# Contents



**Other Models and Alternative Approaches**

This chapter introduces the reader to a fair number of additional regression and autoregression models appropriate for integer–valued time series, switching models, models of hidden periodicities, mixture models, and more.

Some of these models have been known for a long time but their debut in the time series literature is fairly recent. We intend to provide enough useful information without delving deeply into mathematical details.

# 1 Integer Autoregressive and Moving Average Models

## 1.1 Branching Processes With Immigration

An important model for integer-valued time series is the *branching process with immigration*, also known as the *Galton-Watson* process with immigration, defined by the stochastic equation

$$X_n = \sum_{i=1}^{X_{n-1}} Y_{n,i} + I_n, \qquad n = 1, 2, 3, \cdots, \tag{1}$$

where the initial value $X_0$ is a nonnegative integer-valued random variable, and $\sum_1^0 \equiv 0$. The processes $\{Y_{n,i}\}$ and $\{I_n\}$ which drive the system are mutually independent, independent of $X_0$, and each consisting of independently and identically distributed random variable. This defines a Markov chain $\{X_n\}$ with nonnegative integer states, originally introduced and applied by Smoluchowski (1916) in studying the fluctuations in the number of particles contained in a small volume in connection with the second law of thermodynamics [**?**]. Since then, the process has been applied extensively in biological, sociological and physical branching phenomena [**?**],[**?**], [**?**], [**?**] and [**?**]. A good review of early work can be found in [**?**].

In the vernacular of branching processes, $X_n$ is the size of the $n$th generation of a population, $Y_{n,1}, ..., Y_{n,X_{n-1}}$ are the offspring of the $(n-1)$st generation, and $I_n$ is the contribution of immigration to the $n$th generation, that is, the number of immigrants at time $n$. An important role in the behavior of $\{X_n\}$ is played by the mean $m = \mathrm{E}[\mathrm{Y_{n,i}}]$ of the offspring distribution, where the cases $m < 1, m = 1, m > 1$, are referred to as subcritical, critical, and supercritical, respectively. In the subcritical case $\{X_n\}$ has a limiting stationary distribution, while in the supercritical case $\{X_n\}$ explodes at an exponential rate. In the critical case the process is either null recurrent or transient.

The process (1) admits a useful autoregressive representation as follows. Let $\lambda = \mathrm{E}[I_n]$, and let $\mathcal{F}_n$ be generated by the past information $X_0, X_1, X_2, \cdots, X_n$. Then $\mathrm{E}[X_n \mid \mathcal{F}_{n-1}] = mX_{n-1} + \lambda$. Therefore, with

$\epsilon_n \equiv X_n - \mathrm{E}[X_n|\mathcal{F}_{n-1}]$, the stochastic equation (1) is transformed into a stochastic regression model,

$$X_n = mX_{n-1} + \lambda + \epsilon_n, \qquad n = 1, 2, 3, \cdots, \tag{2}$$

where the noise $\{\epsilon_n\}$ is a *martingale difference*), that is, $\{\epsilon_n\}$ is $\mathcal{F}_n$–measurable and $\mathrm{E}[\epsilon_\mathrm{n} \mid \mathcal{F}_{\mathrm{n}-1}] = 0$. This implies that $\mathrm{E}[\epsilon_n\epsilon_k] = 0, n \neq k$, and that sums in terms of $\{\epsilon_n\}$ tend to be normally distributed under fairly general conditions. Another fact is that $\mathrm{E}[\epsilon_n^2 \mid \mathcal{F}_{n-1}] = \mathrm{Var}[Y_{n,i}]X_{n-1} + \mathrm{Var}[I_n]$ is unbounded as $X_n$ increases.

As suggested by (2), the least squares estimators for $m, \lambda$ are obtained by minimizing,

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (X_i - mX_{i-1} - \lambda)^2,$$

and are given by

$$\begin{aligned} \tilde{m} &= \frac{\sum X_i \sum X_{i-1} - n \sum X_i X_{i-1}}{(\sum X_{i-1})^2 - n \sum X_{i-1}^2} \\ \tilde{\lambda} &= \frac{\sum X_{i-1} X_i \sum X_{i-1} - \sum X_{i-1}^2 \sum X_i}{(\sum X_{i-1})^2 - n \sum X_{i-1}^2} \end{aligned}$$

where the summation limits are from $i = 1$ to $i = n$. It turns out that $\tilde{m}$ is consistent in all three cases, while $\tilde{\lambda}$ is not consistent in the critical and supercritical cases.

Improved estimators are obtained by weighted least squares. We write (2) as

$$\frac{X_n}{\sqrt{X_{n-1}+1}} = m\sqrt{X_{n-1}+1} + \frac{(\lambda - m)}{\sqrt{X_{n-1}+1}} + \frac{\epsilon_n}{\sqrt{X_{n-1}+1}}, \tag{3}$$

and estimate $m$ and $\lambda - m$ by minimizing $\sum \delta_n^2$ where $\delta_n = \epsilon_n/\sqrt{X_{n-1}+1}$ to obtain [**?**],

$$\hat{m} = \frac{\sum X_i \sum \frac{1}{X_{i-1}+1} - n \sum \frac{X_i}{X_{i-1}+1}}{\sum (X_{i-1}+1) \sum \frac{1}{X_{i-1}+1} - n^2} \tag{4}$$

$$\hat{\lambda} = \frac{\sum X_{i-1} \sum \frac{X_i}{X_{i-1}+1} - \sum X_i \sum \frac{X_{i-1}}{X_{i-1}+1}}{\sum (X_{i-1}+1) \sum \frac{1}{X_{i-1}+1} - n^2} \tag{5}$$

where again the summation limits are from 1 to $n$.

Then for $0 < m < \infty$, $\hat{m} \to m$ in probability, that is $\hat{m}$ is consistent in all cases, provided $m > 0$, and furthermore, the limiting distribution of $\hat{m}$ is normal in noncritical cases and nonnormal in the critical case. On the other hand, $\hat{\lambda}$ is consistent for $m \leq 1$, but not for $m > 1$, and is asymptotically normal when $m < 1$ or $m = 1$ and $2\lambda > \text{Var}[Y_{n,i}]$ [**?**],[**?**].

### 1.1.1 A Stochastic Model for Rain Rate

There is ample evidence that certain rain characteristics tend to be approximately lognormally distributed in the sense of statistical goodness of fit, pertaining in particular to rainfall amounts and rates under some conditions. See [**?**] and the rainfall references cited in [**?**]. Insight into this problem can be gained from the stochastic process (1) and its manifestation (2) using a heuristic argument as follows [**?**].

Conditional on rain, suppose we observe at discrete time points a rain element or volume in space containing droplets of water having the following dynamics. At time $n-1$ some droplets give rise to a new generation of droplets through a complicated physical process, some droplets leave the volume while new ones, called immigrants, arrive to join the droplets of the new generation. It is really a process of replacement and immigration: each droplet in the volume is replaced by a nonnegative number of droplets where zero can be interpreted as complete departure from the volume, and the totality of these plus the immigrants give rise to a new generation of droplets at the new time step. This process can be described by (1), and switching now to (2), $X_n$ may be interpreted as rain rate, since a multiplication by a constant throughout the equation leaves the process essentially intact.

Under the the continuity assumption

$$|X_n - X_{n-1}| << X_{n-1}$$

and conditional on rain, $\delta_n \equiv (X_n - X_{n-1})/X_{n-1}$ is small, and

$$X_n = (1+\delta_n)(1+\delta_{n-1}) \cdots (1+\delta_1)X_0,$$

or for sufficiently small $\delta_i$,

$$\log(X_n/X_0) \approx \sum_{i=1}^{n} \delta_i$$

Thus, we arrive at the intriguing equation,

$$\log(X_n/X_0) + \sum_{i=1}^{n}[(1-m) - \lambda/X_{i-1}] \approx \sum_{i=1}^{n} \epsilon_i/X_{i-1}.$$

Table 1: Pairs of estimates $(\hat{m}, \hat{\lambda})$ for $4 \times 4 \quad km^2$ and $40 \times 40 \quad km^2$ from 20 different time series of area average rain rate. Source: [**?**].

| $4 \times 4 \quad km^2$ | | | | |
|---|---|---|---|---|
| 0.93, 0.40 | 0.94, 0.38 | 0.94, 0.41 | 0.93, 0.50 | 0.90, 0.68 |
| 0.85, 0.51 | 0.91, 0.40 | 0.93, 0.37 | 0.94, 0.38 | 0.94, 0.42 |
| 0.88, 0.34 | 0.92, 0.32 | 0.95, 0.22 | 0.94, 0.34 | 0.90, 0.61 |
| 0.88, 0.38 | 0.89, 0.39 | 0.94, 0.21 | 0.91, 0.37 | 0.92, 0.50 |
| $40 \times 40 \quad km^2$ | | | | |
| 0.98, 0.05 | 0.97, 0.05 | 0.92, 0.09 | 0.97, 0.05 | 0.96, 0.07 |
| 0.98, 0.07 | 0.99, 0.05 | 0.98, 0.05 | 0.98, 0.07 | 0.98, 0.08 |
| 0.98, 0.08 | 0.99, 0.06 | 0.99, 0.06 | 0.99, 0.06 | 0.99, 0.07 |
| 0.98, 0.08 | 0.98, 0.08 | 0.99, 0.06 | 0.99, 0.08 | 0.99, 0.07 |

So, if there is any hope of seeing lognormality here, at the very least $m$ should be close to 1 and $\lambda$ should be close to 0. This however is verifiable from data by appealing to the estimators (4) and (5).

The Global Atmospheric Research Program's (GARP) Atlantic Tropical Experiment (GATE) was conducted in the summer of 1974 in the eastern Atlantic off the coast of west Africa. During roughly three triweekly periods or phases, detailed rainfall measurements were obtained from precipitation radars on an array of research vessels over a large area of about 400 km in diameter every 15 minutes. The GATE data set consists of a collection of radar reflectivity snapshots which were then converted into rain rates binned into $4 \times 4 \quad km^2$ pixels. For technical details see [**?**].

Twenty time series of length 1716 each of rain rate for individual $4 \times 4 \quad km^2$ pixels, and then of rain rate averaged over larger $40 \times 40 \quad km^2$ pixels have been extracted from the first phase of GATE and the parameters $m, \lambda$ were estimated by the weighted least squares estimators (4) and (5). The results reported in Table 1 show that for area average rain rate obtained from the larger $40 \times 40 \quad km^2$ pixels, $\hat{m}$ tends to be closer to 1 and $\hat{\lambda}$ tends to be closer to 0, than the same quantities obtained from $4 \times 4 \quad km^2$ pixels. This trend is seen very well from Figure as the pixel size increases all the way to $400 \times 400 \quad km^2$. This result suggests that, conditional on rain, a lognormal fit is apparently more appropriate for rain rate averaged over a large area.

[width=5cm, angle=-90]OtherINAR125.eps [width=5cm, angle=-90]OtherINAR175.eps

Figure 1: The monotone increase in $\hat{m}$ (Curve a) and the monotone decrease in $\hat{\lambda}$ (Curve b) as a function of the square root of the area. Source: [**?**].

## 1.2 Integer Autoregressive Models of Order 1

Although the integer autoregressive model of order 1, INAR(1), is a special case of the branching process with immigration (1), it deserves a special consideration due to the *thinning* operation or calculus. The calculus of thinning operators provides further insight into the probabilistic structure of branching with immigration by using the simple device of sums of a random number of Bernoulli random variables.

The thinning operator is defined as follows [**?**, p. 85],[**?**].

**Definition 1.1** Suppose that $X$ is a non–negative integer random variable and let $\alpha \in [0, 1]$. Then, the thinning operator, denoted by $\circ$, is defined as

$$\alpha \circ X = \sum_{i=1}^{X} Y_i,$$

where $\{Y_i\}$ is a sequence of independent and identically distributed Bernoulli random variables–independent of $X$–with success probability $\alpha$. The sequence $\{Y_i\}$ is termed a counting series.

The random variable $\alpha \circ X$ counts the number of successes in a random number of Bernoulli trials where the probability of success $\alpha$ remains constant throughout the experiment so that given $X$, $\alpha \circ X$ is a binomial random variable with parameters $X$ and $\alpha$.

It is easy to see that $0 \circ X = 0$ and $1 \circ X = X$. In addition, the following properties hold.

$$\begin{aligned}
\beta \circ (\alpha \circ X) &\sim (\beta\alpha) \circ X && (6)\\
\mathrm{E}\left[\alpha \circ X \mid X\right] &= \alpha X && (7)\\
\mathrm{E}\left[\alpha \circ X\right] &= \alpha \mathrm{E}\left[X\right] && (8)\\
\mathrm{Var}\left[\alpha \circ X \mid X\right] &= \alpha(1-\alpha)X && (9)\\
\mathrm{Var}\left[\alpha \circ X\right] &= \alpha^2 \mathrm{Var}\left[X\right] + \alpha(1-\alpha)\mathrm{E}\left[X\right] && (10)
\end{aligned}$$

where equation (6) implies equality of distributions.

The notion of binomial thinning is extended to multinomial thinning. Let $\alpha_1, \ldots, \alpha_p$ be positive constants such that $\sum_{i=1}^{p} \alpha_i < 1$. Then the conditional distribution of the vector $(\alpha_1 \circ X, \ldots, \alpha_p \circ X)'$ given $X$, is multinomial with parameters $X$ and $(\alpha_1, \ldots, \alpha_p)$.

Integer autoregressive models imitate the structure of the common autoregressive process, discussed in Appendix... and more throughly in [**?**], in the sense that the thinning operation is applied instead of scalar multiplication [**?**], [**?**], [**?**], [**?**], [**?**].

Let $\alpha \in [0, 1]$ and let $\{\epsilon_t\}$ be a sequence of independent and identically distributed nonnegative integer valued random variables with $\mathrm{E}[\epsilon_t] = \mu$ and $\mathrm{Var}[\epsilon_t] = \sigma^2$. The INAR(1) process $\{X_t\}$, $t = 1, \ldots, N$ is defined by the equation

$$X_t = \alpha \circ X_{t-1} + \epsilon_t, \tag{11}$$

where $\alpha \circ X_{t-1}$ is the sum of $X_{t-1}$ Bernoulli random variables all of which are independent of $X_{t-1}$ (recall Definition 1.1). It should be noted that the Bernoulli variables used in $\alpha \circ X_{t-1}$ are independent of those used in $\alpha \circ X_{t-2}$, etc. Clearly, (11) is a special case of (1).

Figure 2 features realizations of 200 observations from the INAR(1) model for different values of $\alpha$ while the innovation variable $\epsilon_t$ has the Poisson distribution with mean equal to 1. Apparently as $\alpha$ grows, the process tends to be less oscillatory, a fact that is confirmed theoretically by the autocorrelation function (16) of INAR(1).

Distributional including second order properties of the INAR(1) process can be studied by expressing $X_t$ in terms of present and past values of $\epsilon_t$. By repeated substitutions and property (6) we obtain a "moving average" representation,

$$\begin{aligned} X_t &= \alpha \circ X_{t-1} + \epsilon_t \\ &= \alpha \circ (\alpha \circ X_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \alpha^2 \circ X_{t-2} + \alpha \circ \epsilon_{t-1} + \epsilon_t \\ &= \ldots \\ &= \sum_{j=0}^{\infty} \alpha^j \circ \epsilon_{t-j}. \end{aligned} \tag{12}$$

An important consequence of the representation (12) is that for $\alpha \in (0, 1)$ the dependence of $\{X_t\}$ on the sequence $\{\epsilon_t\}$ decays exponentially as $t$ grows.

The mean and variance of the INAR(1) are given by,

$$\mathrm{E}[X_t] = \alpha \mathrm{E}[X_{t-1}] + \mu$$

[width=5cm, angle=-90]OtherINAR.1.25.eps [width=5cm, angle=-90]OtherINAR.1.75.eps

Figure 2: Typical realizations of 200 observations from the INAR(1) model (11) for different values of $\alpha$. Here $\epsilon_t$ Poisson with mean equal to 1. (a) $\alpha = 0.25$. (b) $\alpha = 0.50$. (c) $\alpha = 0.75$. (d) $\alpha = 0.90$.

$$= \alpha^t \mathrm{E}[X_0] + \mu \sum_{j=0}^{t-1} \alpha^j, \tag{13}$$

$$\begin{aligned} \mathrm{Var}[X_t] &= \alpha^2 \mathrm{Var}[X_{t-1}] + \alpha(1-\alpha)\mathrm{E}[X_{t-1}] + \sigma^2 \\ &= \alpha^{2t}\mathrm{Var}[X_0] + (1-\alpha)\sum_{j=1}^{t} \alpha^{2j-1}\mathrm{E}[X_{t-j}] \\ &+ \sigma^2 \sum_{j=1}^{t} \alpha^{2(j-1)}. \end{aligned} \tag{14}$$

As $t$ grows, $\mathrm{E}[X_t] = \mu/(1-\alpha)$, $\mathrm{Var}[X_t] = (\alpha\mu + \sigma^2)/(1-\alpha^2)$, and the autocovariance function evaluated at lag $k$, $c(k)$, is given by

$$c(k) \equiv \mathrm{Cov}[X_t, X_{t-k}] = \alpha^k c_0. \tag{15}$$

Consequently, the autocorrelation function, $\rho(k)$, is

$$\rho(k) = \frac{c(k)}{c(0)} = \alpha^k, \tag{16}$$

so that $\rho(k)$ decays exponentially with the lag $k$ as in AR(1), but unlike the autocorrelation of a stationary AR(1) process, it is always positive for $\alpha \in (0, 1)$.

Under suitable conditions, it can be shown that $X_t$ has a discrete self–decomposable distribution. This in turn implies unimodality properties and characterization of the distribution of $X_t$ through $\epsilon_t$. For example, $X_t$ follows the Poisson distribution if and only if $\epsilon_t$ follows the Poisson distribution [**?**].

### 1.2.1 Poisson INAR(1)

An important special case is that of Poisson INAR(1) [**?**] and [**?**]. That is, $X_t = \alpha \circ X_{t-1} + \epsilon_t$, with $\{\epsilon_t\}$ a sequence of independent and identically distributed Poisson random variables with mean $\mu$. Then $\rho(k) = \alpha^k$,

$$\mathrm{E}[X_t \mid X_{t-1}] = \alpha X_{t-1} + \mu,$$

and

$$\mathrm{Var}\left[X_t \mid X_{t-1}\right] = \alpha(1-\alpha)X_{t-1} + \mu.$$

The conditional distribution of $X_t$ given $X_{t-1}$ is

$$\begin{aligned} p(y \mid x) &= \mathrm{P}\left[X_t = y \mid X_{t-1} = x\right] \\ &= x! \exp(-\mu) \sum_{k=0}^{m} \frac{\alpha^k (1-\alpha)^{x-k} \mu^{y-k}}{k!(x-k)!(y-k)!}, \quad y = 0, 1, \ldots, \end{aligned}$$

where $m = \min(x, y)$. In this case $\{X_t\}$ is a reversible Markov process with transition matrix specified by $p(y \mid x)$, with nonnegative integers $x, y$.

### 1.2.2 Vector INAR(1)

To define the vector INAR(1) process recall the notion of multinomial thinning. That is, suppose $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_p)'$ is a vector of nonnegative components whose sum does not exceed one. Then conditional on $X$, the random vector $\boldsymbol{\alpha} \circ X = (\alpha_1 \circ X, \ldots \alpha_p \circ X)'$ has a multinomial distribution with parameters $X$ and $(\alpha_1, \ldots \alpha_p)'$. For a $p \times p$ matrix $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$, with $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ satisfying the same conditions as those of $\boldsymbol{\alpha}$, and a $p$–dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)'$, define

$$\mathbf{A} \circ \mathbf{X} = \sum_{i=1}^{p} \boldsymbol{\alpha}_i \circ X_i,$$

where each multinomial thinning operator is performed independently. Then, the vector INAR(1) process $\{\mathbf{X}_t\}$ is determined by the stationary solution of

$$\mathbf{X}_t = \mathbf{A} \circ \mathbf{X}_{t-1} + \mathbf{E}_t.$$

where $\{\mathbf{E}_t\}$ is a sequence of independent and identically distributed random vectors.

## 1.3 Estimation for INAR(1) process

Estimation in INAR(1) means estimation in the branching process with immigration in the subcritical case, and this has already been discussed earlier. Still it is interesting to note a few facts regarding estimation in the Poisson INAR(1).

Estimation procedures for the parameters $\alpha$ and $\mu$ of the INAR(1) model (11) assuming that the sequence $\{\epsilon_t\}$ follows the Poisson distribution has

been discussed in [**?**]. Under the Poisson assumption, $\sigma^2 = \mu$ and, equation (15) yields a method of moments estimator for $\alpha$ given by

$$\hat{\alpha} = \frac{\sum_{t=0}^{N-1}(X_t - \bar{X})(X_{t+1} - \bar{X})}{\sum_{t=0}^{N}(X_t - \bar{X})^2},$$

while $\mu$ can be estimated by

$$\hat{\mu} = \frac{\sum_{t=1}^{N} \hat{\epsilon}_t}{N},$$

where $\hat{\epsilon}_t = X_t - \hat{\alpha} X_{t-1}$, for $t = 1, \ldots, N$.

An alternative estimation method is that of conditional least squares. Upon noticing that

$$\mathrm{E}\left[X_t \mid X_{t-1}\right] = \alpha X_{t-1} + \mu,$$

the conditional least squares of the parameters $\alpha$ and $\mu$ are those values that minimize

$$\sum_{t=1}^{N} \left(X_t - \alpha X_{t-1} - \mu\right)^2.$$

Asymptotics properties of the resulting estimators (see Problem 4) are deduced by using results of [**?**], [**?**]. The two methods, moments and least squares, do not require full distributional assumptions.

Maximum likelihood estimation, which requires a full distributional assumption about the innovations, has been suggested in [**?**]. Under the Poisson assumption, the likelihood function of a time series of $(N+1)$ observations from the INAR(1) process is

$$\left(\prod_{t=1}^{N} P_t(X_t)\right) \frac{\left(\mu/(1-\alpha)\right)^{X_0}}{X_0!} \exp\left(-\mu/(1-\alpha)\right), \tag{17}$$

where for t=1,. . . ,N,

$$P_t(y) = \exp(-\mu) \sum_{i=0}^{\min(X_t, X_{t-1})} \frac{\mu^{y-i}}{(y-i)!} \binom{X_{t-1}}{i} \alpha^i (1-\alpha)^{X_{t-1}-i}, \tag{18}$$

Differentiation of (17) gives the unconditional maximum likelihood estimates, while differentiation of $\left(\prod_{t=1}^{N} P_t(X_t)\right)$ yields the conditional maximum likelihood estimates of $\alpha$ and $\mu$ given $X_0$.

## 1.4 Integer Autoregressive Models of Order $p$

A straightforward extension to the integer autoregressive model of order 2, INAR(2), gives

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \epsilon_t,$$

and more generally the $p$th order model, INAR($p$), is defined as

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \epsilon_t \tag{19}$$

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed non negative integer valued random variables with mean $\mu$ and variance $\sigma^2$, and for stability of the process it is required that $\sum_i \alpha_i < 1$ [**?**].

The conditional distribution of $(\alpha_1 \circ X_t, \ldots, \alpha_p \circ X_t)'$ given $X_t$ is multinomial and is independent of the past of the process. In other words given $X_t$, $\alpha_i \circ X_t$ is independent of $X_{t-k}$ and $\alpha_j \circ X_{t-k}$ for $i, j = 1, 2, \ldots, p$ and $k > 0$. The meaning of expression (19) generalizes that of (11) in the sense that the total size of the population at time $t$ is equal to the number of offsprings of the last $p$ generations $\alpha_i \circ X_{t-i}$ plus the immigration process.

Sufficient conditions for the process $\{X_t\}$ to have a limiting distribution are that the roots of

$$\lambda^p - \alpha_1 \lambda^{p-1} - \ldots - \alpha_{p-1}\lambda - \alpha_p = 0, \;\; \alpha_p \neq 0, \tag{20}$$

are less than in 1 in absolute value and $\sum_{j=0}^{\infty}(j+1)^{-1} p_j < \infty$, where $p_j = \sum_{k=j+1}^{\infty} \mathrm{P}[\epsilon_1 = k]$ [**?**]. Moreover, as $t$ grows the mean of the process is

$$\mu_x = \mathrm{E}[X_t] = \frac{\mu}{1 - \sum_{i=1}^{p} \alpha_i} \tag{21}$$

while the autocovariance function satisfies

$$c(k) = \sum_{i=1}^{p} \alpha_i c(k-i) + \sum_{i=k+1}^{p} v(k-i, \alpha_i) + \delta_k(0)\sigma^2, \tag{22}$$

with

$$v(-k, \alpha_i) = \sum_{j=1}^{k-1} \alpha_j v(j-k, \alpha_i) + \alpha_i \left(\delta_i(k) - \alpha_k\right) \mu_x \tag{23}$$

and $v(k-i, \alpha_i)$ is determined from (23) for $k < i$, and $v(k-i, \alpha_i) = 0$ for $k \geq i$. Here $\delta_k(0) = 1$ if $k = 0$. Equation (22) points out that the autocovariance function of the INAR($p$) process has a form similar to that of

a Gaussian ARMA$(p, p-1)$ process due to dependence in $\alpha_i \circ X_{t-i}, i = 1, ..., p$ appearing in different times (see [**?**] and also the discussion in [**?**].) Existence and generalizations of INAR$(p)$ are studied in [**?**], [**?**] and [**?**], while unifying work based on convolution is presented in [**?**].

### 1.5 Regression Analysis of Integer Autoregressive Models

The INAR(1) model (11) has been extended by including explanatory variables. Following [**?**], assume that the observed response process $\{Y_t\}$ is a realization of

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t$$

with $\{\epsilon_t\}$ a sequence of Poisson random variables with mean $\exp(\boldsymbol{\beta}'\mathbf{X}_t)$, where $\{\mathbf{X}_t\}$ is a covariate process. In addition, assume that

$$\alpha = \frac{1}{1 + \exp(-\gamma)}$$

so that $0 < \alpha < 1$. Then,

$$\mathrm{E}\left[Y_t \mid \mathbf{X}_t, Y_{t-1}\right] = \left(\frac{1}{1 + \exp(-\gamma)}\right) Y_{t-1} + \exp(\boldsymbol{\beta}'\mathbf{X}_t),$$

and

$$\mathrm{Var}\left[Y_t \mid \mathbf{X}_t, Y_{t-1}\right] = \left(\frac{\exp(-\gamma)}{(1 + \exp(-\gamma))^2}\right) Y_{t-1} + \exp(\boldsymbol{\beta}'\mathbf{X}_t)$$

by employing (7) and (9). Estimation is based on minimizing with respect to $(\boldsymbol{\beta}, \gamma)$ the unweighted sum of squares

$$\sum_t \left\{ Y_t - \left(\frac{1}{1 + \exp(-\gamma)}\right) Y_{t-1} - \exp(\boldsymbol{\beta}'\mathbf{X}_t) \right\}^2.$$

The corresponding standard errors need to be adjusted though to allow for heteroscedasticity. An alternative method is that of weighted least squares which leads to more efficient estimators. The above specification can be extended by introducing the dynamic parameterization $\alpha_t = 1/(1+\exp(-\boldsymbol{\gamma}'\mathbf{Z}_t)$ with $\mathbf{Z}_t$ another covariate process. Again, least squares or weighted least squares can be used in the estimation of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

### 1.6 Integer Moving Average Models

In the spirit of the INAR$(p)$ process, the integer moving average model of order $q$, abbreviated INMA$(q)$, id defined by the equation [**?**] and [**?**],

$$X_t = \beta_0 \circ \epsilon_t + \beta_1 \circ \epsilon_{t-1} + \ldots + \beta_q \circ \epsilon_{t-q} \tag{24}$$

Figure 3: Typical realizations of 200 observations from the INMA(1) model (11) for different values of $\beta_1$. Here $\epsilon_t$ Poisson with mean equal to 1. (a) $\beta_1 = 0.25$. (b) $\beta_1 = 0.50$. (c) $\beta_1 = 0.75$. (d) $\beta_1 = 0.90$.

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed non negative integer–valued random variables with mean $\mu$ and variance $\sigma^2$, $\beta_0 = 1$ and assume that the $\beta_i$ belong to $[0, 1]$ and that all thinning operations are performed independently. Figure 3 illustrates typical realizations of the INMA(1) process,

$$X_t = \epsilon_t + \beta_1 \circ \epsilon_{t-1}, \tag{25}$$

where the $\epsilon_t$ follows the Poisson distribution with mean 1 and the parameter $\beta_1$ assumes the values 0.25, 0.50, 0.75 and 0.90. As $\beta_1$ grows, the process becomes less oscillatory since the degree of positive correlation between successive observations increses.

It is instructive to consider some properties of the INMA(1) model. Recall the properties of the thinning operator (8) and (10). Then it is easy to see that

$$\begin{aligned} \mathrm{E}\left[X_t\right] &= \mathrm{E}\left[\epsilon_t + \beta_1 \circ \epsilon_{t-1}\right] \\ &= (1 + \beta_1)\mu \end{aligned} \tag{26}$$

and

$$\begin{aligned} \mathrm{Var}\left[X_t\right] &= \mathrm{Var}\left[\epsilon_t + \beta_1 \circ \epsilon_{t-1}\right] \\ &= (1 + \beta_1^2)\sigma^2 + \beta_1(1 - \beta_1)\mu. \end{aligned} \tag{27}$$

By a conditional argument and using the independence of the thinning operations we also obtain

$$c(k) = \beta_1 \sigma^2 \tag{28}$$

for $k = 1$. When $k > 1$, then $\mathrm{Cov}\left[X_t, X_{t-k}\right] = 0$. The autocorrelation function of the INMA(1) process is then

$$\rho(k) = \frac{c(k)}{c(0)} = \frac{\sigma^2 \beta_1}{(1 + \beta_1^2)\sigma^2 + \beta_1(1 - \beta_1)\mu}, \tag{29}$$

for $k = 1$, and is equal to 0 otherwise, similar to the autocorrelation function of the standard MA(1) model. For Poisson distributed errors (29) reduces to [**?**],)

$$\rho(k) = \begin{cases} \beta_1/(1 + \beta_1), & k = 1, \\ 0, & k > 1. \end{cases}$$

It is not hard to generalize formulae (26)–(29) for the case of INMA($q$) model (24). We have,

$$\mathrm{E}\left[X_t\right] = \mu \sum_{i=0}^{q} \beta_i, \tag{30}$$

$$\mathrm{Var}\left[X_t\right] = \sigma^2 \left( \sum_{i=0}^{q} \beta_i^2 \right) + \mu \left( \sum_{i=0}^{q} \beta_i (1 - \beta_i) \right), \tag{31}$$

$$c(k) = \sigma^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k} + \mu \sum_{i=0}^{q-k} \beta_i \left( \beta_k - \beta_{i+k} \right), \tag{32}$$

for $k = 1, 2, \ldots, q$ and 0 otherwise,

$$\rho(k) = \frac{c(k)}{c(0)} = \frac{\sigma^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k} + \mu \sum_{i=0}^{q-k} \beta_i \left( \beta_k - \beta_{i+k} \right)}{\sigma^2 \left( \sum_{i=0}^{q} \beta_i^2 \right) + \mu \left( \sum_{i=0}^{q} \beta_i (1 - \beta_i) \right)} \tag{33}$$

for $k = 1, 2, \ldots, q$ and 0 otherwise. Problem 7 asks the reader to verify all these moment calculations. An extension of the INMA($q$) model and the accompanying estimation problem are discussed in [**?**].

## 1.7 Extensions and Modifications

Among the early works on autoregressive processes with prescribed marginals is that of [**?**] where it is shown that there exists an innovation sequence $\{\epsilon_t\}$ such that $X_n$ from the AR(1) sequence $X_n = \rho X_{n-1} + \epsilon_n, \quad n = 0, \pm 1, \pm 2, ...,$ has a gamma distribution for $0 \leq \rho < 1$. In particular there is a sequence $\{\epsilon_t\}$ of i.i.d. random variables such that the $X_n$ have an exponential distribution and the resulting process is called exponential autoregressive process of order 1, EAR(1). Related works on time series having prescribed marginals such as exponential MA, EMA(1), and exponential ARMA, EARMA(p,q), includes that of [**?**], [**?**], [**?**] [**?**].

In connection with AR(1) having gamma marginals we must mention that under some conditions, in the branching process with immigration (1), which we know admits the AR(1) representation (2), $(1-m)X(m)$ converges in distribution as $m \to 1_-$ to a random variable with a gamma distribution, where $X(m)$ has the limiting stationary distribution of $X_n$ [**?**].

The INAR($p$) and INMA($q$) models can be combined to form the integer autoregressive moving average model of order $(p, q)$, INARMA(p,q). The distributional properties of INAR, INMA, and INARMA depend on the assumptions regarding the innovation process. Besides the Poisson assumption, also binomial, geometric and negative binomial $\epsilon_t$'s have been proposed

[**?**], [**?**], [**?**]. A representative example is the autoregressive negative binomial model, INAR(1)-NB,

$$Y_t = \Pi \circ Y_{t-1} + I_t,$$

where $\Pi$ is a beta random variable and $Y_0, I_t$ are independent negative binomial random variables. The INAR(1)-NB model has been applied to personality factors and emotion experiences in [**?**]. In the same vein, the definition of the binomial thinning operator is extended to the so called hypergeometric thinning operator for the construction of binomial and generalized Poisson models in [**?**], [**?**].

A generalization of binomial thinning is given in [**?**],

$$X_t = A_t(X_{t-1}; \alpha) + \epsilon_t, \quad t = 1, 2, ....,$$

where $A_t$ is a random transformation, and $A_t(X_{t-1}; \alpha)$ and $\epsilon_t$ are independent. Based on this general thinning, a class of stationary moving average processes with margins in the class of infinitely divisible exponential dispersion models was introduced in [**?**].

Let $\{Y_t\}$, $t = 0, 1, ...,$ be a time-homogeneous first-order Markov process on $\mathcal{Y} \subseteq R$. Another recent generalization is that of the first order conditional linear autoregressive process, CLAR(1),

$$m(Y_{t-1}) = \phi Y_{t-1} + \lambda$$

where $m(Y_{t-1}) = \mathrm{E}[Y_t|Y_{t-1}]$, and $\phi, \lambda$ are real numbers. The CLAR(1) class includes many of the non-Gaussian AR(1) models proposed in the literature and allows various generalizations of previous results [**?**]. Interestingly, For $|\phi| < 1$, the autocorrelation $\rho(k) = \phi^k$, $k = 1, 2, ...,$ as in other first order autoregressive processes including the branching process with immigration (1).

## 2 Discrete Autoregressive Moving Average Models

An early attempt to introduce autoregressive and moving average models for discrete-valued time series data was made through the introduction in [**?**] and [**?**] of the so called discrete autoregressive moving average models, DARMA, defined as follows. Let $\{Y_t\}$ be a sequence of independent discrete random variables each having an arbitrary distribution $\pi$ such as Poisson or geometric. By definition, $\mathrm{P}[Y_t = i] = \pi_i$ for $i$ in some countable set. Let $\{U_t\}$ and $\{V_t\}$ be independent sequence of Bernoulli random variables such

that $\mathrm{P}[U_t = 1] = \beta$, $0 \leq \beta \leq 1$, and $\mathrm{P}[V_t = 1] = \rho$, $0 \leq \rho < 1$, and denote by $\{S_t\}$ a sequence of independent and identically distributed random variables supported in $\{0, 1, \ldots, N\}$ with distribution $F$.

We call the process $\{X_t\}$ discrete autoregressive moving average process of order $(1, N+1)$, DARMA(1,N+1), if

$$X_t = U_t Y_{t-S_t} + (1 - U_t) A_{t-(N+1)}, \quad t = 1, 2, \ldots \tag{34}$$

where

$$A_t = V_t A_{t-1} + (1 - V_t) Y_t, \quad , t = -N, -N+1, \ldots. \tag{35}$$

The last expression (35) is termed as discrete autoregressive process of order 1, where $A_t$ equals $A_{t-1}$ with probability $\rho$ or $A_t$ coincides with $Y_t$ with probability $1-\rho$. Similarly, $X_t$ is equal to one of the $Y_t, Y_{t-1}, \ldots, Y_{t-N}$ with probability $\beta$ or $X_t = A_{t-(N+1)}$ with probability $1 - \beta$.

There are several worth mentioning properties of the DARMA$(1,N+1)$ process discussed in detail in [**?**]. When the process starts with $A_{-(N+1)}$ having the distribution $\pi$ independently of $Y_t$ for $t \geq -N$, $\{U_t\}$, $\{V_t\}$, and $\{S_t\}$, then $\{X_t\}$, $t = 1, 2, \ldots$ is stationary with marginal distribution $\pi$ with an autocorrelation structure that depends on $\rho$, $\beta$ and $F$,

$$\begin{aligned} \rho(k) &= \beta^2 \sum_{j=0}^{N-k} F(j) F(j+k) \\ &+ \beta(1-\beta) \left( (1-\rho) \rho^{-(N+1-k)} \sum_{j=N+1-k}^{N} \rho^j F(j) \right) \\ &+ (1-\beta)^2 \rho^k, \end{aligned}$$

for $1 \leq k \leq N$, and

$$\rho(k) = \beta(1-\beta) \rho^{k-(N+1)} \sum_{j=0}^{N} \rho^j (1-\rho) F(j) + (1-\beta)^2 \rho^k,$$

for $k > N$.

It can be shown that $\{X_t\}$ is in general not Markovian but it is so for $\beta = 0$, and that in general it is not time reversible in the sense that $\{X_1, ..., X_k\}$ does not in general have the same distribution as $\{X_{-k}, ..., X_{-1}\}$ For more on the DARMA model and its asymptotic properties regarding estimates of moments, percentiles and quantiles, and a goodness of fit test for the marginal distribution $\pi$ see [**?**]. Applications of DARMA models related to meteorological problems are featured in [**?**], [**?**], [**?**], and [**?**].

An extension of (34) and (35) has been considered in [**?**] where the enlarged discrete autoregressive moving average model of order $(p, N+1)$ is considered in addition to other models. Keeping the same notation as in (34) and (35), the DARMA$(p, N+1)$ model is given by

$$X_t = U_t Y_{t-S_t} + (1-U_t)A_{t-(N+1)}, \quad t = 1, 2, \ldots$$

and

$$A_t = V_t A_{t-D_t} + (1-V_t)Y_t, \quad , t = -N, -N+1, \ldots,$$

where $\{D_t\}$ is a sequence of independent identically distributed random variables from a distribution $G$, taking values $1, 2, \ldots, p$. It can be shown that the correlation structure of this process is similar to that of the ARMA$(p, q)$ model, and that some ad hoc nonparametric estimation methods perform reasonably well compared with maximum likelihood estimators ([**?**]).

A related theory for bivariate exponential and geometric autoregressive moving average models has been developed in [**?**] and [**?**] where the concept of positive dependence is used to show that all these models consist of associated random variables.

# 3 The Mixture Transition Distribution Model

The mixture transition distribution model has been introduced in [**?**] extending previous work from [**?**] as a parsimonious approach for the analysis of higher order Markov chains.

Suppose the process $\{X_t\}$ takes values in $\{1, 2, \ldots, m\}$ satisfying

$$\mathrm{P}\left[X_t \mid X_{t-1}, X_{t-2}, \ldots\right] = \mathrm{P}\left[X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}\right], \tag{36}$$

for some $p \geq 2$. As indicated in Chapter **??**, as $p$ and $m$ grow the number of parameters increases exponentially according to the formula $m^p(m-1)$. The mixture transition distribution model–abbreviated MTD–bypasses this problem by specifying the conditional probability of observing $X_t = i_0$ given the past as a linear combination of contributions from $X_{t-1}, \ldots, X_{t-p}$. More precisely it is assumed that

$$\begin{aligned} \mathrm{P}\left[X_t = i_0 \mid X_{t-1} = i_1, \ldots, X_{t-p} = i_p\right] &= \sum_{j=1}^{p} \lambda_j \mathrm{P}\left[X_t = i_0 \mid X_{t-j} = i_j\right] \\ &= \sum_{j=1}^{p} \lambda_j q_{i_j i_0} \end{aligned} \tag{37}$$

where $i_0, \ldots, i_p$ belong to $\{1, 2, \ldots, m\}$, $q_{i_j i_0}$ are elements of the $m \times m$ transition matrix $\mathbf{Q}$ and the vector of lag parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ satisfies

$$\sum_{j=1}^{p} \lambda_j = 1, \quad \lambda_j \geq 0,$$

so that the right hand side of (37) is between 0 and 1. An alternative set of restriction for $\boldsymbol{\lambda}$ is given by [**?**].

Besides reducing considerably the number of parameters to $m(m-1) + (p-1)$, model (37) enjoys several properties. It can be shown that the limiting behavior of MTD model is the same as the full parameterized higher order Markov chain (see [**?**] and [**?**]). Furthermore, defining the $m \times m$ matrix $\mathbf{B}(k)$ whose elements are

$$b_{ij}(k) = \mathrm{P}\left[X_t = i, X_{t+k} = j\right], \quad i, j = 1, 2, \ldots, m,$$

for $k$ integer, then

$$\mathbf{B}(k) = \sum_{j=1}^{p} \lambda_j \mathbf{Q}\mathbf{B}(k-j). \tag{38}$$

As a result, if $Y_t$ is a random variable whose distribution is given by

$$\mathrm{P}\left[Y_t = i \mid X_t = j\right] = q_{ji}$$

and $\rho(k)$, $\tilde{\rho}(k)$ denote the correlations between $X_{t+k}$ and $X_t$ and between $Y_{t+k}$ and $X_t$, respectively, then

$$\rho(k) = \sum_{j=1}^{p} \lambda_j \tilde{\rho}_{k-j}.$$

That is the autocorrelation satisfies a system similar to the Yule–Walker equations. Various generalizations of the MTD model have been proposed. For example, [**?**] considers the multi–matrix mixture transition distribution model called MTDg. The MTDg model uses a different transition matrix for each lag as follows

$$\mathrm{P}\left[X_t = i_0 \mid X_{t-1} = i_1, \ldots, X_{t-p} = i_p\right] = \sum_{j=1}^{p} \lambda_j q^{(j)}_{i_j i_0}. \tag{39}$$

Model (39) is less parsimonious than model (37) in the sense that it requires $m(m-1)+1$ additional parameters for each lag. However it accommodates a dynamic relation between each lag and time period.

The work in [**?**] and more recently in[**?**] extend definition (37) to arbitrary state space. The spatial MTD model is investigated in [**?**] and [**?**], and the double chain Markov model is studied in [**?**].

### 3.1 Old Faithful Data Revisited

Recall example **??** regarding modeling successive eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming. The analysis, based on regression models for binary time series, suggested that a second order model fits the data reasonably well. This fact is reconfirmed by Table 2 which reports the deviance, AIC and BIC from models (36), (37) and (39) fitted to the Old Faithful data. The second column of Table 2 lists the number of estimated parameters in each model. For example, the first order Markov chain model consists only of one free parameter for the Old Faithful data since 0 is never followed by 0. Therefore the transition probability from 0 to 1 is equal to 1 and the only free parameter is the transition probability from 1 to 1. In a similar manner we obtain the rest of this column's entries. These results are based on the first 259 observations for a fair comparison with Table **??**. The first five lines of Table 2 show that there is a close agreement among the Markov chain models and the corresponding regression models pointing again to the second order Markov chain model as the best candidate for these data.

Table 2: Results from Markov chain and MTD models applied to the Old Faithful Data. $N = 259$.

| Model | Number of Parameters | $D$ | AIC | BIC |
|---|---|---|---|---|
| Independence | 1 | 331.12 | 333.12 | 336.66 |
| Markov chain of order 1 | 1 | 227.37 | 229.37 | 232.92 |
| Markov chain of order 2 | 2 | 215.52 | 219.52 | 226.61 |
| Markov chain of order 3 | 3 | 215.07 | 221.07 | 231.69 |
| Markov chain of order 4 | 5 | 213.95 | 223.95 | 241.65 |
| MTD of order 1 | 1 | 227.37 | 229.37 | 232.92 |
| MTDg of order 1 | 1 | 227.37 | 229.37 | 232.92 |
| MTDg of order 2 | 5 | 215.81 | 225.81 | 243.52 |
| MTDg of order 3 | 6 | 215.52 | 227.52 | 248.77 |
| MTDg of order 4 | 11 | 215.61 | 237.61 | 276.55 |

The estimation results for the MTD model (37) are reported only for $p = 1$, since there was no any further improvement in the deviance, AIC and BIC for higher order models. It is seen that by the AIC result, the MTDg of order 2 is somewhat preferable to the MTD model, but by the BIC criterion there is no advantage to any of the MTDg models. Among

the MTDg models (39), the AIC selects a second order MTDg model while the BIC points to a first order MTDg model. Overall, from all the cases considered, the full second order Markov chain is selected as the preferable model by both AIC and BIC.

### 3.2 Explanatory Analysis of DNA sequence data revisited

Consider example (**??**) concerning DNA sequence data of the gene BNRF1 of the Epstein–Barr virus. Table 3 reports the results of models (36), (37) and (39) and should be compared with Table **??**.

The first line of Table 3 reports results under independence, the selected model under the BIC criterion. The next four rows show the analysis based on full Markov chain modeling, that is, model (36). We see that a Markov chain of order 4 leads to the smallest AIC with 321 parameters, while the BIC selects the first order model with 12 parameters.

MTD fitting points to the first order model. Indeed, an MTD model of order 1 is simply a Markov chain of order 1. It can be seen though that higher order MTD models do not affect the fit considerably since the changes in deviance are rather small. Compared with the output of the multinomial logits model (**??**) for the DNA data, the MTD models reduce both the AIC and BIC criteria. In addition the number of parameters that need to be estimated is appreciably less than the number of parameters that need to be estimated for both the multinomial logits model and the full Markov chain. Note that, as in orders 2 and 3, the equality between the number of parameters for some models may not leave degrees of freedom for testing certain hypotheses.

The multi–lag MTDg model points to the first order Markov chain. Notice again that, as with MTD of order 1, an MTDg model of order 1 is simply a Markov chain of order 1. Regarding the fitted MTDg models, here the number of parameters becomes large compared with those of the MTD model and this leads to an increase in both the AIC and BIC values. Compared with the multinomial logits fit, the AIC and BIC values from the MTDg models are larger.

### 3.3 Soccer Forecasting Data Revisited

The final example in this section concerns the fits of models (36), (37) and (39) to the soccer forecasting data (see Section **??**). Table 4 reports the results of this analysis only for the games played in the first position. We see that the proportional odds model performs better than all the alternatives

Table 3: Results from Markov chain and MTD models applied to gene BNRF1 of the Epstein–Barr virus DNA data. $N = 996$.

| Model | Number of Parameters | $D$ | AIC | BIC |
|---|---|---|---|---|
| Independence | 3 | 2711.31 | 2717.31 | 2732.02 |
| Markov chain of order 1 | 12 | 2677.75 | 2701.75 | 2760.60 |
| Markov chain of order 2 | 48 | 2627.68 | 2723.68 | 2959.06 |
| Markov chain of order 3 | 179 | 2463.22 | 2821.22 | 3698.99 |
| Markov chain of order 4 | 321 | 1808.33 | 2450.33 | 4024.44 |
| MTD of order 1 | 12 | 2677.75 | 2701.75 | 2760.60 |
| MTD of order 2 | 13 | 2677.75 | 2703.75 | 2767.51 |
| MTD of order 3 | 13 | 2677.11 | 2703.11 | 2766.86 |
| MTD of order 4 | 14 | 2676.18 | 2704.18 | 2772.83 |
| MTDg of order 1 | 12 | 2677.75 | 2701.75 | 2760.60 |
| MTDg of order 2 | 25 | 2664.27 | 2714.27 | 2836.87 |
| MTDg of order 3 | 36 | 2647.27 | 2719.27 | 2895.80 |
| MTDg of order 4 | 46 | 2631.12 | 2723.12 | 2948.70 |

considered in the table in the sense of minimizing both the AIC and BIC. To explain this notice the relatively small number of parameters required when fitting a proportional odds model. Furthermore, the results are consistent with the previous analysis. That is, the model of independence fits the soccer data quite well leading once more to the conclusion that the soccer forecasting game is fair.

### 3.4 Estimation in MTD Moels

Estimation of the parameters $\boldsymbol{\lambda}$ and $q_{ij}$ of the mixture transition model (37) is accomplished by maximizing the log–likelihood [**?**], [**?**],

$$\sum_{i_0,\ldots,i_p=1}^{m} n_{i_0,\ldots,i_p} \log\left(\sum_{j=1}^{p} \lambda_j q_{i_j i_0}\right)$$

subject to constraints for $\boldsymbol{\lambda}$. Here $n_{i_0,\ldots,i_p}$ counts the number of sequences $\{X_t = i_0, \ldots, X_{t-p} = i_p\}$. Alternative estimation methods include the minimum $\chi^2$ estimation ([**?**]) and E–M algorithm ([**?**]). Software (MTD and GMTD)for fitting the mixture transition model as described above can be obtained from

Table 4: Results from Markov chain and MTD models applied to the Soccer Forecasting Data for the first position. $N = 289$.

| Model | Number of Parameters | $D$ | AIC | BIC |
|---|---|---|---|---|
| Independence | 2 | 562.43 | 566.43 | 573.74 |
| Markov chain of order 1 | 6 | 558.69 | 570.69 | 592.64 |
| Markov chain of order 2 | 18 | 549.21 | 585.21 | 651.08 |
| MTD of order 1 | 6 | 558.69 | 570.69 | 592.64 |
| MTD of order 2 | 7 | 558.68 | 572.68 | 598.29 |
| MTDg of order 1 | 6 | 558.69 | 570.69 | 592.64 |
| MTDg of order 2 | 12 | 557.84 | 581.84 | 625.75 |

`http://lib.stat.cmu.edu/general`. A thorough review of the mixture transition distribution model for higher order Markov chains and non Gaussian time series can be found in [**?**].

# 4 Hidden Markov Models

Hidden Markov models specify that the observed process is driven by some unobserved process which is assumed to be a Markov chain. To be more specific, assume that $\{X_t\}$, $t = 1, \ldots, N$ denotes a non–negative integer time series and suppose further that $\{A_t\}$ is an irreducible homogeneous Markov chain taking values on $\{1, 2, \ldots, m\}$ with transition probability matrix $\mathbf{Q}$. That is, the $(i, j)$-element of $\mathbf{Q}$ is

$$Q_{ij} = \mathrm{P}\left[A_t = j \mid A_{t-1} = i\right], \quad i, j = 1, 2, \ldots, m. \tag{40}$$

Set

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)' \tag{41}$$

for the stationary distribution of $\{A_t\}$–such a distribution exists because the chain is assumed irreducible. In addition, suppose that conditional on $\{A_t\}$, the random variables $\{X_t\}$, $t = 1, \ldots, N$ are mutually independent such that

$$p_{xa} = \mathrm{P}\left[X_t = x \mid A_t = a\right], \quad x = 0, 1, \ldots, \; a = 1, 2, \ldots, m. \tag{42}$$

In general, the probabilities $p_{xa}$ may depend on $t$. However for our limited exposition we prefer to drop that notation. Examples (42) include Poisson,

binomial and multinomial distributions. Hidden Markov models were introduced in [**?**], [**?**], [**?**] [**?**] and consequently found numerous applications in engineering, speech processing, genetics, econometrics, biochemistry, environmetrics and so on. We do not attempt a comprehensive study of these models and the reader is referred to the texts [**?**], [**?**] and [**?**] for more references and further information on their probabilistic properties and existing estimation methods.

To gain some insight though it is instructive to consider (42) with

$$p_{xa} = \frac{\exp(-\lambda_a)\lambda_a^x}{x!}, \tag{43}$$

that is a Poisson hidden Markov model. Equation (43) implies that given that an unobserved process is in state $a$, the observed process is Poisson with mean $\lambda_a$. It follows that

$$\mathrm{E}\left[X_t \mid A_t = a\right] = \lambda_a$$

and

$$\begin{aligned} \mathrm{E}\left[X_t\right] &= \sum_{a=1}^{m} \mathrm{E}\left[X_t \mid A_t = a\right] \mathrm{P}\left[A_t = a\right] \\ &\sum_{a=1}^{m} \lambda_a \pi_a = \boldsymbol{\lambda}'\boldsymbol{\pi} \end{aligned}$$

upon recalling (41) and defining $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)'$. Furthermore,

$$\begin{aligned} \mathrm{E}\left[X_t^2\right] &= \sum_{a=1}^{m} \mathrm{E}\left[X_t^2 \mid A_t = a\right] \mathrm{P}\left[A_t = a\right] \\ &\sum_{a=1}^{m} \left(\lambda_a^2 + \lambda_a\right) \pi_a \end{aligned}$$

so that

$$\begin{aligned} \mathrm{Var}\left[X_t\right] &= \sum_{a=1}^{m} \left(\lambda_a^2 + \lambda_a\right) \pi_a - \left(\boldsymbol{\lambda}'\boldsymbol{\pi}\right)^2 \\ &= \boldsymbol{\pi}'\boldsymbol{\Lambda}\boldsymbol{\lambda} + \boldsymbol{\lambda}'\boldsymbol{\pi} - \left(\boldsymbol{\lambda}'\boldsymbol{\pi}\right)^2, \end{aligned}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_m\right)$. The autocovariance function of the Poisson hidden Markov model is

$$\begin{aligned} c(k) &= \mathrm{E}\left[X_t X_{t+k}\right] - \mathrm{E}\left[X_t\right] \mathrm{E}\left[X_{t+k}\right] \\ &= \sum_{a=1}^{m} \sum_{b=1}^{m} \lambda_a \lambda_b \pi_a q_{ab}^{(k)} - \left(\boldsymbol{\lambda}'\boldsymbol{\pi}\right)^2, \end{aligned} \tag{44}$$

where $q_{ab}^{(k)}$ is the $(a,b)$-element of $\mathbf{Q}^k$, for $k$ positive. The last equation holds since

$$\begin{aligned}\mathrm{E}\left[X_t X_{t+k}\right] &= \sum_{a=1}^{m}\sum_{b=1}^{m}\mathrm{E}\left[X_t X_{t+k} \mid X_t = a, X_{t+k} = b\right]\mathrm{P}\left[X_t = a, X_{t+k} = b\right] \\ &= \sum_{a=1}^{m}\sum_{b=1}^{m}\lambda_a\lambda_b\pi_a q_{ab}^{(k)}.\end{aligned}$$

Thus, (44) becomes

$$c(k) = \boldsymbol{\pi}'\boldsymbol{\Lambda}\mathbf{Q}^k\boldsymbol{\lambda} - \left(\boldsymbol{\lambda}'\boldsymbol{\pi}\right)^2 \tag{45}$$

and consequently the autocorrelation function is

$$\rho(k) = \frac{\boldsymbol{\pi}'\boldsymbol{\Lambda}\mathbf{Q}^k\boldsymbol{\lambda} - \left(\boldsymbol{\lambda}'\boldsymbol{\pi}\right)^2}{\boldsymbol{\pi}'\boldsymbol{\Lambda}\boldsymbol{\lambda} + \boldsymbol{\lambda}'\boldsymbol{\pi} - \left(\boldsymbol{\lambda}'\boldsymbol{\pi}\right)^2}. \tag{46}$$

Similar results are obtained by specifying (42) to be the probability mass function of the binomial, multinomial or any other probability distribution.

From a statistical point of view, interest is focused on the estimation of model parameters based on the observed data $\{X_t\}$. In this case the unknown parameters consist of the transition probabilities (40) of the hidden Markov chain and any other parameters introduced by (42). Observe that (41) depends on (40) since $\mathbf{Q}\boldsymbol{\pi} = \boldsymbol{\pi}$. The Poisson example shows that the unknown parameters in the model are all the elements of the transition matrix plus $\lambda_1, \lambda_2, \ldots, \lambda_m$.

To estimate the parameters of a hidden Markov model, consider the likelihood of the observed data $\{X_t\}$, for $t = 1, 2, \ldots, N$

$$L = \mathrm{P}\left[X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N\right].$$

Conditioning on the event $\{A_1 = a_1, \ldots, A_N = a_N\}$, for the $a_j$'s in $\{1, 2, \ldots, m\}$ leads to

$$\begin{aligned}L &= \sum_{a_1=1}^{m}\cdots\sum_{a_N=1}^{m}\mathrm{P}\left[X_1 = x_1, \ldots, X_N = x_N \mid A_1 = a_1, \ldots, A_N = a_N\right] \\ &\qquad\qquad \times\mathrm{P}\left[A_1 = a_1, \ldots, A_N = a_N\right] \\ &= \sum_{a_1=1}^{m}\cdots\sum_{a_N=1}^{m}\left(p_{x_1 a_1}\ldots p_{x_N a_N}\right)\left(\pi_{a_1} q_{a_1 a_2}\ldots q_{a_{N-1}a_N}\right) \\ &= \sum_{a_1=1}^{m}\cdots\sum_{a_N=1}^{m}\pi_{a_1} p_{x_1 a_1} q_{a_1 a_2}\ldots p_{x_N a_N} q_{a_{N-1}a_N}\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\pi}' \mathbf{V}_1 \mathbf{Q} \mathbf{V}_2 \ldots \mathbf{Q} \mathbf{V}_N \mathbf{1} \\
&= \boldsymbol{\pi}' \left\{ \prod_{t=1}^{N} \mathbf{V}_t \mathbf{Q} \right\} \mathbf{1} \qquad (47)
\end{aligned}
$$

where $\mathbf{V}_t = \text{diag}\,(p_{t1}, p_{t2}, \ldots, p_{tm})$, for $t = 1, \ldots, N$. To derive (47), we note that the second equality is a simple consequence of the conditional independence while the other equalities are obtained by rearranging terms expressing them in matrix notation–see for example [**?**]. Computational methods for the evaluation of the maximum likelihood estimators as well as numerical complications are discussed in [**?**, Section 2.7] where the authors also provide a set of Fortran routines for fitting hidden Markov models at `http://www.statoek.wiso.uni-goettingen.de/links`. Asymptotic properties of maximum likelihood estimators have been recently studied in [**?**] [**?**], [**?**] and [**?**].

# 5 Variable Mixture Models

A broad class of models is formed by switching between models where the switching mechanism is controlled by a hidden process, a hidden Bernoulli process in the case of two components. The resulting marginal probability distribution of the observed process is a mixture of the component distributions, and the problem is to decide which distribution is applicable (the regime probability) through a regression structure. Such models are called *variable mixture models* because the parameters of the observed process keep changing over time due to switching as opposed to, for example, ARMA processes where the parameters are fixed and do not change with time [**?**]. The variable mixture models may be used as alternatives to threshold models where the threshold is either fixed or random, and to hidden Markov models. Useful references on mixture models include [**?**], [**?**], [**?**], [**?**], [**?**].

To define the variable mixture models suppose that $\{Y_t\}$ denotes a univariate time series and $\{\mathbf{Z}_{t-1}\}$, $t = 1, \ldots, N$ is a vector of random time dependent covariates. Furthermore, assume that $\{\mathbf{W}_{t-1}, X_{t0}, X_{t1}\}$ are covariates composed of known functions of $\mathbf{Z}_{t-1}$. Let $\{I_t\}$ be an *unobserved* Bernoulli process taking the values 0,1, and assume that $Y_t$ can obey two different regimes/models where $Y_t$ is generated by the regime 1 distribution if $I_t = 1$ and $Y_t$ is generated by the regime 0 distribution if $I_t = 0$ where

$$
\mathrm{P}_{\boldsymbol{\gamma}} \left[ I_t = 1 \mid \mathcal{F}_{t-1} \right] = F(\mathbf{W}'_{t-1} \boldsymbol{\gamma}) \qquad (48)
$$

and $\boldsymbol{\gamma}$ represents an unknown vector of regression parameters. Typical

choices for $F(\cdot)$ include the logistic (logistic mixture) and standard normal (probit mixture) cdf's. The main ideas of variable mixture models can be put forth in terms of mixtures of two components only.

The previous description can be summarized as follows in the case of two components. The conditional density of $Y_t$ given the indicator $I_t$, the past $\mathcal{F}_{t-1}$, and regime specific covariates and parameters $\alpha_i$, denoted by $f(y_t; \boldsymbol{\theta} \mid \mathcal{F}_{t-1})$, is expressed as,

$$\begin{aligned} g(y_t; \alpha_1 \mid x_{t1}) \quad &\text{if} \quad I_t = 1, \\ g(y_t; \alpha_0 \mid x_{t0}) \quad &\text{if} \quad I_t = 0, \end{aligned} \tag{49}$$

where the functions $g(y_t; \alpha_1 \mid x_{t1})$ and $g(y_t; \alpha_0 \mid x_{t0})$ are probability densities. The components of model (49) may be generalized linear models and in particular autoregressive processes [**?**], [**?**].

Variable mixture models have been considered previously in [**?**], [**?**], where logistic regression is used within the transition matrix of a hidden Markov model, and in [**?**] and [**?**] where logistic mixtures are employed in the mixing of hazard rates. Asymptotic results and testing for the presence of a mixture have been addressed at length in [**?**] for logistic mixtures.

### 5.1 Threshold Models

A closely related idea is that of threshold models developed in [**?**], [**?**]. In the case of two components, the two state self exciting threshold autoregressive (SETAR) model is defined by the switching mechanism,

$$Y_t = \begin{cases} \xi_{01} + \xi_{11} Y_{t-1} + ... + \xi_{p1} Y_{t-p} + \sigma_1 \cdot \epsilon_t \text{ if } Y_{t-d} > \tau \\ \xi_{00} + \xi_{10} Y_{t-1} + ... + \xi_{p0} Y_{t-p} + \sigma_0 \cdot \epsilon_t \text{ if } Y_{t-d} \leq \tau \end{cases}$$

where $d, p, \tau$ are the delay, lag length, and threshold parameters, respectively. These are unknown parameters that can be estimated by conditional least squares. The SETAR model may include more components (states) and also covariates with coefficients that change depending upon whether or not the threshold is exceeded. SETAR goes under the rubric of nonlinear models.

### 5.2 Partial Likelihood Inference

Write $\boldsymbol{\theta} = (\alpha_1, \alpha_0, \boldsymbol{\gamma})'$. Then the conditional distribution of $Y_t$ given the observed past, $\mathcal{F}_{t-1}$, is given by

$$\begin{aligned} f(y_t; \boldsymbol{\theta} \mid \mathcal{F}_{t-1}) \quad &= \quad g(y_t; \alpha_1 \mid x_{t1}) \mathrm{P}_{\boldsymbol{\gamma}} \left[ I_t = 1 \mid \mathcal{F}_{t-1} \right] \\ &+ \quad g(y_t; \alpha_0 \mid x_{t0}) \mathrm{P}_{\boldsymbol{\gamma}} \left[ I_t = 0 \mid \mathcal{F}_{t-1} \right]. \end{aligned} \tag{50}$$

Therefore, the partial likelihood function evaluated at $\boldsymbol{\theta}$ is given by

$$\prod f(y_t; \boldsymbol{\theta} \mid \mathcal{F}_{t-1}).$$

The parameter $\theta$ is estimated by the EM algorithm (see [**?**], [**?**] or [**?**]). The specific algorithm for logistic mixtures is discussed in detail in [**?**].

### 5.3 Comparison With the Threshold Model

It has been noted in [**?**] that simple threshold models may yield inconsistent results if the threshold is measured with error. An example of this is the model

$$Y_t = \begin{cases} .6Y_{t-1} + .5 * \epsilon_t & \text{if } W_{t-1} < .1 \\ .1Y_{t-1} + .3 * \epsilon_t & \text{if } W_{t-1} \geq .1 \end{cases}$$

where $W_t = Y_t + .8\eta_t$ and both $\eta_t$ and $\epsilon_t$ are unobserved i.i.d. $\mathcal{N}(0,1)$, and $\epsilon_t$ is independent of $\eta_s$ for all $s$. That is $W_{t-1}$ is not observed, and only $Y_{t-1}$ is available. Not knowing that the threshold is noisy, the statistician might fit the SETAR model

$$Y_t = \begin{cases} \beta_1 Y_{t-1} + \sigma_1 * \epsilon_t & \text{if } Y_{t-1} < \tau \\ \beta_0 Y_{t-1} + \sigma_0 * \epsilon_t & \text{if } Y_{t-1} \geq \tau \ . \end{cases} \tag{51}$$

An alternative model is the logistic or probit mixture model for the same data. Accordingly, let $\alpha_i = (\beta_i, \sigma_i)$, $i = 0, 1$, and write

$$Y_t = \begin{cases} \beta_1 Y_{t-1} + \sigma_1 * \epsilon_t & \text{if } I_t = 1 \\ \beta_0 Y_{t-1} + \sigma_0 * \epsilon_t & \text{otherwise,} \end{cases} \tag{52}$$

with $\mathrm{P}_{\boldsymbol{\gamma}}[I_t = 1 \mid Y_{t-1}] = F(\gamma_0 + \gamma_1 Y_{t-1})$, and with $F(\cdot)$ either the logistic or the standard normal cdf and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$. In a simulation study the estimates derived under (52) were more precise than those obtained under (51) [**?**]. Notice that the SETAR model is a limiting case of variable mixtures when $\gamma_1 \to -\infty$ and $\gamma_0 = -\tau\gamma_1$.

## 6 ARCH models

*Autoregressive conditionally heteroscedastic* (ARCH) models were introduced by [**?**] to account for changes in volatility, or variability, in time series data. They have been found useful in numerous applications, especially in the context of financial time series which often exhibit large variability.

## 6.1 The ARCH(1) model

Suppose that $\{Y_t\}$, $t = 1, \ldots, N$ denotes the observed response time series. The ARCH(1) model is specified by the following

$$Y_t = \sigma_t \epsilon_t, \tag{53}$$

and

$$\sigma_t^2 = \beta_0 + \beta_1 Y_{t-1}^2 \tag{54}$$

where the coefficient $\beta_1$ is assumed positive and $\{\epsilon_t\}$ is sequence of i.i.d. standard normal random variables. A straightforward consequence of (53) and (54) is that the conditional distribution of $Y_t$ given $Y_{t-1} = y_{t-1}$ is normal with mean 0 and variance $\beta_0 + \beta_1 y_{t-1}^2$. Moreover, subtracting (54) from the square of (53), we obtain

$$Y_t^2 - \left(\beta_0 + \beta_1 Y_{t-1}^2\right) = \sigma_t^2 \left(\epsilon_t^2 - 1\right)$$

or equivalently

$$Y_t^2 = \beta_0 + \beta_1 Y_{t-1}^2 + u_t, \tag{55}$$

where $u_t = \sigma_t^2 \left(\epsilon_t^2 - 1\right)$, a scaled $\mathcal{X}_1^2$ random variable with shifted mean equal to zero. It turns out that both processes $\{Y_t\}$ and $\{u_t\}$ are martingale differences.

Representation (55) shows that the sequence $\{Y_t^2\}$, $t = 1, \ldots, N$ follows an AR(1) process–a fact which immediately points to the following:

$$\begin{aligned} \mathrm{E}\left[Y_t^2\right] &= \mathrm{Var}\left[Y_t\right] = \frac{\beta_0}{1-\beta_1}, \\ \mathrm{Var}\left[Y_t^2\right] &= \mathrm{E}\left[Y_t^4\right] = \frac{3\beta_0^2}{(1-\beta_1)^2} \frac{1-\beta_1^2}{1-3\beta_1^2}, \end{aligned}$$

provided that $0 \leq \beta_1 < 1$, $3\beta_1^2 < 1$ and the variance of $\{u_t\}$, $t = 1, \ldots, N$ is finite. Thus the marginal distribution of $\{Y_t\}$ is *leptokurtic* ("fat tailed" ) since the *kurtosis* is equal to

$$\frac{\mathrm{E}\left[Y_t^4\right]}{\left[\mathrm{E}\left[Y_t^2\right]\right]^2} = 3\frac{1-\beta_1^2}{1-3\beta_1^2}.$$

If $3\beta_1^2 \geq 1$, then the process $\left\{Y_t^2\right\}$ is strictly stationary with infinite variance.

### 6.2 Maximum Likelihood Estimation

Estimation of the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is carried out by maximizing the conditional log–likelihood

$$l(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{t=1}^{N}\left\{\log(\beta_0 + \beta_1 y_{t-1}^2) + \frac{y_t^2}{\beta_0 + \beta_1 y_{t-1}^2}\right\}$$

derived by (53) and (54). For more on estimation see [**?**].

### 6.3 Extensions of ARCH models

There are several extensions of the ARCH(1) model. For example, the so called ARCH($p$) model is specified by equation (53) and the following

$$\sigma_t^2 = \beta_0 + \beta_1 Y_{t-1}^2 + \ldots + \beta_p Y_{t-p}^2.$$

It is also feasible to postulate an ARMA model for the mean and an ARCH model for the errors. Moreover, the so called GARCH (generalized ARCH) models ([**?**]) enlarge the class of ARCH models by stipulating an autoregressive process for $\sigma_t^2$. Specifically, the GARCH(1,1) model stipulates that

$$\sigma_t^2 = \beta_0 + \beta_1 Y_{t-1}^2 + \gamma_1 \sigma_{t-1}^2,$$

in addition to (53). Other extensions include the exponential GARCH (EGARCH) models and the integrated GARCH models (IGARCH), to mention a few. The reader is referred to [**?**], [**?**] and [**?**] for more details.

## 7 Sinusoidal Regression Model

The classical problem of estimating the frequencies $\omega_1, ..., \omega_p$ in the sinusoidal regression model,

$$Z_t = \sum_{j=1}^{p}(A_j \cos(\omega_j t) + B_j \sin(\omega_j t)) + \zeta_t$$

where, $t = 0, \pm 1, \pm 2, \cdots$, has been studied from many different angles for over one hundred years [**?**], [**?**], [**?**]. The amplitude and frequency estimates may be obtained by nonlinear least squares provided the initial guess of the frequencies in a Newton-Raphson type algorithm is sufficiently precise as explained below.

In many cases the frequencies are located "far apart" from each other so that it is sufficient, due to suitable linear operations, to concentrate on the special case of a "single frequency in noise,"

$$Y_t = \beta \cos(\omega_1 t + \phi) + \epsilon_t \tag{56}$$

where $\beta$ is a positive constant, $\omega_1 \in (0, \pi)$, $\phi$ uniformly distributed in $(0, \pi]$, and where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean 0 and variance $\sigma_\epsilon^2$, independent of $\phi$.

In addition to nonlinear least squares, another well known method in frequency detection and estimation is based on the *periodogram*. Denoted by $I_N(\omega)$, the periodogram is defined by the transformation,

$$I_N(\omega) = \frac{2}{N} \left| \sum_{t=1}^{N} Y_t \exp(-i\omega t) \right|^2, \qquad \omega \in [-\pi, \pi] \tag{57}$$

where $Y_1, ..., Y_N$ is the time series to be transformed. Periodogram analysis consists of the search for significant peaks in $I_N(\omega)$ by treating $\omega$ as a continuous variable or, much more often, as a discrete variable taking the values in $[0, \pi]$,

$$\frac{2\pi k}{N}, \qquad k = 0, 1, \cdots, \left[ \frac{N}{2} \right],$$

known as Fourier frequencies. In the latter case the periodogram can be computed very rapidly by the *fast Fourier transform*, FFT, a great computational device [**?**]. When a time series has a significant sinusoidal component with frequency $\omega_0 \in [0, \pi]$, then the periodogram exhibits a peak at that frequency with a high probability. In the single frequency regression model (56) with a fixed frequency and phase, the periodogram, apart from a constant, acts as the sample correlation between the observed series $Y$ and the fitted data $\hat{Y}$ obtained from a linear least squares fit. The optimal frequency is then the one that maximizes this sample correlation, or the periodogram.

Given a time series $Y_1, ..., Y_N$, the more traditional methods of nonlinear least squares and periodogram maximization result in efficient estimates with precision $O(N^{-3/2})$, provided the optimization is initialized with precision $o(N^{-1})$. This requirement of a very precise initial guess of the frequencies may not be feasible in practice, and alternative methods are called for. The contraction mapping (CM) method which estimates the frequency by fixed point iterations relaxes the requirement of a very precise initial guess. With this method, under some conditions guesses precise to order $O(1)$ can

still lead to estimates arbitrarily close to being efficient, and at the same time the method is computationally simple and fast [**?**], [**?**], [**?**], [**?**], [**?**].

With $\alpha \in (-1, 1)$, define $Y_t(\alpha)$ by operating on $Y_t$,

$$Y_t(\alpha) = (1 + \eta^2)\alpha Y_{t-1}(\alpha) - \eta^2 Y_{t-2}(\alpha) + Y_t \tag{58}$$

where $\eta \in (0, 1)$ is the *contraction parameter*. Put $Y_{-1}(\alpha) = Y_0(\alpha) = 0$. Define the sample autocorrelation

$$\hat{\rho}_1(\alpha) = \frac{\sum_{t=1}^{N-1} Y_t(\alpha) Y_{t-1}(\alpha)}{\sum_{t=0}^{N-1} Y_t^2(\alpha)} \tag{59}$$

Then the CM algorithm is given by,

$$\alpha_{k+1} = \hat{\rho}_1(\alpha_k), \qquad k = 0, 1, 2, \cdots \tag{60}$$

For an initial guess $\theta_0$ of $\omega_1$, the algorithm starts with $\alpha_0 = \cos(\theta_0)$, and $\eta$ close to 1, for example $\eta = 0.98$. We obtain $\{Y_t(\alpha_0)\}$ from (58), then $\hat{\rho}_1(\alpha_0)$ from (59), then get a new $\alpha_1$ from (60), $\alpha_1 = \hat{\rho}_1(\alpha_0)$, we increase $\eta$ slightly and start anew with $\alpha_1$ and a new $\eta$. This iterative scheme gives a sequence $\alpha_0, \alpha_1, \alpha_2 ...$, which converges to $\alpha^*$. Under certain conditions, $\alpha^* = \alpha_N^*$ converges almost surely to $\cos(\omega_1)$ as $N$ increases. The desired estimator is given by

$$\hat{\omega}_1 = \arccos(\alpha^*).$$

It has been shown recently in [**?**] under regularity conditions that if $\eta$ is chosen such that $(1 - \eta)^2 N \to 0$ as $N \to \infty$, then $(1 - \eta)^{-1/2} N(\hat{\omega}_1 - \omega_1)$ converges in distribution to $\mathcal{N}(0, \gamma^{-1})$ as $N \to \infty$, where $\gamma = \frac{1}{2}\beta^2/\sigma_\epsilon^2$. The implication of this is that by a judicious choice of $\eta$, the precision of the CM estimate can be made arbitrarily close to that achieved by periodogram maximization and nonlinear least squares. Based on this and other results the authors suggest a two-step procedure for the implementation of the CM algorithm. See [**?**] and [**?**] for many more important related references from the statistical and engineering literature.

The following S-Plus function `KY.AR2()` gives the code for the CM algorithm. The arguments of `KY.AR2()` are $z, \theta_0, \eta, inc, niter$, where $z$ is a vector containing the data, $\theta_0$ is the initial guess of the frequency of interest, $\eta$ is the initial value of the contraction parameter, *inc* is the increment of $\eta$ at each iteration, and *niter* is the number of iterations of the algorithm. The output includes the sequence of estimates, and the sample variance of the filtered process. An appreciable increase in the variance accompanies detection. The algorithm makes use of the S-Plus functions *acf(), filter()*.

```
KY.AR2 <- function(z,theta0,eta,inc,niter){
y <- rep(0,length(z))
r <- rep(0,niter); OMEGA <- rep(0,niter)
r [1] <- cos(theta0); OMEGA[1] <- theta0
cat(c("Initial frequency guess is", OMEGA[1]),fill=T)
cat(c("eta","         r(k)","         Omega(k)",
"          Var(y)"), fill=T)
for(k in 2:niter){# eta increments by inc
eta <- eta+inc
if((eta < 0)|(eta >1))
  stop("eta must be between 0 and 1")
FiltCoeff <- c((1+eta^2)*r[k-1],-(eta^2))
y <- filter(z,FiltCoeff, "rec")
# CM Iterations-------------------
rrr <- acf(y)      # motif() must be on
r[k] <- rrr$acf[2] # Gives acf(1)!!!
# -------------------------------
OMEGA[k] <- acos(r[k])
cat(c(eta,r[k],OMEGA[k],var(y)),fill=T)}}
```

The algorithm can be used also for multiple frequencies by starting the algorithm with different initial guesses $\theta_0$ of $\omega_1$, that is, centering the filter at different regions in $(0, \pi)$.

As an example consider the sum of two sinusoids plus noise with frequencies $\omega_1 = 0.513$, $\omega_2 = 0.771$,

$$Y_t = 0.5\cos(0.513t + \phi_1) + \cos(0.771t + \phi_2) + 2.2\epsilon_t$$

where the $\epsilon_t$ are i.i.d. $\mathcal{N}(0, 1)$ random variables, and $t = 1, ..., 1500$. The signal to noise ratio defined as $10\log 10((.5^2/2 + 1^2/2)/2.2^2) = -8.890$ is relatively low.

Storing the data in a vector $z$ and calling `KY.AR2(z,0.48,0.98,0.0015,10)` gives Table 5. The first iteration is not shown, only the last nine. Similarly, calling `KY.AR2(z,0.88,0.98,0.001,10)` starting at $\theta_0 = 0.88$ and changing the increment to 0.001 gives Table 6. The estimates are 0.5135, 0.7709, respectively, so that in both cases the error is of order $10^{-4}$. The FFT evaluated at Fourier frequencies gives 0.5152, 0.7749, respectively, or errors of order $10^{-3}$.

It is interesting to observe in Tables 5 and 6 the steady increase in the variance of the the filtered series $Y_t(\alpha)$ as the CM algorithm builds up the power of the frequency to be detected.

Table 5: Nine iterations of the CM algorithm in the estimation of $\omega_1 = 0.513$. Convergence of $\hat{\omega}$ starting at $\theta_0 = 0.48$, and $\eta = 0.98$ increasing by 0.0015. The final estimate is $\hat{\omega} = 0.5135$. Error: 0.0005.

| $\eta$ | $\alpha(k)$ | $\omega(k)$ | $\text{Var}(Y_t(\alpha))$ |
|---|---|---|---|
| 0.9815 | 0.8807 | 0.4932 | 425.958 |
| 0.9830 | 0.8755 | 0.5042 | 574.342 |
| 0.9845 | 0.8723 | 0.5107 | 856.165 |
| 0.9860 | 0.8711 | 0.5131 | 1134.483 |
| 0.9875 | 0.8708 | 0.5138 | 1365.735 |
| 0.9890 | 0.8707 | 0.5139 | 1666.432 |
| 0.9905 | 0.8708 | 0.5138 | 2106.870 |
| 0.9920 | 0.8709 | 0.5136 | 2783.643 |
| 0.9935 | 0.8710 | 0.5135 | 3892.713 |

Table 6: Nine iterations of the CM algorithm in the estimation of $\omega_2 = 0.771$. Convergence of $\hat{\omega}$ starting at $\theta_0 = 0.88$, and $\eta = 0.98$ increasing by 0.001. The final estimate is $\hat{\omega} = 0.7709$. Error: 0.0001.

| $\eta$ | $\alpha(k)$ | $\omega(k)$ | $\text{Var}(Y_t(\alpha))$ |
|---|---|---|---|
| 0.981 | 0.6518 | 0.8607 | 102.987 |
| 0.982 | 0.6672 | 0.8403 | 128.128 |
| 0.983 | 0.6822 | 0.8199 | 162.341 |
| 0.984 | 0.6973 | 0.7990 | 215.022 |
| 0.985 | 0.7104 | 0.7806 | 371.580 |
| 0.986 | 0.7162 | 0.7723 | 988.001 |
| 0.987 | 0.7171 | 0.7710 | 1555.238 |
| 0.988 | 0.7172 | 0.7708 | 1817.274 |
| 0.989 | 0.7172 | 0.7709 | 2130.545 |

# 8 Problems and Complements

1. Prove equations (6)–(10).

2. Prove the following expressions:

   (a)
   $$\mathrm{E}\left[\alpha \circ X\right]^2 = \alpha(1-\alpha)\mathrm{E}\left[X\right] + \alpha^2 \mathrm{E}\left[X^2\right].$$

   (b)
   $$\mathrm{E}\left[\alpha \circ X - \alpha \circ Z\right]^2 = \alpha(1-\alpha)\mathrm{E} \mid X - Z \mid + \alpha^2 \mathrm{E}\left[X - Z\right]^2,$$
   with $\alpha \circ X = \sum_{i=1}^{X} Y_i$ and $\alpha \circ Z = \sum_{i=1}^{Z} Y_i$.

   (c)
   $$\mathrm{E}\left[(\alpha \circ X)(\beta \circ Z)\right] = \alpha\beta \mathrm{E}\left[XZ\right],$$
   where $\alpha \circ X = \sum_{i=1}^{X} Y_i$, $\beta \circ Z = \sum_{i=1}^{Z} Y_i^\star$, $\{Y_i\}$ is independent of $\{Y_i^\star\}$ and $(X, Z)$ is independent of $\{Y_i\}$ and $\{Y_i^\star\}$.

3. Derive expression (15) by using the representation (12).

4. a. Fix $m, \lambda$ and simulate the process (1), for $n = 1, ..., N$, $N = 100, 200, ..., 2000$. Obtain estimates for $m, \lambda$ by minimizing $\sum_{i=1}^{N} (\epsilon_n^*)^2$ in
   $$\frac{X_n}{\sqrt{X_{n-1}+1}} = m\frac{X_{n-1}}{\sqrt{X_{n-1}+1}} + \frac{\lambda}{\sqrt{X_{n-1}+1}} + \epsilon_n^*$$
   where $\epsilon_n^* = \epsilon_n / \sqrt{X_{n-1}+1}$.

   b. Consider the design matrix,
   $$\mathbf{X}_N = \begin{pmatrix} X_1/\sqrt{X_1+1} & 1/\sqrt{X_1+1} \\ X_2/\sqrt{X_2+1} & 1/\sqrt{X_2+1} \\ . & . \\ . & . \\ . & . \\ X_N/\sqrt{X_N+1} & 1/\sqrt{X_N+1} \end{pmatrix}$$
   and define $\mathbf{A} \equiv \mathbf{X}_N'\mathbf{X}_N$. Let $\lambda_{min}(N)$ and $\lambda_{max}(N)$ be the smallest and largest eigenvalues of $\mathbf{A}$, respectively. Verify that $[\log \lambda_{max}(N)]/\lambda_{min}(N)$ decreases as $N$ increases, and interpret the results relative to the convergence of your least squares estimates [**?**].

5. More on the estimation of $m$ in (1) [**?**]. By a simulation study compare the estimate of $m$ given by (4) with the estimator

$$1 - \frac{1}{2}\frac{\sum_{i=1}^{N}(X_{i-1} - X_i)^2}{\sum_{i=1}^{N}(X_{i-1} - \bar{X}_N)^2},$$

$\bar{X}_N = (X_1 + \cdots + X_N)/N$, assuming the immigration distribution is Poisson, $m < 1$, and that the process is stationary.

6. Recall the INAR(1) process (11) and assume that the innovations follow the Poisson distribution with mean $\mu$.

   (a) By differentiating

   $$\sum_{t=1}^{N} (X_t - \alpha X_{t-1} - \mu)^2,$$

   show that the conditional least squares estimators of $\alpha$ and $\mu$ are given by

   $$\hat{\alpha} = \frac{\sum_{t=1}^{N} X_t X_{t-1} - \left(\sum_{t=1}^{N} X_t \sum_{t=1}^{N} X_{t-1}\right)/N}{\sum_{t=1}^{N} X_{t-1}^2 - \left(\sum_{t=1}^{N} X_{t-1}\right)^2/N}$$

   and

   $$\hat{\mu} = \frac{1}{N}\left(\sum_{t=1}^{N} X_t - \hat{\alpha}\sum_{t=1}^{N} X_{t-1}\right),$$

   respectively.

   (b) Verify the conditions of [**?**, th. 3.2] to conclude that the conditional least squares derived above, are asymptotically normally distributed. In addition calculate the asymptotic covariance matrix.

7. For the integer moving average model of order $q$ (24), show the following:

   (a) Verify equations (30)–(33).

   (b) Assume that the sequence $\{\epsilon_t\}$ is Poisson with mean $\mu/(\sum_{i=0}^{q}\beta_i)$. Then the autocorrelation function is given by

   $$\rho(k) = \begin{cases} \sum_{i=0}^{q-k}\beta_i\beta_{i+k}/\sum_{i=0}^{q}\beta_i, & k = 0, 1, \ldots, q \\ 0, & k > q \end{cases}$$

(c) If $\{\epsilon_t\}$ is Poisson with mean $\mu/\sum_{i=0}^{q}\beta_i$, then $X_t$ is Poisson with mean $\mu$.

8. By combining the INAR(1) model and the INMA($q$) we obtain an autoregressive moving average process of order $(1, q)$. Indeed, let $\epsilon_t$ be a sequence of independent and identically distributed Poisson random variables with mean $(1-\alpha)\mu$. and put

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t, \tag{61}$$

$$X_t = Y_{t-q} + \sum_{k=1}^{q} \beta_k \circ \epsilon_{t+1-k}, \tag{62}$$

with $0 < \alpha < 1$, $0 < \beta_k < 1$ for $k = 1, \ldots, q$ and $\sum_{k=1}^{q}\beta_k < 1$. In addition assume that $Y_0$ is a Poisson random variable with mean $\mu$ which is independent of $\{\epsilon_t\}$. Show that the autocorrelation function of the process specified by (61) and (62) is given by

$$\rho(k) = \begin{cases} \frac{\left\{\alpha^k + (1-\alpha)\left(\sum_{i=1}^{q-k}\beta_i\beta_{i+k} + \sum_{i=q-k+1}^{q}\beta_i\alpha^{k-1+i-q}\right)\right\}}{(1+(1-\alpha)\sum_k \beta_k)}, & k = 1, \ldots, q \\ \frac{\left(\alpha^q + (1-\alpha)\sum_{i=1}^{q}\beta_i\alpha^{i-1}\right)\alpha^{k-q}}{(1+(1-\alpha)\sum_k \beta_k)}, & k > q \end{cases}$$

9. The discrete autoregressive process of order 1 has the following form

$$A_t = V_t A_{t-1} + (1 - V_t)Y_t,$$

with the same notation as in (35).

(a) Simulate a realization of the process $\{A_t\}$ of length 200. Use the Poisson distribution for $\pi$.

(b) Show that the autocorrelation function of the process $\{A_t\}$ is $\rho^k$.

(c) Suppose that for a realization of length $T$ from $\{A_t\}$, we denote by $m_{ij}$ the number of times that the process moves from state $i$ to $j$ and $m_j$ stands for the number of times that the process is in state $j$. Show that the log–likelihood function for a DAR(1) process is given by

$$\begin{aligned} L &= \sum_{i=0}^{\infty}\sum_{j\neq i} m_{ij}\log\left[(1-\rho)\pi_j\right] + \sum_{i=0}^{\infty} m_{ii}\log\left[1-(1-\rho)(1-\pi_i)\right] \\ &+ \sum_{i=0}^{\infty} I(X_1 = i)\log p_i, \end{aligned}$$

where $I(B)$ is the indicator function of the event $B$. Hence, taking derivative with respect to $x = 1 - \rho$, the maximum likelihood estimator of $x$, if it exists, is given by

$$1 - \frac{1}{m-1} \sum_{i=0}^{\infty} m_{ii} \frac{1}{1 - x[1 - \pi_i]} = 0.$$

10. (a) For the mixture transition model (37), prove that it satisfies the system of equations (38).

    (b) If $l = 2$, then equations (38) have unique solution when $0 < \lambda_1 \leq 1$.

11. Suppose that $\{Y_t\}$, $t = 1, \ldots, N$ is a time series taking values in an arbitrary space. The Gaussian mixture transition distribution model (GMTD) is given by ([**?**])

$$\begin{aligned} F(y_t \mid y_1, \ldots y_{t-1}) &= \alpha_0 \Phi \left( \frac{y_t - \sum_{j=1}^{p} \phi_{0j} y_{t-j}}{\sigma_0} \right) \\ &+ \sum_{i=1}^{p} \alpha_i \Phi \left( \frac{y_t - \phi_i y_{t-i}}{\sigma_i} \right) \end{aligned} \tag{63}$$

where $F$ is the conditional distribution function of $Y_t$ given the past of the process, $\alpha_i \geq 0$, $i = 0, \ldots, p$ such that $\sum_{i=0}^{p} \alpha_i = 1$ and $\Phi$ is the cumulative distribution function of a Gaussian random variable. Suppose that $Y_t$ is second order stationary, prove the following:

    (a) If $\rho(l)$ denotes the lag-$l$ autocorrelation, then

$$\rho(l) = \sum_i (\alpha_0 \phi_{0i} + \alpha_i \phi_i) \, \rho(\mid l - i \mid),$$

    for $l = 1, \ldots, p$. These equations, for different values of $l$, resemble the Yule–Walker equations for the AR($p$) model upon noticing that the coefficients $\alpha_0 \phi_{0i} + \alpha_i \phi_i$ are replaced by the lag-$i$ coefficient of the AR($p$) process.

    (b) Suppose that $\alpha_0 = 0$ and $p = 2$. Derive the admissible region for the autocorrelations $\rho(1)$ and $\rho(2)$.

12. For a hidden Markov model satisfying the relations (40)–(42), show that

$$\mathrm{E}\left[g(X_t)\right] = \sum_{a=1}^{m} \mathrm{E}\left[g(X_t) \mid A_t = a\right] \pi_a,$$

and

$$\mathrm{E}\left[g(X_t, X_{t+k})\right] = \sum_{a=1}^{m}\sum_{b=1}^{m} \mathrm{E}\left[g(X_t, X_{t+k}) \mid A_t = a, A_{t+k} = b\right] \pi q_{ij}^{(k)},$$

provided that all expectations exist.

13. Define the binomial hidden Markov model by specifying $p_{xa}$ of (42) to be the binomial distribution with parameters $n$ and $\zeta_a$ ($0 < \zeta_a < 1$). Calculate the mean, variance, autocovariance and autocorrelation functions of the binomial hidden Markov Model.

14. Based on (47), show the following

    (a) The marginal distribution of $X_t$ is given by

    $$\mathrm{P}\left[X_t = x\right] = \boldsymbol{\pi}\mathbf{V}_x\mathbf{1}'.$$

    (b) The joint distribution of $(X_t, X_{t+1})$ is

    $$\mathrm{P}\left[X_t = u, X_{t+1} = v\right] = \boldsymbol{\pi}\mathbf{V}_u\mathbf{Q}\mathbf{V}_v\mathbf{1}'.$$

    (c) The one–step–ahead distribution is

    $$\mathrm{P}\left[X_{t+1} = x_{t+1} \mid X_1 = x_1, \ldots, X_t = x_t\right] = \frac{\boldsymbol{\pi}\mathbf{V}_1\mathbf{Q}\mathbf{V}_2 \ldots \mathbf{Q}\mathbf{V}_{t+1}\mathbf{1}'}{\boldsymbol{\pi}\mathbf{V}_1\mathbf{Q}\mathbf{V}_2 \ldots \mathbf{Q}\mathbf{V}_t\mathbf{1}'}$$