 1.000), and the sample covariance matrix is:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & .003303 & .001540 & 0 \\ 0 & .001540 & .002065 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The estimated standard deviations of the estimators are therefore (.000 .057 .045 .000). (The zero standard errors are of course not typical: they are a consequence of the rather special nature of the model from which we are generating the series.) As a further indication of the behaviour of the estimators we present in Table 4.3 selected percentiles of the bootstrap sample of values of $\hat{\gamma}_2$ and $\hat{p}_1$. From these bootstrap results it appears that, for this application, the maximum likelihood estimators have fairly small standard deviations and are not obviously asymmetric. It should, however, be borne

in mind that the estimate of the distribution of the estimators which is provided by the parametric bootstrap is derived under the assumption that the model fitted to the data is correct.

Table 4.3: Geyser data. Percentiles of bootstrap sample of estimators of parameters of two-state hidden Markov model.

percentile:	5th	25th	median	75th	95th
$\hat{\gamma}_2$	.709	.793	.828	.856	.886
$\hat{p}_1$	.139	.191	.218	.244	.273

There is, however, a further class of models which generalizes both the two-state second-order Markov chain and the two-state hidden Markov model as described above. This is the class of two-state second-order hidden Markov models on state-space $\{0, 1\}$. Such models (inter alia) are described in some detail in section 3.2. By using the recursion for $\nu_t(i, j)$ given in that section, with the appropriate scaling, it is almost as straightforward to compute the likelihood of a second-order model as a first-order one and to fit models by maximum likelihood. In the present example the resulting probabilities of a long eruption are .072 (state 1) and 1.000 (state 2). The underlying process (or parameter process) is a two-state second-order Markov chain with associated first-order Markov chain having transition probability matrix

$$\begin{pmatrix} 1-a & a & 0 & 0 \\ 0 & 0 & .717 & .283 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & .441 & .559 \end{pmatrix}. \tag{4.4}$$

Here a is any real number between 0 and 1, and the four states used for this purpose are, in order: (1,1), (1,2), (2,1), (2,2). The log-likelihood is

-126.90. (Clearly the state (1,1) can be disregarded above without loss of information, in which case the first row and first column are deleted from the matrix (4.4).) It should be noted that the second-order Markov chain used here as underlying process is the general four-parameter model, not the Pegram-Raftery submodel, which has three parameters. From the comparison which follows it will be seen that a hidden Markov model based on a Pegram-Raftery second-order chain is in this case not worth pursuing, because with a total of five parameters it cannot produce a log-likelihood value better than -126.90. (The two four-parameter models fitted produce values of -127.31 and -127.12.)

We now compare all the models considered on the basis of their unconditional log-likelihoods, denoted by l, and AIC and BIC. For instance, in the case of model (4.1), the first-order Markov chain,

$$l = \log(194/299) + 105\log(105/194) + 89\log(89/194) = -134.2426.$$

The comparable figure for model (4.2) is -134.2423: in view of the minute difference we shall here ignore the distinction between estimation by conditional and by unconditional maximum likelihood. For the second-order Markov chain model (4.3), l is given by

$$\log(104/297)+35\log(35/104)+69\log(69/104)+35\log(35/89)+54\log(54/89),$$

and equals -127.12 . The criteria AIC and BIC are here given by $-2l + 2k$ and $-2l + k\log 299$ respectively, where k denotes the number of parameters estimated. Table 4.4 presents a comparison of six types of model, from which it emerges that, on the basis of AIC and BIC, only the second-order Markov chain and the two-state (first-order) hidden Markov model are worth considering. In the comparison, both of these models are taken to have four parameters, because without knowledge of the data we cannot know, for

instance, that the sequence (short, short) is not possible.

Table 4.4: Geyser data. Comparison of models on the basis of AIC and BIC.

model	no. parameters	l	AIC	BIC
Markov chain	2	−134.24	272.48	279.88
second-order M. chain	4	−127.12	262.24	277.04
2-state hidden Markov	4	−127.31	262.62	277.42
3-state hidden Markov	9	−126.85	271.70	305.00
4-state hidden Markov	16	−126.59	285.18	344.39
2-state second-order HM	6	−126.90	265.80	288.00

While it is true that the second-order Markov chain seems a slightly better model on the basis of the model selection exercise described above, and possibly on the basis of the ACF, both are reasonable models capable of describing the principal features of the data without using an excessive number of parameters. The hidden Markov model perhaps has the advantage of relative simplicity, given its nature as a Markov chain with some noise in one of the states. Azzalini and Bowman note that their second-order Markov chain model requires a more sophisticated interpretation than does their first-order model, but do not provide such an interpretation. Either a longer series of observations or a convincing geophysical interpretation for one model rather than the other would be needed to take the discussion further.

We conclude this section by demonstrating how the two-state hidden Markov model may be used to provide forecasts. As it happens, the last observation in the series $\{D_t\}$ (the 299th) is zero, so that with probability one the next one is 1. We therefore give here the 2-step-ahead and joint 2- and 3-step-ahead forecast distributions implied by the model. (Higher-order joint distributions and forecasts further into the future involve no essentially

different features.) The relevant probabilities are given by ratios of likelihood values, as described in section 2.6. Given the full history, the probability that $D_{301} = 1$ (and $D_{300} = 1$) is .359. The probabilities that, given the history, $D_{301} = i$ and $D_{302} = j$ (and $D_{300} = 1$) are given in Table 4.5. For comparison we state here the corresponding figures for the second-order Markov chain model (4.3). The conditional probability that $D_{301} = 1$ is .337, and the joint forecast distribution of D_{301} and D_{302} is also given in Table 4.5.

Table 4.5: Geyser data. Joint 2- and 3-step-ahead forecast distributions for the two-state hidden Markov model (left) and the second-order Markov chain model (right).

	$j=0$	1
$i=0$	.000	.641
1	.111	.248

	$j=0$	1
$i=0$	.000	.663
1	.132	.204

In order to assess the variability of the forecast distribution supplied by the two-state hidden Markov model, the bootstrap sample already described was used as follows. Each of the 100 series generated gives rise to a model, and each such model was used to compute a joint 2- and 3-step-ahead forecast distribution conditional on the actual series observed. The sample mean of the forecast distribution is $\begin{pmatrix} .000 & .641 \\ .115 & .244 \end{pmatrix}$, or in row vector form (.000 .641 .115 .244), the corresponding sample covariance matrix is

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & .001525 & -.000620 & -.000905 \\ 0 & -.000620 & .000940 & -.000319 \\ 0 & -.000905 & -.000319 & .001224 \end{pmatrix},$$

and the sample standard deviations are $\begin{pmatrix} .000 & .039 \\ .031 & .035 \end{pmatrix}$. The forecast distribution is thereby seen to be reasonably stable in this example.

4.3 Births at Edendale hospital 1970–1986

Haines, Munoz and van Gelderen (1989) have described the fitting of Gaussian ARIMA models to various discrete-valued time series related to births occurring during a 16-year period at Edendale Hospital in Natal, South Africa. The data include (inter alia) monthly totals of mothers delivered and deliveries by various methods at the Obstetrics Unit of that hospital in the period from February 1970 to January 1986 inclusive. For the data, which were provided by Dr L.M. Haines, see Appendix B.

4.3.1 Models for the proportion Caesarean

One of the series considered by Haines et al., to which they fitted two models, was the number of deliveries by Caesarean section. From their models they drew the conclusions (in respect of this particular series) that there is a clear dependence of present on past observations, and that there is a clear linear upward trend. In this subsection we describe the fitting of (discrete-valued) Markov regression and hidden Markov models to this series. These models are of course rather different from those fitted by Haines et al. in that the latter, being based on the normal distribution, are continuous-valued. Furthermore, the discrete-valued models make it possible to model the proportion (as opposed to the number) of Caesareans performed in each month. Of the models proposed here one type is observation-driven and the other parameter-driven. The most important conclusion drawn from the discrete-valued models, and one which the Gaussian ARIMA models did not provide,

is that there is a strong upward time trend in the proportion Caesarean.

The two models which Haines et al. fitted to the time series of Caesareans performed, and which they found to fit very well, may be described as follows. Let Z_t denote the number of Caesareans in month t, and let the process $\{a_t\}$ be Gaussian white noise, i.e. uncorrelated random shocks distributed normally with zero mean and common variance σ_a^2. The first model fitted is the ARIMA(0,1,2) model with constant term:

$$\nabla Z_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}. \tag{4.5}$$

The maximum likelihood estimates of the parameters, with associated standard errors, are: $\hat{\mu} = 1.02 \pm 0.39, \hat{\theta}_1 = 0.443 \pm 0.097, \hat{\theta}_2 = 0.393 \pm 0.097$ and $\hat{\sigma}_a^2 = 449.25$. The second model is an AR(1) with linear trend:

$$Z_t = \beta_0 + \beta_1 t + \phi Z_{t-1} + a_t, \tag{4.6}$$

with parameter estimates as follows: $\hat{\beta}_0 = 120.2 \pm 8.2, \hat{\beta}_1 = 1.14 \pm 0.15, \hat{\phi} = 0.493 \pm 0.092$ and $\hat{\sigma}_a^2 = 426.52$.

Both of these models provide support for the conclusion of Haines et al. that there is a dependence of present on past observations, and a linear upward trend. Furthermore, the models are nonseasonal: the Box-Jenkins methodology used found no seasonality in the Caesareans series. The X-11-ARIMA seasonal adjustment method employed in an earlier study (Munoz, Haines and van Gelderen, 1987) did however find some evidence, albeit weak, of a seasonal pattern similar to a pattern observed in the 'total deliveries' series. This latter series shows marked seasonality, with a peak in September, and in Haines et al. (1989) is modelled by the seasonal ARIMA model $(0,1,1) \times (0,1,1)_{12}$.

It is of some interest to model the proportion, rather than the number, of Caesareans in each month. It could be the case, for instance, that any trend, dependence or seasonality apparently present in the number of Caesareans is largely inherited from the total deliveries, and a constant proportion Caesarean is an adequate model. On the other hand, it could be the case that there is an upward trend in the proportion of the deliveries that are by Caesarean and this accounts at least partially for the upward trend in the number of Caesareans. The two classes of model that we discuss in this subsection condition on the total number of deliveries in each month and seek to describe the principal features of the proportion Caesarean.

Although sixteen years of data were available, only the final eight years were used by Haines et al. in fitting the models defined in equations (4.5) and (4.6) to the Caesareans series. This was because the model structure of some of the series they considered was not stable over the full sixteen-year period. We have also used the last eight years of data to fit the models we consider. Note however that Haines et al. did use all sixteen years' data to fit their model for the total deliveries.

Now let n_t denote the total number of deliveries in month t. A very general possible model for $\{Z_t\}$ which allows for trend, dependence on previous observations and seasonality in the proportion Caesarean is as follows. Suppose that, conditional on the history $Z^{(t-1)} = \{Z_s : s \leq t-1\}$, Z_t is distributed binomially with parameters n_t and p_t, where for some positive integer q:

$$\begin{aligned} \text{logit } p_t &= \alpha_1 + \alpha_2 t + \beta_1(Z_{t-1}/n_{t-1}) + \beta_2(Z_{t-2}/n_{t-2}) + \cdots \\ &\quad +\beta_q(Z_{t-q}/n_{t-q}) + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12). \end{aligned} \tag{4.7}$$

This is a Markov regression model and generalizes the model Cox (1981)

refers to as an 'observation-driven linear logistic autoregression', in that it incorporates trend and seasonality and is based on a binomial rather than a Bernoulli distribution. Clearly it is possible to add further terms to the above expression for logit p_t to allow for the effect of any further covariates thought relevant, e.g. the number or proportion of deliveries by various instrumental techniques.

It is easy to compute estimates of the parameters α_1, α_2, etc. of the model (4.7) by using a program such as GLIM or GENSTAT. If for instance no observations earlier than Z_{t-1} appear in the model, the product

$$\prod_{t=97}^{192} \binom{n_t}{z_t} p_t^{z_t}(1-p_t)^{n_t-z_t}$$

is the likelihood of $\{Z_t : t = 97, \ldots, 192\}$, conditional on Z_{96}. Maximization thereof with respect to $\alpha_1, \alpha_2, \beta_1, \gamma_1$ and γ_2 yields estimates of these parameters, and can be accomplished by performing a logistic regression of Z_t on t, Z_{t-1}/n_{t-1}, $\sin(2\pi t/12)$ and $\cos(2\pi t/12)$.

In the search for a suitable model the following explanatory variables were in fact considered: t (i.e. the time in months, with February 1970 as month 1), t^2, the proportion Caesarean lagged one, two, three or twelve months, sinusoidal terms at the annual frequency (as in (4.7)), the calendar month, the proportion and number of deliveries by forceps or vacuum extraction, and the proportion and number of breech births. Models were compared on the basis of AIC and BIC. These criteria are defined as $-2l + 2k$ and $-2l + k \log n$ respectively, where l is the maximized log likelihood of the model under discussion, k the number of parameters and n the number of observations. GLIM, the program used, does not provide l, but it does provide the (scaled) deviance, which is closely related. It is in fact twice the difference between the maximum log likelihood achievable in a full model

and that achieved in the model under investigation (McCullagh and Nelder, 1989, p. 33). Here the maximum log likelihood of a full model is -321.04, from which it follows that $-l = 321.04 + \frac{1}{2} \times$ deviance, and AIC and BIC are therefore easily computed. The best models found with between one and four explanatory variables are listed in Table 4.6, as well as the model with constant term only and the model with Z_{t-1}/n_{t-1} as the only explanatory variable. BIC indicated that very little would be gained by inclusion of a fifth explanatory variable in the model.

Table 4.6: Births data. Models fitted by GLIM to the the logit of the proportion Caesarean.

Explanatory variables	Deviance	$-2l$	AIC	BIC	Coefficients (constant first)
–	324.05	966.13	968.13	970.69	−1.073
z_{t-1}/n_{t-1}	224.89	866.97	870.97	876.10	−1.813, 2.899
t	208.92	851.00	855.00	860.13	−1.583, .003439
$t, z_{t-1}/n_{t-1}$	191.70	833.78	839.78	847.47	−1.822, .002372, 1.554
$t, z_{t-1}/n_{t-1}$, no. forceps deliveries in month t	183.63	825.71	833.71	843.97	−1.592, .001536, 1.409, −.002208
$t, z_{t-1}/n_{t-1}$, no. forceps deliveries in month t, October indicator	175.60	817.68	827.68	840.50	−1.583, .001422, 1.431, −.002393, .08962

The strongest conclusion we may draw from these models is that there is indeed a marked upward time trend in the proportion Caesarean. Secondly, there is positive dependence on the proportion Caesarean in the previous month. The negative association with the number (or proportion) of forceps deliveries is not surprising in view of the fact that delivery by Caesarean and by forceps are in some circumstances alternative techniques. As regards seasonality, the only possible seasonal pattern found in the proportion Caesarean is the positive 'October effect'. Among the calendar months only October stood out as having some explanatory power. As can be seen from Table 4.6, the indicator variable specifying whether the month was October was included in the 'best' set of four explanatory variables.

Since the main conclusion emerging from the above logistic-linear models is that there is a marked upward time trend in the proportion Caesarean, it is of interest also to fit hidden Markov models with and without time trend. The hidden Markov models we use in this application have two states and are defined as follows. Suppose $\{C_t\}$ is a stationary homogeneous Markov chain on state space $\{1,2\}$, with transition probability matrix Γ. Suppose also that, conditional on the Markov chain, Z_t has a binomial distribution with parameters n_t and p_i, where $C_t = i$. A model without time trend assumes that p_1 and p_2 are constants, and a possible model which allows p_i to depend on t has logit ${}_tp_i = \alpha_i + \beta t$.

Maximization of the likelihood of the last eight years' observations, subject to the bounds of 0 and 1 on the probabilities involved, was performed by the Nelder-Mead simplex method (Press et al., 1986). The model without time trend yielded $-l = 420.48$, where as before l denotes the maximized log likelihood. The details of the model with time trend are as follows, with t

denoting the time in months and February 1970 being month 1:

$$\Gamma = \begin{pmatrix} .838 & .162 \\ .262 & .738 \end{pmatrix}, \text{ logit } {}_tp_1 = -1.634 + .003297t, \qquad (4.8)$$

$$\text{logit } {}_tp_2 = -1.456 + .003297t, \ -l = 402.35.$$

This model can be described as consisting of a Markov chain with two fairly persistent states, along with their associated time-dependent probabilities of delivery being by Caesarean, the (upward) time trend being the same for the two states. State 1 is rather more likely than state 2 because the stationary distribution is (.618 .382), and has a lower probability of delivery being by Caesarean. It may or may not be possible to interpret the states as (for instance) busy and non-busy periods in the Obstetrics Unit of the hospital, but without further information, e.g. on staffing levels, such an interpretation would be speculative.

If one wishes to use model (4.8) to forecast the proportion Caesarean at time 193 for a given number of deliveries, what is needed is the one-step-ahead forecast distribution of Z_{193}, i.e. the distribution of Z_{193} conditional on $Z_{97}, \ldots, Z_{192}$. This is given by the likelihood of $Z_{97}, \ldots, Z_{193}$ divided by that of $Z_{97}, \ldots, Z_{192}$. More generally, the the k-step-ahead forecast distribution, i.e. the conditional probability that $Z_{192+k} = z$, is given by a ratio of likelihoods, as described in section 2.6.

The difference in likelihood between the hidden Markov models with and without time trend is convincing evidence of an upward trend in the proportion Caesarean, and confirms the main conclusion drawn above from the logistic-linear models. Haines et al. concluded that there is an upward trend in the number of Caesareans: since the last eight years of the total deliveries series apparently has an upward trend (see Fig. 1 of Haines et al., or sub-

section 4.3.2 of this thesis), our conclusion is consistent with theirs. It does not seem possible, however, to draw any conclusion about the proportion Caesarean from their ARIMA models alone.

It is of interest also to compare the fit of the hidden Markov model (4.8) to the data with that of the logistic autoregressive models. Here it should be noted that the hidden Markov model produces lower values of AIC (814.70) and BIC (827.52) than does the logistic autoregressive model with four explanatory variables (827.68 and 840.50 respectively), *and* makes use of less information. It does not use z_{96}, nor does it use information on forceps deliveries or the calendar month. It seems therefore that hidden Markov models have considerable potential as simple yet flexible models for examining dependence on covariates (such as time) in the presence of autocorrelation.

4.3.2 Models for the total number of deliveries

If one wishes to project the number of Caseareans, however, one needs in addition to the model for the proportion Caesarean a model for the total number of deliveries, which is a series of unbounded counts. The model of Haines et al. for the total deliveries was the seasonal ARIMA model $(0,1,1)\times(0,1,1)_{12}$ without constant term, and for this series (unlike the others) they used all sixteen years' data to fit the model.

In this work an attempt was first made to model the monthly total of deliveries by means of two-state Poisson-HM models. Two such models were considered: one with a single linear trend in the log of the conditional mean, i.e. a model with

$$\log {}_t\lambda_i = a_i + bt \; ; \tag{4.9}$$

and one incorporating in addition sinusoidal terms at the annual frequency:

$$\log {}_t\lambda_i = a_i + bt + c\cos(2\pi t/12) + d\sin(2\pi t/12) \ . \qquad (4.10)$$

Both models fitted are unsuccessful in the sense that they effectively degenerate to one-state models: in each case one of the two off-diagonal transition probabilities is so close to zero as to be negligible, and the stationary distribution assigns probability one to one of the two states.

Various logistic-linear models were then fitted by GLIM, and it is notable that GLIM yielded (inter alia) precisely the two models described above, with parameter estimates agreeing to four significant figures, but required fewer parameters to do so. (For instance, the general two-state hidden Markov model with conditional mean given by equation (4.10) has a total of seven parameters, and the corresponding GLIM model has four.) Table 4.7 compares the models fitted by GLIM to the total deliveries, and from that table it can be seen that BIC selects the model incorporating time trend, sinusoidal components at the annual frequency and the number of deliveries in the previous month. The details of this model are as follows. Conditional on the history, the number of deliveries in month t (N_t) is distributed Poisson with mean ${}_t\lambda$, where

$$\log {}_t\lambda = 5.781 + .002436t - .03652\cos(2\pi t/12) - .02164\sin(2\pi t/12) + .0005737 n_{t-1} \ .$$

(Here, as before, February 1970 is month 1 but the model was fitted to the numbers of deliveries in months 97–192 only.)

Although both the hidden Markov models and the logistic-linear autoregressions revealed time trend and seasonality in this case, only the latter were able to detect dependence on the previous observation. This suggests that the dependence which is present in the total deliveries series is simply of

Table 4.7: Births data. Models fitted by GLIM to the log of the mean no. of deliveries.

Explanatory variables	Deviance	$-2l$	AIC	BIC
t *	545.98	1348.9	1352.9	1358.1
t, sinusoidal terms **	428.70	1231.7	1239.7	1249.9
t, sinusoidal terms, n_{t-1}	356.10	1159.1	1169.1	1181.9
t, sinusoidal terms, n_{t-1}, n_{t-2}	353.91	1156.9	1168.9	1184.3
sinusoidal terms, n_{t-1}	464.83	1267.8	1275.8	1286.1
t, n_{t-1}	410.69	1213.6	1219.6	1227.3
n_{t-1}	499.47	1302.4	1306.4	1311.6

* This is the model identical to the hidden Markov model (4.9).

** This is identical to the model (4.10).

a kind that can be detected by appropriate observation-driven models, but not by parameter-driven models.

4.3.3 Conclusion

The conclusion is therefore twofold. If a model for the number of Caesareans, given the total number of deliveries, is needed, the binomial-hidden Markov model with time trend is best of all of those considered (including various logistic-linear autoregressive models). If on the other hand a model for the total deliveries is needed (e.g. as a building-block in projecting the number

of Caesareans) a first-order logistic-linear autoregressive model incorporating also time trend and sinusoidal components seems best, and is certainly superior to the Poisson-hidden Markov models fitted.

4.4 An application to animal behaviour: locomotory behaviour of *Locusta migratoria*

Discrete-time Markov chain models, of first order or higher, are quite commonly used in the study of behaviour sequences of species as diverse as blowflies (Cane, 1978), beavers (Rugg and Buech, 1990), and Rhesus monkeys (Cane, 1978). The transition probabilities are usually estimated by conditional maximum likelihood, i.e. by equating them to the relevant relative frequencies. The two main purposes of such modelling are firstly to provide a fairly simple summary of the behaviour sequence observed, and secondly to provide a basis for comparisons between subjects or between the behaviours of a single subject under different conditions. It seems usually to be assumed in such applications that the transition probabilities are stationary, i.e. that a homogeneous Markov chain is appropriate as a model. Clearly this assumption would not be justified if there were a trend in environmental conditions or in the motivational state of the animal under observation. A further limitation of the usual approach is that it is not easily applicable to joint modelling of several subjects with possibly differing transition probabilities. (If it may reasonably be assumed that several independent subjects possess the same transition probabilities, these probabilities can then be estimated by pooling the transition counts across subjects.) Expansion of the state-space of the Markov chain to accommodate several subjects greatly increases its dimension and a fortiori the number of transition probabilities being estimated, and it is doubtful whether such a complex model would be

useful in the role of summarizing behaviour. Hidden Markov models, on the other hand, can cope both with time trend and with expansion to several subjects without an explosion in the number of parameters, and furthermore they provide an alternative to Markov chains of order greater than one if it is thought that the Markov property (of order one) is too restrictive an assumption and a longer memory is needed for a model to be realistic.

In this section we describe multivariate and univariate hidden Markov models fitted to the simultaneous behaviour of 24 locusts (*Locusta migratoria*) in an experiment carried out by Dr D. Raubenheimer of the Zoology department of the University of Cape Town. The experiment studied the effect of hunger on locomotory behaviour, but the purpose was in fact to calibrate the experimental methods used by assessing their sensitivity to a known effect. The subjects were all three days into the fifth stadium, and the experiment was conducted as follows. The odd-numbered subjects were allowed to feed ad libitum for $5\frac{1}{2}$ hours before observation commenced, and the even-numbered subjects were deprived of food for the same period. Food and water were not available during observation. Within each of the two groups the subjects were alternately male and female. During the observation period a beep sounded at 30-second intervals. The 24 subjects were observed sequentially as soon as possible thereafter, and their behaviour classified by the observer as locomoting, quiescent or (rarely) grooming. The number of observations made on each subject was 161. For the purpose of this analysis the categories used were locomoting (denoted by 1) and not locomoting (0). For the data, see Appendix B.

As a preliminary, one-step transition counts were found for each of the 24 subjects, and homogeneous Markov chain models fitted thereby. Although these models do indeed point to differences between the fed and the starved

subjects (all but one of the fed subjects having a lower unconditional probability of locomoting than all the starved subjects), there are two facts which suggest that a homogeneous Markov chain may not be an appropriate model. Firstly, in most of the 24 cases the sample ACF of the observed binary sequence is very far from being of the form α^k. Secondly, the level of activity across all subjects increases with the passage of time, which suggests that there may be a time trend in the transition probabilities. At the very least therefore we can conclude that a *stationary* homogeneous Markov chain is not appropriate.

It was therefore decided to fit a two-state multivariate hidden Markov model with time trend to each of the two groups of twelve subjects, and to compare these models with similar ones lacking a time trend, in order to assess whether inclusion of such a trend appreciably improves the fit. The model without trend has 26 parameters. For the fed subjects, the multivariate model without time trend has transition probability matrix $\begin{pmatrix} .995 & .005 \\ .009 & .991 \end{pmatrix}$, corresponding stationary distribution (.627 .373), and log likelihood (l) equal to -371.81. The 24 'state-dependent' probabilities are the probabilities of locomotion for each of the twelve subjects in each of the two states. The model fitted has the property that, for nine of the twelve subjects, state 1 has associated with it a lower probability of locomotion than has state 2 — but of course it must be remembered that the numbering of states is arbitrary.

The more general model, which allows for a single time trend (i.e. the same trend for all subjects in the two states), assumes that the probability of locomotion of subject j in state i at time t is ${}_tp_{ij}$, where for $i = 1, 2$ and

$j = 1, 2, \ldots, 12$

$$\text{logit } {}_tp_{ij} = a_{ij} + bt.$$

Such a model has 27 parameters, and the model fitted has a log likelihood of -365.40, whence by AIC and BIC it is preferable to the model without time trend. (In passing we note that a model allowing b to vary between states was also fitted, but because it resulted in a very small improvement in the log likelihood, to -365.20, it was not considered further.) The 27-parameter model fitted is as follows. The underlying Markov chain has transition probability matrix $\begin{pmatrix} .994 & .006 \\ .011 & .989 \end{pmatrix}$ and stationary distribution (.654 .346). The time trend parameter b is .01405, and the parameters a_{ij} are given for $i = 1, 2$ and $j = 1, 2, \ldots, 12$ in Table 4.8. There is no clear pattern of these parameters being greater for one state than another: for seven of the twelve subjects a_{1j} exceeds a_{2j}.

The same sequence of models was fitted to the behaviour of the twelve starved subjects. The 26-parameter model had $l = -1116.1$, the 27-parameter $l = -1104.5$, and the 28-parameter $l = -1103.6$. As in the case of the fed subjects, therefore, we concentrate on the 27-parameter model. The Markov chain of that model has transition probability matrix $\begin{pmatrix} .966 & .034 \\ .064 & .936 \end{pmatrix}$ and stationary distribution (.653 .347). The time trend parameter b is .01054, and the parameters a_{ij} for this model are also given in Table 4.8. It will be noted that, for all but one of the starved subjects, a_{1j} exceeds a_{2j} and ${}_tp_{1j}$ therefore exceeds ${}_tp_{2j}$ for all t.

We now compare the two sets of subjects on the basis of these two 27-parameter models. Because the same time trend parameter applies whether the state is 1 or 2, it is meaningful to compare time trends between the two

Table 4.8: Locust data, fed subjects (left) and starved subjects (right). Parameters a_{ij} of multivariate hidden Markov models with single time trend in each case.

j	a_{1j}	a_{2j}	j	a_{1j}	a_{2j}
1	−5.841	−23.47	1	−.002285	−2.292
2	−3.996	−1.139	2	−0.9990	−1.656
3	−4.210	−4.797	3	1.803	−1.386
4	−4.193	−4.402	4	0.9171	−0.2047
5	−5.841	−4.086	5	−0.5881	0.09890
6	−5.841	−5.566	6	−2.092	−21.37
7	−3.066	−3.098	7	−.3111	−1.001
8	−4.722	−23.49	8	−1.332	−1.748
9	−25.41	−3.425	9	−1.229	−2.268
10	−3.559	−1.641	10	0.1437	−0.3940
11	−2.904	−2.966	11	−1.325	−2.325
12	−5.841	−23.48	12	−1.471	−1.324

sets of subjects even though the underlying Markov chains differ. (As it happens, the stationary distributions of the two Markov chains are practically identical, but that is not necessary for a comparison to be meaningful.) Although the time trend is positive in both cases (i.e. the probability of locomotion increases with the passage of time), it is greater (on a logistic scale) for the fed subjects than for the starved ones. This is plausible: the subjects fed beforehand are experiencing a greater change in their condition during the observation period than are the subjects starved beforehand. Furthermore one can note that there appears to be among the starved subjects a greater consistency of behaviour than is the case for the fed ones: the evidence for this statement is the property noted above that, for all but one of the starved subjects, ${}_tp_{1j}$ exceeds ${}_tp_{2j}$. For the fed subjects the situation is less clear-cut, in that a_{1j} exceeds a_{2j} for seven of the twelve subjects. A

convincing explanation for this difference between the fed and the starved subjects is that the fed subjects have entered observation at varying stages of the 'cycle of satiety': variation of this kind is unlikely to apply in the case of the starved subjects.

Although in the case of the data considered here there is separate information on each of the subjects, there are experimental situations in which it is known only, for each time, *how many* of the subjects have a given characteristic: that is, only 'macro data' are available. It is therefore interesting to fit univariate models to the number locomoting in each group, to make comparisons between these univariate models for the two groups, and to compare the conclusions with those drawn from the multivariate models discussed above.

We therefore consider, for each group of twelve subjects, univariate models which pool the movements of the subjects. First a simple two-state binomial-hidden Markov model was fitted to the series of length 161 giving for each time point the number of subjects (out of twelve) observed to be locomoting. In the notation of Chapter 2, these models have n_t equal to 12 for all t, and they can of course be generalized by incorporating time trends in the state-dependent probabilities. For the fed and starved subjects respectively, the models without time trend achieved log likelihood values of -169.80 and -314.07. The corresponding models with a single time trend applying to both states achieved values of -163.93 and -304.49. Models allowing for differing time trends in the two states were also fitted, but produced only very minor improvements in the log likelihood (to -163.90 and -303.66) and were not considered further. The models with a single time trend are clearly superior to those without time trend, on the basis of AIC or BIC: see Table 4.9.

Table 4.9: Locust data. Comparison of univariate models pooling movements within groups.

model	no. parameters	l	AIC	BIC
fed subjects, no time trend	4	−169.80	347.6	359.9
fed subjects, one trend	5	−163.93	337.9	353.3
starved subjects, no trend	4	−314.07	636.1	648.5
starved subjects, one trend	5	−304.49	619.0	634.4

In detail, the models with a single time trend are as follows. For the fed subjects, the underlying Markov chain has transition probability matrix $\begin{pmatrix} .963 & .037 \\ .020 & .980 \end{pmatrix}$, stationary distribution (.355 .645), and probabilities ${}_tp_i$ of locomotion at time t in state i specified by

$$\text{logit } {}_tp_1 = -4.745 + .01788t$$

and

$$\text{logit } {}_tp_2 = -4.050 + .01788t \ .$$

For the starved subjects, the t.p.m. is $\begin{pmatrix} .986 & .014 \\ .013 & .987 \end{pmatrix}$, the stationary distribution is (.478 .522), and

$$\text{logit } {}_tp_1 = -1.280 + .008747t$$

and

$$\text{logit } {}_tp_2 = -.5103 + .008747t \ .$$

Again the time trend is positive for both fed and starved subjects, and greater (on a logistic scale) for the fed subjects. The initial probability of locomotion is very much smaller for the fed subjects (.009 or .017, depending on the state) than it is for the starved ones (.218 or .375). The corresponding

probabilities for time 161 are .134 and .237 (fed) and .532 and .711 (starved).

We may conclude therefore that both types of model (i.e. the multivariate and the univariate models) can provide a relatively simple and meaningful summary, and are useful as bases for comparison between groups. The great advantage that the hidden Markov models have over straightforward Markov chain models is that they can in fairly simple fashion accommodate several subjects and dependence on time (or any other covariate, for that matter).

Whether the states of the Markov chain (in either the multivariate binary model or the univariate binomial-HM model) will support a substantive interpretation is a question perhaps best left to the entomologist. In the case of the multivariate model it seems, however, that it may be difficult to provide an explanation in terms of the motivational states of the subjects. An explanation in terms of common environmental conditions which form the basis for simultaneous behaviour seems more feasible. For a single subject either kind of explanation may be feasible. For the purposes of summary and comparison of behaviour, however, the hidden Markov models do appear to be useful even without such substantive interpretation.

4.5 Wind direction at Koeberg

South Africa's only nuclear power station is situated at Koeberg on the west coast, about 30 km north of Cape Town. Wind direction, wind speed, rainfall and other meteorological data are collected continuously by the Koeberg weather station with a view to their use in radioactive plume modelling, inter alia. Four years of wind direction data were made available by Mr G. Fick and Mr F. Potgieter of the Koeberg weather station, and this section describes an attempt to model the wind direction at Koeberg by means of

hidden Markov models for categorical time series.

The data consist of hourly values of the average wind direction over the preceding hour at 35 m above ground level. The period covered is 1 May 1985 to 30 April 1989 inclusive. The average referred to is a vector average, which allows for the circular nature of the data, and is given in degrees. There are in all 35064 observations: there are no missing values. Before any models were fitted the hourly averages were classified into the 16 conventional directions N, NNE, ..., NNW, coded 1 to 16 in that order. For the data, see Appendix B.

The first model fitted was a simple multinomial-hidden Markov model with two states and no seasonal components, the case $m = 2$ and $q = 16$ of the categorical model described in subsection 3.3.3. In this model there are 32 parameters to be estimated: two transition probabilities to specify the Markov chain, and fifteen probabilities for each of the two states, subject to the sum of the fifteen not exceeding one. As usual, parameter estimation was performed by maximizing the likelihood with the help of the NAG subroutine E04UCF. The results are as follows. The underlying Markov chain has transition probability matrix $\begin{pmatrix} .964 & .036 \\ .031 & .969 \end{pmatrix}$ and stationary distribution (.462 .538) , and the sixteen probabilities associated with each of the two states are displayed in Table 4.10. A graph of these two sets of probabilities appears as Figure 4.1. The value of the unconditional log likelihood achieved by this model is -75832.1.

The model successfully identifies two apparently meaningful 'climate states' which are very different at least as regards the likely wind directions in those states. In state 1 the most likely direction is NNW, and the

Table 4.10: Koeberg wind data. Probabilities of each direction in the simple two-state hidden Markov model.

1	N	.129	.000
2	NNE	.048	.000
3	NE	.059	.001
4	ENE	.044	.026
5	E	.006	.050
6	ESE	.001	.075
7	SE	.000	.177
8	SSE	.000	.313
9	S	.001	.181
10	SSW	.004	.122
11	SW	.034	.050
12	WSW	.110	.008
13	W	.147	.000
14	WNW	.130	.000
15	NW	.137	.000
16	NNW	.149	.000

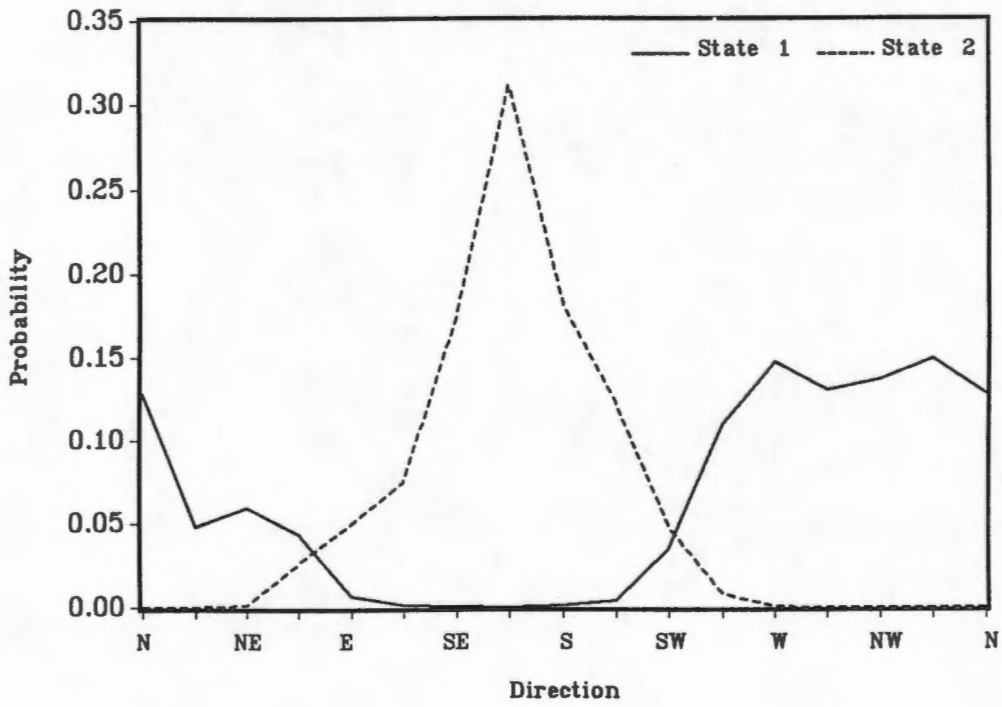

Figure 4.1: Koeberg wind data. Probabilities of each direction in the simple two-state hidden Markov model.

probability falls away on either side of NNW, reaching a level of less than .01 for all directions (clockwise) from E to SSW inclusive. In state 2 the peak is at direction SSE, and the probability falls away sharply on either side, being less than .01 for all directions from WSW to NE inclusive.

Two generalizations of this type of model were also fitted: firstly a model based on a three-state Markov chain rather than a two-state one, and secondly a model based on a two-state Markov chain but incorporating both a daily cycle and an annual cycle. We shall now describe these two models, and in due course compare all the models considered on the basis of AIC and BIC.

The three-state hidden Markov model has 51 parameters: six to specify the Markov chain, and fifteen probabilities associated with each of the states. Essentially this model splits state two of the two-state model into two new states, one of which peaks at SSE and the other at SE. The transition probability matrix is

$$\begin{pmatrix} .957 & .030 & .013 \\ .015 & .923 & .062 \\ .051 & .077 & .872 \end{pmatrix}$$

and the stationary distribution is (.400 .377 .223). The sixteen probabilities associated with each of the three states are displayed in Table 4.11, and the unconditional log likelihood is -69525.9.

As regards the model which adds daily and annual cyclic effects to the simple two-state hidden Markov model, it was decided to build these effects into the Markov chain rather than into the state-dependent probabilities: this was because for a two-state chain we can model cyclic effects parsimoniously by assuming that the two off-diagonal transition probabilities

$$\mathrm{P}(C_t \neq i \mid C_{t-1} = i) = {}_t\gamma_i$$

Table 4.11: Koeberg wind data. Probabilities of each direction in the three-state hidden Markov model.

1	N	.148	.000	.001
2	NNE	.047	.000	.016
3	NE	.016	.000	.097
4	ENE	.003	.000	.148
5	E	.001	.000	.132
6	ESE	.000	.000	.182
7	SE	.000	.023	.388
8	SSE	.000	.426	.033
9	S	.000	.257	.002
10	SSW	.002	.176	.000
11	SW	.002	.089	.000
12	WSW	.111	.028	.000
13	W	.169	.002	.000
14	WNW	.151	.000	.000
15	NW	.159	.000	.001
16	NNW	.173	.000	.000

are given by

$$\text{logit}\,{}_t\gamma_i = a_i + b_i \cos(2\pi t/24) + c_i \sin(2\pi t/24) + d_i \cos(2\pi t/8766) + e_i \sin(2\pi t/8766) \tag{4.11}$$

for $i = 1, 2$ and $t = 2, 3, \ldots, T$. A similar model for each of the state-dependent probabilities in each of the two states would involve many more parameters, or a problem of asymmetry which does not arise if the above method is used, or both. As discussed in section 3.7, the estimation technique has to be modified when the underlying Markov chain is not assumed to be homogeneous. The estimates in this case were based on the initial state of the Markov chain being state 2, and the log of the likelihood conditioned on that initial state is -75658.5. (Conditioning on state 1 yielded a slightly inferior value for the likelihood, and similar parameter estimates.) The models for the two off-diagonal transition probabilities are given in Table 4.12, in the notation of equation (4.11), and the probabilities associated with each state are given in Table 4.13. From this last table it will be noted that the general pattern of the state-dependent probabilities is very similar to the pattern seen in the simple two-state model without any cyclical components.

Table 4.12: Koeberg wind data. Models, incorporating cyclical components, for the off-diagonal transition probabilities of the hidden Markov model.

i	a_i	b_i	c_i	d_i	e_i
1	−3.349	.1975	−.6946	−.2078	−.4008
2	−3.523	−.2723	.8007	.08209	−.08858

The three models described above were compared with each other and with a saturated sixteen-state Markov chain model, on the basis of AIC and BIC. The transition probabilities defining the Markov chain model were estimated by conditional maximum likelihood, as described in section 1.2, and

Table 4.13: Koeberg wind data. Probabilities of each direction in the two-state hidden Markov model with cyclical components.

1	N	.127	.000
2	NNE	.047	.000
3	NE	.057	.002
4	ENE	.027	.040
5	E	.004	.052
6	ESE	.001	.076
7	SE	.001	.179
8	SSE	.000	.317
9	S	.001	.183
10	SSW	.007	.121
11	SW	.059	.026
12	WSW	.114	.003
13	W	.145	.000
14	WNW	.128	.000
15	NW	.135	.000
16	NNW	.147	.000

are displayed in Table 4.14. The comparison appears as Table 4.15.

What is of course striking is that the saturated Markov chain model is so much better for this data-set than the hidden Markov models that the large number of parameters of the Markov chain (240) is virtually irrelevant when comparisons are made by AIC or BIC. It is therefore interesting to compare certain properties of the Markov chain model with the corresponding properties of the simple two-state hidden Markov model (e.g. unconditional and conditional probabilities). The unconditional probabilities of each direction were computed for the two models, and are virtually identical. However, examination of the conditional probability $\mathrm{P}(S_t = 16 \mid S_{t-1} = 8)$, where S_t denotes the direction at time t, suggests a difference between the models. For the Markov chain model this probability is zero, since no such transitions were observed. For the hidden Markov model it is, in the notation of section 3.3.3,

$$\frac{\delta P_{(8)} \Gamma P_{(16)} \mathbf{1}'}{\delta P_{(8)} \mathbf{1}'} = \frac{.5375 \times .3131 \times .0310 \times .1494}{.5375 \times .3131} = .0046.$$

Although small, this probability is not insignificant: in 5899 transitions from state 8 (the actual number observed, out of a total of 35063 transitions) one would, on the basis of the hidden Markov model, expect 27 transitions to be to state 16. There were none observed. The explanation is that the hidden Markov model makes 180-degree switches in direction quite possible: every time the state changes (which happens at any given time point with probability in excess of .03) the most likely wind direction changes by 180 degrees. This is inconsistent with the observed gradual changes in direction: the matrix of observed transition counts is heavily dominated by diagonal and near-diagonal elements (and, because of the circular nature of the categories, elements in the corners most distant from the principal diagonal). Changes through 180 degrees were in general rarely observed.

Table 4.14: Koeberg wind data. Transition probability matrix of saturated Markov chain model.

row 1	. 610	.080	.022	.008	.003	.001	.001	.002	
		.004	.000	.003	.003	.014	.020	.037	.190
row 2	.241	.346	.163	.037	.015	.001	.003	.003	
		.003	.003	.009	.010	.029	.037	.036	.064
row 3	.056	.164	.468	.134	.028	.006	.009	.006	
		.004	.006	.006	.018	.032	.018	.019	.025
row 4	.013	.033	.163	.493	.144	.032	.017	.011	
		.011	.017	.007	.012	.014	.014	.010	.010
row 5	.009	.007	.048	.249	.363	.138	.053	.027	
		.033	.022	.011	.006	.008	.007	.013	.009
row 6	.004	.009	.010	.051	.191	.423	.178	.051	
		.025	.023	.008	.007	.004	.006	.004	.004
row 7	.001	.001	.003	.006	.023	.141	.607	.160	
		.036	.008	.006	.003	.002	.001	.001	.001
row 8	.001	.001	.001	.003	.005	.016	.140	.717	
		.094	.013	.005	.003	.002	.001	.000	.000
row 9	.004	.001	.002	.004	.005	.008	.025	.257	
		.579	.077	.017	.009	.005	.003	.001	.000
row 10	.002	.002	.002	.001	.006	.005	.010	.036	
		.239	.548	.093	.041	.010	.005	.002	.000
row 11	.005	.002	.003	.005	.003	.003	.008	.012	
		.038	.309	.397	.151	.038	.012	.008	.005
row 12	.004	.002	.002	.003	.001	.003	.002	.005	
		.017	.056	.211	.504	.149	.019	.016	.007
row 13	.010	.005	.005	.003	.004	.001	.002	.005	
		.004	.013	.028	.178	.561	.138	.030	.013
row 14	.013	.005	.004	.003	.003	.003	.001	.001	
		.001	.007	.008	.027	.188	.494	.199	.043
row 15	.031	.009	.007	.004	.005	.002	.001	.001	
		.002	.001	.003	.011	.043	.181	.509	.190
row 16	.158	.023	.009	.005	.001	.001	.002	.001	
		.000	.002	.002	.004	.017	.054	.162	.559

Table 4.15: Koeberg wind data. Comparison of four models.

model	no. parameters	log likelihood	AIC	BIC
2-state HM	32	−75832.1	151728	151999
3-state HM	51	−69525.9	139154	139585
2-state HM with cycles saturated	40	−75658.5*	151397	151736
Markov chain	240	−48301.7	97083	99115

* conditional on state two being the initial state

The above discussion suggests that, if daily figures are examined rather than hourly, the observed process will be more amenable to modelling by means of a hidden Markov model, because abrupt changes of direction will be more likely in the daily data. Markov chain and two-state hidden Markov models were therefore fitted to the series of length 1461 beginning with the first observation and including every 24th observation thereafter. For these data the hidden Markov model proved to be superior to the Markov chain even on the basis of AIC, which penalizes extra parameters less here than does BIC. (See Table 4.16.) Although a daily model is of little use in the context of the main application intended (radioactive plume modelling) because of the necessarily short time scale involved in such modelling, there are other applications for which a daily model is exactly what one needs. Forecasting of the wind direction a day ahead is the obvious example.

Since the (first-order) Markov chain model does not allow for dependence beyond first-order, the question that arises is whether *any* model for the hourly data which allows for higher-order dependence is superior to the Markov chain model: the hidden Markov models considered clearly are not.

Table 4.16: Koeberg wind data (daily). Comparison of two models fitted.

model	no. parameters	log likelihood	AIC	BIC
2-state HM	32	−3461.88	6987.75	7156.93
saturated MC	240	−3292.52	7065.04	8333.89

A saturated second-order Markov chain model would have a ridiculous number of parameters, and the Pegram model for (e.g.) a second-order Markov chain could not reflect the property that, if a transition is made out of a given state, its near neighbours are more likely destinations than are more distant states. The Raftery models do not suffer from that disadvantage, and in fact it is very easy to find a lag-2 Raftery model which is convincingly superior to the first-order Markov chain model: in the notation of section 1.3, take Q to be the transition probability matrix of the first-order Markov chain, and perform a simple line-search to find that value of λ_1 which maximizes the resulting conditional likelihood. This turns out to be .925. With these values for Q and λ_1 as starting values, the conditional likelihood was then maximized with respect to the 241 parameters, subject to the assumption that $0 \leq \lambda_1 \leq 1$. (This assumption avoids the necessity of imposing 16^3 non-linear constraints on the maximization, and seems plausible in the context of hourly wind directions showing a high degree of persistence.) The resulting value for λ_1 is .9125, and the resulting matrix Q does not differ much from its starting value, the transition probability matrix displayed in Table 4.14. The log likelihood achieved is −48049.8: see Table 4.17 for a comparison of likelihood values and the usual model selection criteria, from which the Raftery model emerges as superior to the Markov chain.

The general conclusion we may draw from the above analysis is that, both for the hourly and the daily wind direction data, it is possible to fit a

Table 4.17: Koeberg wind data (hourly). Comparison of first-order Markov chain with Raftery models.

model	no. parameters	log likelihood	AIC	BIC
Saturated MC	240	−48301.7	97083.4	99115.0
Raftery model with starting values	241	−48087.8*	96657.5	98697.6
Raftery model fitted by max. likelihood	241	−48049.8*	96581.6	98621.7

* conditioned on the first two states

model which is superior (in terms of the model selection criteria used) to a saturated first-order Markov chain. In the case of the hourly data, a lag-2 Raftery model (i.e. a particular kind of second-order Markov chain model) is preferred. In the case of the daily data, a simple two-state hidden Markov model for categorical time series, as introduced in subsection 3.3.3, performs better than the Markov chain. One further approach to the hourly data that might well turn out to be superior to any of the models considered above is to develop a parsimonious class of models for Markov chains with transition probability matrix of the form seen in Table 4.14. This could be extended to higher-order Markov chains in exactly the same way as Raftery's models generalize the saturated (first-order) Markov chain.

4.6 Evapotranspiration

Kedem (1976) has proposed a second-order Markov chain model for a binary time series derived from a set of evapotranspiration data. The data are $\{Z_t : t = 1, 2, \ldots, 96\}$, rates of evapotranspiration in 96 consecutive months at a site in Israel, and are given in full in Kedem's paper. Kedem defines

$W_t = Z_t - Z_{t-12}$ and

$$X_{t-12} = \begin{cases} 1 & \text{if } W_t \geq \overline{W} \\ 0 & \text{if } W_t < \overline{W}. \end{cases}$$

The resulting series $\{X_t\}$ has 84 values, and is given by:

11111 11100 01111 10000 00011 11000 10000 01111
11110 01100 00000 11111 10000 00000 11111 11000 0000 .

Kedem performs a likelihood-ratio test of the hypothesis of first order against second order, and concludes that a second-order Markov chain is an appropriate model. He does however note that the significance level used is high (.2). In order to discover whether a better model than a second-order Markov chain could be found, hidden Markov models (with two, three and four states) were therefore fitted to the series. All the models (first- and second-order Markov chains and the three hidden Markov models) were then compared by BIC, which is a consistent estimator of Markov chain order (Katz, 1981). The results are as in Table 4.18, with l denoting the log of the (unconditional) likelihood. The AIC values are included for completeness.

Table 4.18: Evapotranspiration data. Comparison of models.

model	no. parameters	l	AIC	BIC
Markov chain	2	−39.935	83.87	88.73
2nd order MC	4	−38.203	84.41	94.13
2-state HM	4	−39.905	87.81	97.53
3-state HM	9	−37.845	93.69	115.6
4-state HM	16	−35.187	102.4	141.3

Since the Markov chain and second-order Markov chain were estimated by conditional maximum likelihood, and the other models by unconditional,

this comparison on the basis of the unconditional likelihood is slightly unfair to the first two models. What is interesting, however, is that the simplest model of them all, the first-order Markov chain, emerges as the 'best' — contrary to Kedem's conclusion that a second-order Markov chain is needed. Furthermore, the two-state binomial-hidden Markov model fitted is simply the Markov chain with transition probability matrix $\begin{pmatrix} .823 & .177 \\ .185 & .815 \end{pmatrix}$. (Although this is not identical to the Markov chain model yielding the value $l = -39.935$, i.e. the Markov chain with t.p.m. $\begin{pmatrix} \frac{35}{42} & \frac{7}{42} \\ \frac{8}{41} & \frac{33}{41} \end{pmatrix}$, the difference is explained by the use of conditional maximum likelihood in one case and unconditional in the other.) The fact that the best two-state hidden Markov model that can be found is just a Markov chain strengthens the conclusion that nothing more complicated than a Markov chain is justified, although the use of unconditional maximum likelihood seems preferable because the assumption of stationarity is reasonable for these data.

4.7 Thinly traded shares on the Johannesburg Stock Exchange

One of the difficulties encountered in statistical analyses of the price series of shares listed on the Johannesburg Stock Exchange is that many of the shares are only thinly traded. The market is heavily dominated by institutional investors, and if for any reason a share happens not to be an 'institutional favourite' there will very likely be days, or even weeks, during which no trading of that share takes place. One approach to this problem is to model the presence or absence of trading quite separately from the modelling of the price achieved when trading does take place. This is analogous to the modelling of the sequence of wet and dry days separately from the modelling of

the amounts of precipitation occurring on the wet days. It is therefore natural to consider, as models for the trading pattern of one or several shares, hidden Markov models of the kind used by Zucchini and Guttorp (1991) in the context of daily precipitation at one or several sites.

In order to assess the usefulness of such models for thin trading, trading data for six thinly traded shares were obtained from Mr D. Bowie (Dept. of Statistical Sciences, University of Cape Town), and various models, including two-state hidden Markov models, were fitted and compared. (For the data, see Appendix B.) Of the six shares, three are from the coal sector and three from the diamonds sector. The coal shares are Amcoal, Vierfontein and Wankie, and the diamond shares are Anamint, Broadacres and Carrigs. In all six cases the data cover the period from 5 October 1987 to 3 June 1991 (inclusive), during which time there were 910 days on which trading could take place. The data are therefore a multivariate binary time series of length 910.

The first two univariate models fitted to each of the six shares were a model assuming independence of successive observations and a Markov chain (the latter fitted by conditional maximum likelihood). In all six cases, however, the sample ACF bore little resemblance to the ACF of the Markov chain model, and the Markov chain was therefore considered unsatisfactory. Second-order Markov chains and two-state binomial-hidden Markov models, with and without trend, were also fitted, the former by conditional maximum likelihood and the latter by unconditional. The resulting log-likelihood and BIC values are shown in Table 4.19.

From that table we see that, of the five univariate models considered, the two-state hidden Markov model with a time trend fares best for four of

Table 4.19: Minus log-likelihood values and BIC values achieved by five types of univariate model for six thinly traded shares.

Values of $-l$:

model:	Amcoal	Vierf'n	Wankie	Anamint	Broadac	Carrigs
independence	543.51	629.04	385.53	612.03	599.81	626.88
Markov chain	540.89	611.07	384.57	582.64	585.76	570.25
second-order M. chain	539.89	606.86	383.87	576.99	580.06	555.67
2-state HM, no trend	533.38	588.08	382.51	572.55	562.96	533.89
2-state HM with trend	528.07	577.51	381.28	562.55	556.21	533.88

Values of BIC:

model:	Amcoal	Vierf'n	Wankie	Anamint	Broadac	Carrigs
independence	1093.83	1264.89	777.88	1230.88	1206.43	1260.58
Markov chain	1095.41	1235.77	782.77	1178.91	1185.15	1154.13
second-order M. chain	1107.03	1240.97	794.99	1181.23	1187.37	1138.59
2-state HM, no trend	1094.01	1203.41	792.27	1172.35	1153.17	1095.03
2-state HM with trend	1090.22	1189.08	796.63	1159.17	1146.49	1101.83

the six shares: Amcoal, Vierfontein, Anamint and Broadacres. Of these four shares, all but Anamint show a negative trend in the probability of trading taking place, and Anamint a positive trend. In the case of Wankie, the model assuming independence of successive observations is chosen by BIC, and in the case of Carrigs a hidden Markov model without time trend is chosen.

Since a stationary hidden Markov model is chosen for Carrigs, it is interesting to compare the ACF of that model with the sample ACF and with the ACF of the competing Markov chain model. For the hidden Markov model the ACF is $\rho_k = .3517 \times .9127^k$, and for the Markov chain it is $\rho_k = .3499^k$. Table 4.20 displays the first eight terms in each case. It is clear that the hidden Markov model comes much closer to matching the sample ACF than does the Markov chain model: a two-state hidden Markov model can model slow decay in ρ_k from any starting value ρ_1, but a two-state Markov chain cannot.

Table 4.20: Trading of Carrigs Diamonds. First eight terms of the sample ACF compared with the autocorrelations of two possible models.

sample ACF	.349	.271	.281	.237	.230	.202	.177	.200
ACF of Markov chain	.350	.122	.043	.015	.005	.002	.001	.000
ACF of HM model	.321	.293	.267	.244	.223	.203	.186	.169

Two-state multivariate hidden Markov models of two kinds were then fitted to each of the two groups of three shares: a model without time trend, and one which has a single (logistic-linear) time trend common to the two states and to the three shares in the group. The first type of model has eight parameters, the second has nine. These models were then compared with each other and with the 'product models' obtained by combining indepen-

dent univariate models for the individual shares. The three types of product model considered were those based on independence of successive observations and those obtained by using the univariate hidden Markov models with and without trend. The results are displayed in Table 4.21.

Table 4.21: Comparison of various multivariate models for the three coal shares and the three diamond shares.

Coal shares:

model:	no. parameters	$-l$	BIC
3 'independence' models	3	1558.08	3136.60
3 univariate HM models, no trend	12	1503.97	3089.69
3 univariate HM models with trend	15	1486.86	3075.93
multivariate HM model, no trend	8	1554.01	3162.52
multivariate HM, single trend	9	1538.14	3137.60

Diamond shares:

model:	no. parameters	$-l$	BIC
3 'independence' models	3	1838.72	3697.88
3 univariate HM models, no trend	12	1669.40	3420.56
3 univariate HM models with trend	15	1652.64	3407.48
multivariate HM model, no trend	8	1590.63	3235.77
multivariate HM, single trend	9	1543.95	3149.22

It is clear that, for the coal shares, the multivariate modelling has not been a success: the model consisting of three independent univariate hidden Markov models with trend is 'best'. For the diamond shares, the multivariate hidden Markov model with trend is best of those considered. We therefore give below, in Tables 4.22 and 4.24 respectively, the three univariate models for the coal shares and the multivariate model for the diamond shares. In

Table 4.22 ${}_tp_i$ is the probability that the relevant share is traded on day t if the state of the underlying Markov chain is i, and logit ${}_tp_i = a_i + bt$. Similarly, in Table 4.24 ${}_tp_{ij}$ is the probability that share j is traded on day t if the state is i, and logit ${}_tp_{ij} = a_{ij} + bt$.

Table 4.22: Univariate hidden Markov models (with trend) for the three coal shares.

share	t.p.m.	a_1	a_2	b
Amcoal	$\begin{pmatrix} .774 & .226 \\ .019 & .981 \end{pmatrix}$	−.3316	1.826	−.001488
Vierfontein	$\begin{pmatrix} .980 & .020 \\ .091 & .909 \end{pmatrix}$	.6063	3.358	−.001792
Wankie	$\begin{pmatrix} .807 & .193 \\ .096 & .904 \end{pmatrix}$	−5.028	−0.9428	−.0006810

The parametric bootstrap, with a sample size of 100, was used to investigate the distribution of the estimators in the models displayed in Table 4.22. In this case the estimators show much more variability than do the estimators used for the geyser data analyzed in section 4.2. Table 4.23 gives for each of the three coal shares the bootstrap sample means, medians and standard deviations for the estimators of the five parameters. It will be noted that the estimators of a_1 and a_2 seem particularly variable. It is however true that, except in the middle of the range, very large differences on a logistic scale correspond to small ones on a probability scale: two models with very different values of a_1 (for instance) may therefore produce almost identical distributions for the observations. For all three shares the trend parameter b, which is probably the parameter of most substantive interest, seems to be

more reliably estimated than the other parameters.

Table 4.23: Coal shares. Means, medians and standard deviations of bootstrap sample of estimators of parameters of two-state hidden Markov models with time trend.

share		$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}$
Anamint	sample mean	.251	.048	−1.40	2.61	−.00180
	sample median	.228	.024	−.204	1.95	−.00163
	sample s.d.	.154	.063	4.95	3.18	.00108
Vierfontein	sample mean	.023	.102	.599	4.06	−.00187
	sample median	.020	.097	.624	3.39	−.00190
	sample s.d.	.015	.046	.204	3.70	.00041
Wankie	sample mean	.145	.148	−14.3	.089	−.00081
	sample median	.141	.089	−20.9	−.874	−.00074
	sample s.d.	.085	.167	10.3	4.17	.00070

Table 4.24: Multivariate hidden Markov model (with trend) for the three diamond shares.

t.p.m. $= \begin{pmatrix} .998 & .002 \\ .001 & .999 \end{pmatrix}$, $b = -.003160$, and the parameters a_{ij} are given by:

share	a_{1j}	a_{2j}
Anamint	1.756	1.647
Broadacres	.3642	.6939
Carrigs	1.920	−.9648

As regards the multivariate hidden Markov model for the three diamond shares (Table 4.24), it is perhaps surprising that the model is so much improved by the inclusion of a single (negative) time trend parameter: in the

corresponding univariate models the time trend was positive for one share, negative and of similar magnitude for another share, and negligible for the remaining share. Another criticism to which this multivariate model is open is that the off-diagonal elements of the transition probability matrix of its underlying Markov chain are so close to zero as to be almost negligible: on average only one or two changes of state would take place during a sequence of 910 observations. Furthermore it is not possible to interpret the two states as conditions in which trading is in general more likely and conditions in which it is in general less so. This is because the probability of trading ($_tp_{ij}$) is not consistently higher for one state i than the other.

In view of the relative lack of success of multivariate hidden Markov models in this application, these models are not pursued here. The above discussion does however serve as an illustration of the methodology, and suggests that such multivariate models are a potentially useful tool in studies of this sort. They could, for instance, be used to model occurrences other than the presence or absence of trading, e.g. the price (or volume) rising above some level of particular interest.

4.8 Firearm and non-firearm homicides and suicides, Cape Town, 1986–1991

In South Africa, as in the U.S.A., gun control is a subject of much public interest and debate. Furthermore there is in South Africa an apparently increasing tendency for violent crime to involve firearms. In a project intended to study this and related issues, Dr L.B. Lerer (of the Dept. of Forensic Medicine and Toxicology, University of Cape Town) and two medical students, Ms Z. Gawlowski and Ms R. Phillips, collected data relating to homi-

cides and suicides from the South African Police mortuary in Salt River, Cape Town. Records relating to over 90% of the homicide and suicide cases occurring in metropolitan Cape Town are kept at this mortuary. The remaining cases are dealt with at the Tygerberg hospital. It is believed, however, that the exclusion of the Tygerberg data does not materially affect the conclusions.

The data consist of all the homicide and suicide cases appearing in the deaths registers relating to the six-year period from 1 January 1986 to 31 December 1991. In each such case the following information was recorded: the deaths register reference, the date of death, the sex of the deceased, a racial classification (African, 'coloured' or white), and the cause of death. The categories used for the cause of death were: firearm homicide, non-firearm homicide, firearm suicide, non-firearm suicide, and 'legal intervention homicide'. This last category refers to homicide by members of the police or army in the course of their work. Clearly some of the information recorded in the deaths registers could be inaccurate, e.g. a homicide recorded as a suicide, or a legal intervention homicide recorded as belonging to another category. This has to be borne in mind in drawing any conclusions from the data.

One question of interest that was examined by means of hidden Markov models was whether there is an upward trend in the proportion of all the deaths recorded that are firearm homicides. This is of course quite distinct from the question of whether there is an upward trend in the *number* of firearm homicides. The latter kind of trend could be caused by an increase in the population exposed to risk of death, without there being any change in the proportion. This distinction is important because of the rapid urbanization which has recently taken place in South Africa and has caused the population in and around Cape Town to increase dramatically.

Four models were fitted to the 313 weekly totals of firearm homicides (given the weekly totals of all deaths). For these totals, see Appendix B. The four models are: a two-state binomial-hidden Markov model with constant 'success probabilities' p_1 and p_2, a similar model with a linear time trend (the same for both states) in the logits of those probabilities, a model allowing differing time trend parameters in the two states, and finally a model which assumes that the success probabilities are piecewise constant, with a single change-point at time 287, 26 weeks before the end of the six-year period studied. The time of the change-point was chosen because of the known upsurge of violence in some of the areas adjacent to Cape Town, in the second half of 1991. Much of this violence was associated with the 'taxi wars', a dispute between rival groups of public transport operators.

The models were compared on the basis of AIC and BIC. The results are shown in Table 4.25. Broadly, the conclusion from BIC is that a single (upward) time trend is better than either no trend or two trend parameters, but the model with a change-point is the best of the four. The details of this model are as follows. The underlying Markov chain has transition probability matrix

$$\begin{pmatrix} .658 & .342 \\ .254 & .746 \end{pmatrix}$$

and stationary distribution (.426 .574). The probabilities p_1 and p_2 are given by (.050 .116) for weeks 1–287, and by (.117 .253) for weeks 288–313. From this it appears that the proportion of the deaths that are firearm homicides was substantially greater by the second half of 1991 than it was earlier, and that this change is better accommodated by a discrete shift in the probabilities p_1 and p_2 than by gradual movement with time, at least gradual movement of the kind incorporated into the models with time trend. (In passing, this use of a discrete shift further illustrates the flexibility of hidden

Markov models.) One other model was also fitted: a model with change point at the end of week 214. That week included 2 February 1990, on which day President De Klerk made a speech which is widely regarded as a watershed in South Africa's recent history. That model yielded a log-likelihood of -579.83, and AIC and BIC values of 1171.67 and 1194.14. Such a model is therefore superior to the models with time trend, but inferior to the model with the change-point at the end of week 287, and was therefore not considered further.

Table 4.25: Comparison of various binomial-hidden Markov models fitted to the weekly totals of firearm homicides given the weekly totals of all deaths.

model with:	no. parameters	$-l$	AIC	BIC
p_1 and p_2 constant	4	590.26	1188.52	1203.50
one time trend parameter	5	584.34	1178.67	1197.40
two trend parameters	6	581.87	1175.75	1198.23
change-point at time 287	6	573.27	1158.55	1181.03

In order to model the number (rather than the proportion) of firearm homicides, Poisson-HM models were also fitted. The four models fitted in this case were: a two-state model with constant conditional means λ_1 and λ_2, a similar model with a single linear trend in the logs of those means, a model with a quadratic trend therein, and finally a model allowing for a change-point at time 287. A comparison of these models is shown in Table 4.26. The conclusion is that, of the four models, the model with a quadratic trend in the conditional means is best. In detail, that model is as follows. The underlying Markov chain has transition probability matrix

$$\begin{pmatrix} .881 & .119 \\ .416 & .584 \end{pmatrix}$$

and stationary distribution (.777 .223). The conditional means are given by

$$\log \lambda_1 = .4770 - .004858t + .00002665t^2 ,$$

where t is the week number, and

$$\log \lambda_2 = 1.370 - .004858t + .00002665t^2 .$$

The fact that this smooth trend works better here than does a discrete shift may possibly be explained by population increase due to migration, especially towards the end of the six-year period.

Table 4.26: Comparison of various Poisson-hidden Markov models fitted to the weekly totals of firearm homicides.

model with:	no. parameters	$-l$	AIC	BIC
λ_1 and λ_2 constant	4	626.64	1261.27	1276.26
log-linear trend	5	606.82	1223.65	1242.38
log-quadratic trend	6	602.27	1216.55	1239.02
change-point at time 287	6	605.56	1223.12	1245.60

A question of interest that arises from the apparently increased proportion of firearm homicides is whether there is any similar tendency in respect of suicides. Here the most interesting comparison is between firearm homicides as a proportion of all homicides and firearm suicides as a proportion of all suicides. Binomial-hidden Markov models of various types were therefore used to model these proportions, and the results are given in Tables 4.27 and 4.28.

The chosen models for these two proportions are therefore as follows. For

Table 4.27: Comparison of binomial-hidden Markov models for firearm homicides given all homicides.

model with:	no. parameters	$-l$	AIC	BIC
p_1 and p_2 constant	4	590.75	1189.49	1204.48
one time trend parameter	5	585.59	1181.17	1199.90
two trend parameters	6	583.98	1179.95	1202.43
change-point at time 287	6	575.04	1162.07	1184.55

Table 4.28: Comparison of binomial-hidden Markov models for firearm suicides given all suicides.

model with:	no. parameters	$-l$	AIC	BIC
p_1 and p_2 constant	4	289.93	587.86	602.84
one time trend parameter	5	289.22	588.45	607.18
two trend parameters	6	288.30	588.61	611.09
change-point at time 287	6	289.21	590.42	612.90

the firearm homicides the Markov chain has transition probability matrix

$$\begin{pmatrix} .695 & .305 \\ .283 & .717 \end{pmatrix}$$

and stationary distribution (.481 .519). The probabilities p_1 and p_2 are given by (.060 .140) for weeks 1–287, and by (.143 .283) for weeks 288–313. The unconditional probability that a homicide involved the use of a firearm is therefore .102 before the change-point, and .216 thereafter. For the firearm suicides, the transition probability matrix is

$$\begin{pmatrix} .854 & .146 \\ .117 & .883 \end{pmatrix}$$

and the stationary distribution is (.446 .554). The probabilities p_1 and p_2 are given by (.186 .333), and the unconditional probability that a suicide involves a firearm is .267. The conclusion is that the proportion of homicides that involve firearms does indeed seem to be higher after June 1991, but that there is no evidence of a similar upward shift (or trend) in respect of the proportion of suicides that involve firearms.

As a final illustration of the application of hidden Markov models to this data-set we describe here two multinomial-hidden Markov models for the weekly totals in each of the five categories of death. These models are of the kind introduced in section 3.3. Each has two states. One model has constant 'success probabilities' and the other allows for a change in these probabilities at time 287. The model without change-point has ten parameters: two to determine the Markov chain, and four independently determined probabilities for each of the two states. The model with change-point has eighteen parameters, since there are eight independent probabilities relevant to the period before the change-point, and eight after. For the model without change-point, $-l$ (apart from the constant term involving the multinomial

coefficients) is 6463.7, and for the model with change-point it is 6429.3. The corresponding AIC and BIC values are 12947.5 and 12985.0 (without change-point), and 12894.6 and 12962.0 (with). The model with the change-point at time 287 is therefore preferred, and we give it in full here. The underlying Markov chain has transition probability matrix

$$\begin{pmatrix} .541 & .459 \\ .097 & .903 \end{pmatrix}$$

and stationary distribution (.174 .826). Table 4.29 displays, for the period up to the change-point and the period thereafter, the probability of each category of death in state 1 and in state 2, and the corresponding unconditional probabilities. The most noticeable difference between the period before the change-point and the period thereafter is the sharp increase in the unconditional probability of category 1 (firearm homicide), with corresponding decreases in all the other categories.

Clearly the above brief discussion does not attempt to pursue all the questions of interest arising from this data-set that may be answered by the fitting of hidden Markov (or other) time series models. It is felt, however, that the models described, and the conclusions that may be drawn, are sufficiently illustrative of the technique to make clear its utility and flexibility. A fuller analysis and discussion of these data is in preparation (Lerer and MacDonald, 1992).

Table 4.29: Multinomial-hidden Markov model with change-point at time 287. Probabilities associated with each category of death, before and after the change-point.

Weeks 1–287:

	category 1	2	3	4	5
in state 1	.124	.665	.053	.098	.059
in state 2	.081	.805	.024	.074	.016
unconditional	.089	.780	.029	.079	.023

Weeks 288–313:

	category 1	2	3	4	5
in state 1	.352	.528	.010	.075	.036
in state 2	.186	.733	.019	.054	.008
unconditional	.215	.697	.018	.058	.013

Categories:
1 firearm homicide
2 non-firearm homicide
3 firearm suicide
4 non-firearm suicide
5 legal intervention homicide.

4.9 Conclusion

It seems evident that, at least for some time to come, the analysis of discrete-valued time series will continue to be far less unified than that of continuous-valued series. In the latter case many practical problems can be tackled by a unified strategy based on one class of models, the Gaussian ARMA models. For discrete-valued series, on the other hand, a wide range of models and techniques seems necessary. This chapter has explored the use of just one class of possible models, in a variety of fields of application: medical and sociological applications, finance, animal behaviour, geophysics and climatology. Hidden Markov models of widely differing kinds have been illustrated: models for bounded and for unbounded counts, models with and without time trend, models with change-points, multivariate models, models for categorical series, models with cyclical components and models with a second-order Markov chain as parameter process. From Chapter 3 it will be apparent that the selection of applications presented does not by any means exhaust the variations and techniques that are possible. Although there are many applications for which a parameter-driven model will be inappropriate, the versatility and flexibility of hidden Markov models does make them a promising approach to the modelling of those discrete-valued time series for which parameter-driven models are appropriate.

There are of course some important statistical questions relating to the use of hidden Markov models which have not been addressed in this thesis: here we include the distribution of goodness-of-fit statistics for the models, the asymptotic properties of the maximum likelihood estimators of the parameters, and the use of likelihood ratio statistics for tests of hypotheses on nested models. It is however hoped that the theory presented in Chapters 2 and 3 and the illustrative applications presented in this chapter will persuade

the reader that hidden Markov models are a useful addition to the techniques available to the statistician who meets discrete-valued time series in her or his work.

Appendix A

Proofs of certain results used in the derivation of the Baum-Welch algorithm

The purpose of this appendix is to derive the four properties stated without proof in section 2.2 and used there in the derivation of the Baum-Welch algorithm. All four results refer to the following situation. The processes $\{C_t : t \in \mathbf{N}\}$ and $\{S_t : t \in \mathbf{N}\}$ are finite state-space processes. $\{C_t\}$ is a homogeneous Markov chain and $\{S_t\}$ has (for all T) the property that, conditional on $C^{(T)}$, the random variables $S_1, S_2, \ldots, S_T$ are mutually independent and the (conditional) distribution of S_t is given by $\mathrm{P}(S_t \mid C_t)$. That is, this distribution depends only on C_t and not on any $C_k, k \neq t$.

The technique of proof is in general as follows:

(a) express the probability of interest in terms of probabilities conditional on $C^{(T)}$, i.e. conditional on all of $C_1, \ldots, C_T$;

(b) use the fact that, conditional on $C^{(T)}$, the random variables $S_1, \ldots, S_T$ are independent, with the distribution of each S_t depending only on

the corresponding C_t;

(c) use the Markov property of $\{C_t\}$ if necessary.

We establish first the property (2.1), and then derive property (2.4) from it. To establish property (2.1) we use Propositions 1, 2 and 3 given below.

Proposition 1 *For all t and l such that $1 \leq t \leq l \leq T$:*

$$\mathrm{P}(S_l, S_{l+1}, \ldots, S_T \mid C_t, \ldots, C_T) = \mathrm{P}(S_l, \ldots, S_T \mid C_l, \ldots, C_T).$$

Proof: The left-hand side above can be written as:

$$\frac{1}{\mathrm{P}(C_t, \ldots, C_T)} \sum_{c_1, \ldots, c_{t-1}} \mathrm{P}(S_l, \ldots, S_T \mid C^{(T)})\, \mathrm{P}(C^{(T)}),$$

there being no summation in the case $t=1$. By (b) we have

$$\mathrm{P}(S_l, \ldots, S_T \mid C^{(T)}) = \mathrm{P}(S_l \mid C_l) \ldots \mathrm{P}(S_T \mid C_T),$$

which can be taken outside the summation. The resulting sum then reduces to $\mathrm{P}(C_t, \ldots, C_T)$, and the left-hand side is seen to be just

$$\mathrm{P}(S_l \mid C_l) \ldots \mathrm{P}(S_T \mid C_T),$$

which is independent of t. The right-hand side, being the case $t = l$ of the left-hand side, equals the same expression. □

Proposition 2 *For $t = 1, 2, \ldots, T-1$:*

$$\mathrm{P}(S_{t+1}, \ldots, S_T \mid C^{(t)}) = \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t).$$

Proof: The left-hand side can be written as

$$\frac{1}{\mathrm{P}(C^{(t)})} \sum_{c_{t+1}, \ldots, c_T} \mathrm{P}(C^{(T)})\, \mathrm{P}(S_{t+1}, \ldots, S_T \mid C^{(T)}).$$

Now apply Proposition 1 (twice) and the Markov property of $\{C_t\}$ to see that this equals

$$\sum_{c_{t+1},\ldots,c_T} \mathrm{P}(C_{t+1},\ldots,C_T \mid C_t)\ \mathrm{P}(S_{t+1},\ldots,S_T \mid C_t,\ldots,C_T).$$

The summand is $\mathrm{P}(S_{t+1},\ldots,S_T,C_t,\ldots,C_T)/\mathrm{P}(C_t)$, and the sum is therefore equal to $\mathrm{P}(S_{t+1},\ldots,S_T,C_t)/\mathrm{P}(C_t)$, as required. □

Proposition 3 *For $t = 1,2,\ldots,T$:*

$$\mathrm{P}(S_1,\ldots,S_t \mid C^{(T)}) = \mathrm{P}(S_1,\ldots,S_t \mid C^{(t)}).$$

Proof: Apply (b) in respect of the conditioning on $C^{(T)}$ to see that the left-hand side equals $\mathrm{P}(S_1 \mid C_1)\ldots\mathrm{P}(S_t \mid C_t)$. Apply (b) in respect of the conditioning on $C^{(t)}$ to see that the right-hand side equals the same expression. □

Proposition 4 (Property (2.1)) *For $t = 1,2,\ldots,T$:*

$$\mathrm{P}(S_1,\ldots,S_T \mid C_t) = \mathrm{P}(S_1,\ldots,S_t \mid C_t)\ \mathrm{P}(S_{t+1},\ldots,S_T \mid C_t).$$

Proof: Making use of the mutual independence of $S_1,\ldots,S_T$ given $C^{(T)}$, write the left-hand side as

$$\frac{1}{\mathrm{P}(C_t)}\sum\nolimits^{(1)}\sum\nolimits^{(2)}\mathrm{P}(C^{(T)})\ \mathrm{P}(S_1,\ldots,S_t \mid C^{(T)})\ \mathrm{P}(S_{t+1},\ldots,S_T \mid C^{(T)}),$$

where $\sum^{(1)}$ denotes summation over $c_1,\ldots,c_{t-1}$, and $\sum^{(2)}$ over $c_{t+1},\ldots,c_T$. Then apply Propositions 3 and 2 to show that this equals

$$\begin{aligned} &\frac{1}{\mathrm{P}(C_t)}\left(\sum\nolimits^{(1)}\mathrm{P}(S_1,\ldots,S_t,C_1,\ldots,C_t)\right)\mathrm{P}(S_{t+1},\ldots,S_T \mid C_t) \\ = \ &\frac{1}{\mathrm{P}(C_t)}\mathrm{P}(S_1,\ldots,S_t,C_t)\ \mathrm{P}(S_{t+1},\ldots,S_T \mid C_t). \end{aligned}$$

□

Proposition 5 (Property (2.4)) *For* $t = 1, 2, \ldots, T$*:*

$$\mathrm{P}(S_t, \ldots, S_T \mid C_t) = \mathrm{P}(S_t \mid C_t)\ \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_t).$$

Proof: Sum the result of Proposition 4 with respect to $s_1, \ldots, s_{t-1}$. □

Proposition 6 (Property (2.2)) *For* $t = 1, 2, \ldots, T-1$*:*

$$\mathrm{P}(S_1, \ldots, S_T \mid C_t, C_{t+1}) = \mathrm{P}(S_1, \ldots, S_t \mid C_t)\ \mathrm{P}(S_{t+1}, \ldots, S_T \mid C_{t+1}).$$

Proof: Write the left-hand side as

$$\frac{1}{\mathrm{P}(C_t, C_{t+1})} \sum\nolimits^{(1)} \sum\nolimits^{(2)} \mathrm{P}(C^{(T)})\ \mathrm{P}(S_1, \ldots, S_t \mid C^{(T)})\ \mathrm{P}(S_{t+1}, \ldots, S_T \mid C^{(T)}),$$

where $\sum^{(1)}$ denotes summation over $c_1, \ldots, c_{t-1}$ and $\sum^{(2)}$ over $c_{t+2}, \ldots, c_T$. By Propositions 3 and 1 respectively, the last two factors in the above expression reduce to $\mathrm{P}(S_1, \ldots, S_t \mid C^{(t)})$ and $\mathrm{P}(S_{t+1}, \ldots, S_T \mid C_{t+1}, \ldots, C_T)$. The Markov property of $\{C_t\}$ is then used, and after some routine manipulations of conditional probabilities it emerges that the above expression is equal to :

$$\begin{aligned} & \mathrm{P}(S_1, \ldots, S_t \mid C_t) \frac{1}{\mathrm{P}(C_{t+1})} \sum\nolimits^{(2)} \mathrm{P}(S_{t+1}, \ldots, S_T, C_{t+1}, \ldots, C_T) \\ = \ & \mathrm{P}(S_1, \ldots, S_t \mid C_t)\ \mathrm{P}(S_{t+1}, \ldots, S_T, C_{t+1}) / \mathrm{P}(C_{t+1}), \end{aligned}$$

as required. □

Proposition 7 (Property (2.3)) *For all* t *and* l *such that* $1 \leq t \leq l \leq T$*:*

$$\mathrm{P}(S_l, \ldots, S_T \mid C_t, \ldots, C_l) = \mathrm{P}(S_l, \ldots, S_T \mid C_l).$$

Proof: If $\sum^{(1)}$ denotes summation over $c_1, \ldots, c_{t-1}$ and $\sum^{(2)}$ over $c_{l+1}, \ldots, c_t$, the left-hand side can be written as

$$\frac{1}{\mathrm{P}(C_t, \ldots, C_l)} \sum\nolimits^{(2)} \sum\nolimits^{(1)} \mathrm{P}(S_l, \ldots, S_T \mid C^{(T)})\ \mathrm{P}(C^{(T)}).$$

By Proposition 1

$$\mathrm{P}(S_l, \ldots, S_T \mid C^{(T)}) = \mathrm{P}(S_l, \ldots, S_T \mid C_l, \ldots, C_T),$$

and the above expression for the left-hand side therefore equals

$$\sum^{(2)} \mathrm{P}(S_l, \ldots, S_T \mid C_l, \ldots, C_T)\ \mathrm{P}(C_{l+1}, \ldots, C_T \mid C_t, \ldots, C_l).$$

By the Markov property of $\{C_t\}$, this equals

$$\begin{aligned} & \sum^{(2)} \mathrm{P}(S_l, \ldots, S_T \mid C_l, \ldots, C_T)\ \mathrm{P}(C_{l+1}, \ldots, C_T \mid C_l) \\ = \ & \frac{1}{\mathrm{P}(C_l)} \sum^{(2)} \mathrm{P}(S_l, \ldots, S_T, C_l, \ldots, C_T) \\ = \ & \mathrm{P}(S_l, \ldots, S_T, C_l)/\mathrm{P}(C_l), \end{aligned}$$

i.e. the right-hand side. □

Appendix B

The data-sets

In this appendix we give some relevant information about five of the data-sets discussed in Chapter 4, and indicate in particular how the data are stored on the disk accompanying the thesis. Anyone wishing to use any of these data is requested to consult the person who provided the data. These people are identified in the text.

There are five ASCII files on the disk: BIRTHS.DAT, FEED2.01, WINDALL.DAT, JSEDAY2.DAT and CODALL.DAT. We describe these files in that order.

- BIRTHS.DAT contains the 'Edendale births' data analyzed in section 4.3. There are 192 rows of data. Row i contains the data for month i, February 1970 being month 1. For each month there are the following items, in order: the month number, the number of mothers delivered, the number of Caesarean sections performed, the number of breech births, the number of forceps deliveries, the number of vacuum extraction deliveries, the number of stillbirths, the number of neonatal deaths and the number of maternal deaths. The number of maternal deaths for month 10 is missing.

- FEED2.01 contains the locust data analyzed in section 4.4. There are 161 rows and 24 columns of data. Row i refers to time i and column j to subject j. The symbol '1' in position (i, j) indicates that subject j was locomoting at time i, and the symbol '0' that that subject was not locomoting.

- WINDALL.DAT contains the wind direction data analyzed in section 4.5. Apart from the first row, this file contains 8766 rows of data. In each such row there are five items. Read from left to right, these are: a sequence number, and four consecutive hourly observations of wind direction, classified into directions 1 to 16 as described in the text. The sequence numbers correspond to row numbers in smaller files which were merged to produce WINDALL.DAT, and run from 1 to 3660, then from 1 to 4386, and then from 1 to 720.

- JSEDAY2.DAT contains the thin trading data analyzed in section 4.7. There are 910 rows and 6 columns of data. Row i refers to trading day i, and the six columns (from left to right) refer to the shares Amcoal, Vierfontein, Wankie, Anamint, Broadacres and Carrigs. A '1' denotes that trading took place, a '0' that trading did not take place.

- CODALL.DAT contains the homicide and suicide data analyzed in section 4.8. There are 313 rows, corresponding to weeks in the period 1986–1991 (inclusive). Each row contains (from left to right): the week number, the numbers of deaths falling into categories 1–5, the total number of deaths, and the proportions of the deaths falling into categories 1–5. For the definitions of the categories, see the text.

References

Adke, S.R. and Deshmukh, S.R. (1988). Limit distribution of a high order Markov chain. *J. R. Statist. Soc.* B **50**, 105–108.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Al-Osh, M.A. and Alzaid, A.A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Ser. Anal.* **8**, 261–275.

Al-Osh, M.A. and Alzaid, A.A. (1988). Integer-valued moving average (INMA) process. *Statist. Papers* **29**, 281–300.

Al-Osh, M.A. and Alzaid, A.A. (1991). Binomial autoregressive moving average models. *Stoch. Models* **7**, 261–282.

Albert, P.S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics* **47**, 1371–1381.

Alzaid, A.A. and Al-Osh, M.A. (1988). First-order integer-valued autoregressive (INAR(1)) process: distributional and regression properties. *Statist. Neerl.* **42**, 53–61.

Alzaid, A.A. and Al-Osh, M.A. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *J. Appl. Prob.* **27**, 314–324.

Azzalini, A. (1982). Approximate filtering of parameter driven processes. *J. Time Ser. Anal.* **3**, 219–223.

Azzalini, A. (1983). Maximum likelihood estimation of order m for stationary stochastic processes. *Biometrika* **70**, 381–387.

Azzalini, A. and Bowman, A.W. (1990). A look at some data on the Old Faithful geyser. *Appl. Statist.* **39**, 357–365.

Basawa, I.V. and Prakasa Rao, B.L.S. (1980). *Statistical Inference for Stochastic Processes.* Academic Press, London.

Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Proc. Third Symposium on Inequalities*, ed. O. Shisha, Academic Press, New York, 1–8.

Baum, L.E. and Eagon, J.A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73**, 360–363.

Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37**, 1554–1563.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41**, 164–171.

Baum, L.E. and Sell, G.R. (1968). Growth transformations for functions on manifolds. *Pacific J. Math.* **27**, 211–227.

Billingsley, P. (1961). Statistical methods in Markov chains. *Ann. Math. Statist.* **32**, 12–40.

Bisgaard, S. and Travis, L.E. (1991). Existence and uniqueness of the solution of the likelihood equations for binary Markov chains. *Statist. Prob. Letters* **12**, 29–35.

Blight, P.A. (1989). Time series formed from the superposition of discrete renewal processes. *J. Appl. Prob.* **26**, 189–195.

Block, H.W., Langberg, N.A. and Stoffer, D.S. (1988). Bivariate exponential and geometric autoregressive and autoregressive moving average models. *Adv. Appl. Prob.* **20**, 798–821.

Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis, Forecasting and Control*, revised edition. Holden-Day, Oakland, California.

Brockwell, P.J. and Davis, R.A. (1987). *Time Series: Theory and Methods.* Springer, New York.

Buishand, T.A. (1978). The binary DARMA(1,1) process as a model for wet-dry sequences. Technical note 78-01, Dept. of Mathematics, Statistics Division, Agricultural University, Wageningen, The Netherlands.

Campillo, F. and Le Gland, F. (1989). MLE for partially observed diffusions: direct maximization vs the EM algorithm. *Stoch. Processes Appl.* **33**, 245–274.

Cane, V.R. (1978). On fitting low-order Markov chains to behaviour sequences. *Anim. Behav.* **26**, 332–338.

Chang, T.J., Delleur, J.W. and Kavvas, M.L. (1987). Application of discrete autoregressive moving average models for estimation of daily runoff. *J. Hydrol.* **91**, 119–135.

Chang, T.J., Kavvas, M.L. and Delleur, J.W. (1984a). Modeling of sequences of wet and dry days by binary discrete autoregressive moving average processes. *J. Clim. Appl. Meteorol.* **23**, 1367–1378.

Chang, T.J., Kavvas, M.L. and Delleur, J.W. (1984b). Daily precipitation modeling by discrete autoregressive moving average processes. *Water Resour. Res.* **20**, 565–580.

Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94.

Cox, D.R. (1981). Statistical analysis of time series: some recent developments. *Scand. J. Statist.* **8**, 93–115.

Cox, D.R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5**, 169–174.

Cox, D.R. and Lewis, P.A.W. (1966). *The Statistical Analysis of Series of Events.* Methuen, London.

Cox, D.R. and Miller, H.D. (1965). *The Theory of Stochastic Processes.* Chapman and Hall, London.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*, second edition. Chapman and Hall, London.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

Devore, J.L. (1976). A note on the estimation of parameters in a Bernoulli model with dependence. *Ann. Statist.* **4**, 990–992.

Diggle, P. and Westcott, M. (1985). Contribution to the discussion of Lawrance and Lewis (1985). *J. R. Statist. Soc.* B **47**, 192–193.

Du Jin-Guan and Li Yuan (1991). The integer-valued autoregressive (INAR-(p)) model. *J. Time Ser. Anal.* **12**, 129–142.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Ephraim, Y. and Rabiner, L.R. (1990). On the relations between modeling approaches for speech recognition. *IEEE Trans. Inform. Theory* **36**, 372–380.

Fahrmeir, L. and Kaufmann, H. (1987). Regression models for non-stationary categorical time series. *J. Time Ser. Anal.* **8**, 147–160.

Fildes, R. (1991). Review of Harvey (1989) and West and Harrison (1989). *J. Opl. Res. Soc.* **42**, 1031–1033.

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

Gill, P.E., Hammarling, S.J., Murray, W., Saunders, M.A. and Wright, M.H. (1986). *User's Guide for LSSOL.* Department of Operations Research, Stanford University, Report SOL 86-1.

Grimmett, G.R. and Stirzaker, D.R. (1982). *Probability and Random Processes.* Oxford University Press, Oxford.

Grunwald, G.K. (1987). Time series models for continuous proportions. Ph.D. dissertation, Dept. of Statistics, University of Washington.

Guttorp, P. (1986). On binary time series obtained from continuous time point processes describing rainfall. *Water Resour. Res.* **22**, 897–904.

Guttorp, P., Newton, M.A. and Abkowitz, J.L. (1990). A stochastic model for haematopoiesis in cats. *IMA J. of Math. Appl. in Medicine and Biology* **7**, 125–143.

Guttorp, P. and Thompson, M.L. (1990). Nonparametric estimation of intensities for sampled counting processes. *J. R. Statist. Soc.* B **52**, 157–173.

Haines, L.M., Munoz, W.P. and van Gelderen, C.J. (1989). ARIMA modelling of birth data. *J. Appl. Statist.* **16**, 55–67.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge.

Harvey, A.C. and Fernandes, C. (1989a). Time series models for count or qualitative observations. *J. Bus. Econ. Statist.* **7**, 407–422.

Harvey, A.C. and Fernandes, C. (1989b). Time series models for insurance claims. *J. Inst. Act.* **116**, 513–528.

Holden, R.T. (1987). Time series analysis of a contagious process. *J. Amer. Statist. Ass.* **82**, 1019–1026.

Jacobs, P.A. and Lewis, P.A.W. (1978a). Discrete time series generated by mixtures I: Correlational and runs properties. *J. R. Statist. Soc.* B **40**, 94–105.

Jacobs, P.A. and Lewis, P.A.W. (1978b). Discrete time series generated by mixtures II: Asymptotic properties. *J. R. Statist. Soc.* B **40**, 222–228.

Jacobs, P.A. and Lewis, P.A.W. (1978c). Discrete time series generated by mixtures III: Autoregressive processes (DAR(p)). Technical report NPS55-78-022, Naval Postgraduate School, Monterey, California.

Jacobs, P.A. and Lewis, P.A.W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal.* **4**, 19–36.

Juang, B.H. (1985). Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T Tech. J.* **64**, 1235–1249.

Juang, B.H. and Rabiner, L.R. (1991). Hidden Markov models for speech recognition. *Technometrics* **33**, 251–272.

Kanter, M. (1975). Autoregression for discrete processes mod 2. *J. Appl. Prob.* **12**, 371–375.

Katz, R.W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics* 23, 243–249.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *Ann. Statist.* **15**, 79–98.

Kedem, B. (1976). Sufficient statistics associated with a two-state second-order Markov chain. *Biometrika* **63**, 127–132.

Kedem, B. (1980). *Binary Time Series.* Marcel Dekker, New York.

Keenan, D.M. (1982). A time series analysis of binary data. *J. Amer. Statist. Ass.* **77**, 816–821.

Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley, Chichester.

Kemeny, J.G., Snell, J.L. and Knapp, A.W. (1976). *Denumerable Markov Chains.* Springer, New York.

Kendall, M.G. and Stuart, A. (1979). *The Advanced Theory of Statistics, Vol. 2*, fourth edition. Griffin, London.

Klimko, L.A. and Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* **6**, 629–642.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. *Ann. Statist.* **1**, 373–379.

Langberg, N.A. and Stoffer, D.S. (1987). Moving-average models with bivariate exponential and geometric distributions. *J. Appl. Prob.* **24**, 48–61.

Lawrance, A.J. (1976). On conditional and partial correlation. *Amer. Statistician* **30**, 146–149.

Lawrance, A.J. (1982). The innovation distribution of a gamma distributed autoregressive process. *Scand. J. Statist.* **9**, 234–236.

Lawrance, A.J. and Lewis, P.A.W. (1980). The exponential autoregressive-moving average EARMA(p, q) process. *J. R. Statist. Soc.* B **42**, 150–161.

Lawrance, A.J. and Lewis, P.A.W. (1981). A new autoregressive time series model in exponential variables (NEAR(1)). *Adv. Appl. Prob.* **13**, 826–845.

Lawrance, A.J. and Lewis, P.A.W. (1985). Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. *J. R. Statist. Soc.* B **47**, 165–183.

Lerer, L.B. and MacDonald, I.L. (1992). Homicides and suicides in Cape Town, 1986–1991: a statistical study. In preparation.

Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, to appear.

Levinson, S.E., Rabiner, L.R. and Sondhi, M.M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov

process to automatic speech recognition. *Bell System Tech. J.* **62**, 1035–1074.

Lewis, P.A.W. (1985). Some simple models for continuous variate time series. *Water Resour. Bull.* **21**, 635–644.

Lewis, P.A.W., McKenzie, E. and Hugus, D.K. (1989). Gamma processes. *Stoch. Models* **5**, 1–30.

Li, W.K. (1991) Testing model adequacy for some Markov regression models for time series. *Biometrika* **78**, 83–89.

Li, W.K. and Kwok, M.C.O. (1990). Some results on the estimation of a higher order Markov chain. *Commun. Statist.-Simul. Comp.* **19**, 363–380.

Liang, K.-Y. and Zeger, S.L. (1989). A class of logistic regression models for multivariate binary time series. *J. Amer. Statist. Ass.* **84**, 447–451.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Lloyd, E.H. (1980). *Handbook of Applicable Mathematics, Vol. 2: Probability.* Wiley, New York.

MacDonald, I.L. (1989). Discrete-valued time series and generalized linear models. Paper presented at the 1989 annual conference of the South African Statistical Association, University of the Witwatersrand, Johannesburg.

MacDonald, I.L. (1990). A new class of models for discrete-valued time series. Paper presented at the 1990 annual conference of the South African Statistical Association, University of Cape Town.

Martin, R.D. and Raftery, A.E. (1987). Robustness, computation, and non-Euclidean (*sic*) models. *J. Amer. Statist. Ass.* **82**, 1044–1050.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, second edition. Chapman and Hall, London.

McInnes, F. and Jack, M. (1988). Automatic speech recognition using word reference patterns. *Speech Technology: a survey*, eds. M. Jack and J. Laver, Edinburgh University Press, Edinburgh, 1–68.

McKenzie, E. (1981). Extending the correlation structure of exponential autoregressive-moving-average processes. *J. Appl. Prob.* **18**, 181–189.

McKenzie, E. (1985a). Contribution to the discussion of Lawrance and Lewis (1985). *J. R. Statist. Soc.* B **47**, 187–188.

McKenzie, E. (1985b). Some simple models for discrete variate time series. *Water Resour. Bull.* **21**, 645–650.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv. Appl. Prob.* **18**, 679–705.

McKenzie, E. (1987). Innovation distributions for gamma and negative binomial autoregressions. *Scand. J. Statist.* **14**, 79–85.

McKenzie, E. (1988a). The distributional structure of finite moving-average processes. *J. Appl. Prob.* **25**, 313–321.

McKenzie, E. (1988b). Some ARMA models for dependent sequences of Poisson counts. *Adv. Appl. Prob.* **20**, 822–835.

Mehran, F. (1989). Analysis of discrete longitudinal data: infinite-lag Markov models. *Statistical Data Analysis and Inference*, ed. Y. Dodge, Elsevier Science Publishers, Amsterdam, 533–541.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc.* B **51**, 127–138.

Munoz, W.P., Haines, L.M. and van Gelderen, C.J. (1987). An analysis of the maternity data of Edendale Hospital in Natal for the period 1970–1985. Part 1: Trends and seasonality. Internal report, Edendale Hospital.

Nádas, A. (1983). Hidden Markov chains, the forward-backward algorithm, and initial statistics. *IEEE Trans. Acoust., Speech, Signal Processing* **31**, 504–506.

Noble, B. (1969). *Applied Linear Algebra.* Prentice-Hall, Englewood Cliffs, New Jersey.

Numerical Algorithms Group (1990). *User's Manual, NAG Fortran Library.* Numerical Algorithms Group, Oxford.

Pegram, G.G.S. (1980). An autoregressive model for multilag Markov chains. *J. Appl. Prob.* **17**, 350–362.

Poritz, A.B. (1988). Hidden Markov models: a guided tour. *Proc. 1988 Int. Conf. Acoust., Speech, Signal Processing,* IEEE Press, New York, 7–13.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986). *Numerical Recipes: the Art of Scientific Computing.* Cambridge University Press, Cambridge.

Priestley, M.B. (1981). *Spectral Analysis and Time Series.* Academic Press, London.

Raftery, A.E. (1985a). A model for high-order Markov chains. *J. R. Statist. Soc.* B **47**, 528–539.

Raftery, A.E. (1985b). A new model for discrete-valued time series: autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni*, **3–4**, 149–162.

Rugg, D.J. and Buech, R.R. (1990). Analyzing time budgets with Markov chains. *Biometrics* **46**, 1123–1131.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Seneta, E.E. (1981). *Non-negative Matrices and Markov Chains.* Springer, New York.

Steutel, F.W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Prob.* **7**, 893–899.

Stoffer, D.S. (1985). Central limit theorems for finite Walsh-Fourier transforms of weakly stationary time series. *J. Time Ser. Anal.* **6**, 261–267.

Stoffer, D.S. (1987). Walsh-Fourier analysis of discrete-valued time series. *J. Time Ser. Anal.* **8**, 449–467.

Stoffer, D.S. (1990). Multivariate Walsh-Fourier analysis. *J. Time Ser. Anal.* **11**, 57–73.

Stoffer, D.S. (1991). Walsh-Fourier analysis and its statistical applications. *J. Amer. Statist. Ass.* **86**, 461–485.

Teicher, H. (1954). On the multivariate Poisson distribution. *Skand. Aktuarietidskr.* **37**, 1–9.

Thompson, E.A. (1983). Optimal sampling for pedigree analysis: parameter estimation and genotypic uncertainty. *Theor. Pop. Biol.* **24**, 39–58.

West, M. and Harrison, P.J. (1989). *Bayesian Forecasting and Dynamic Models.* Springer, New York.

West, M., Harrison, P.J. and Migon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Ass.* **80**, 73–97.

Wichmann, B.A. and Hill, I.D. (1982). An efficient and portable pseudo-random number generator. *Appl. Statist.* **31**, 188-190. Correction, *Appl. Statist.* **33** (1984), 123.

Winkler, R.L. (1989). Contribution to the discussion of Harvey and Fernandes (1989a). *J. Bus. Econ. Statist.* **7**, 419–422.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621–629.

Zeger, S.L. and Liang, K.-Y. (1991). Feedback models for discrete and continuous time series. *Statistica Sinica* **1**, 51–64.

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44**, 1019–1031.

Zucchini, W. and Guttorp, P. (1991). A hidden Markov model for space-time precipitation. *Water Resour. Res.* **27**, 1917–1923.