CATEGORICAL TIME SERIES

by

Xiaoa Zhen

(Under the direction of Ishwar V. Basawa)

Abstract

Standard time series models typically assume that the data are continuous. If the available data consist of counts of observations in a finite number of categories, the usual autoregressive moving average (ARMA) models cannot be applied to fit the count data. For each time t, the count vector $N(t) = (N_1(t), N_2(t), ..., N_m(t))'$ can be modeled by a multinomial distribution. The vector time series $\{N(t)\}$, t=0,1,2,..., is then a categorical time series. The main goal of this dissertation is to present new models for $\{N(t)\}$, by introducing dependence over time.

The new models in this dissertation include (i) binary models obtained by clipping (or grouping) Gaussian processes, (ii) observation driven state space models both for binary and multi-category models, and (iii) models for dependent contingency tables. These models are applied to real data sets to illustrate the models and methods developed.

Index words: Binary data, Categorical time series, Partial likelihood, Maximum likelihood, State space models, Dependent contingency tables

CATEGORICAL TIME SERIES

by

Xiaoa Zhen

B.E., Tianjin Polytechnic University, 1994

China

M.S.,the University of Georgia, 2007

A Thesis Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2008

CATEGORICAL TIME SERIES

by

Xiaoa Zhen

Approved:

Major Professor: Ishwar V. Basawa

Committee: Gauri Datta
Daniel Hall
Jaxk Reeves
Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

Dedication

To my parents, my husband and my son

Acknowledgments

At the very first, I am deeply indebted to my dedicated advisor Prof. Basawa whose patience, meticulous guidance and encouragement accompanied me in all the time of research for and writing of this dissertation. I have learnt from him a lot not only about dissertation research, but also the professional ethics.

Especially, I want to extend my gratitude to all my committee members: Dr.Gauri Datta, Dr.Daniel Hall, Dr.Jaxk Reeves and Dr.Lynne Seymour, who devote their time serving as my committee members and never hesitate to offer their helps, supports, and valuable suggestions.

Many thanks also go to my fellow colleagues and friends here at STATISTICS department for their kindness, cooperations and helps.

Finally, I want to thank my husband yunzhou and my son shaozi, without them this dissertation would not be possible!

Table of Contents

List of Tables

Chapter 1

Introduction

This dissertation presents results on time series models for categorical data and related inference as well as prediction problems. A sequence of random vector variables $\{Y_t\} = (Y_{t,1}, Y_{t,2}, ...Y_{t,m})'$, t=1,2,..., is a categorical time series if

$$Y_{t,j} = \begin{cases} 1 & \text{when jth category is observed, j=1,...,m} \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Note that $\sum_{j=1}^{m} Y_{t,j} = 1$. Generalized linear models for $\{Y_t\}$ conditional on past data and possible covariates will be discussed.

For m=2, we have a univariate binary sequence $\{Y_t\}$. As an illustration, we shall focus here on the special case m=2. Examples of a binary process include level crossings in a Gaussian process, crossing certain temperature levels in global warming studies, exceeding regulated pollution levels in the atmosphere or in water, occurrences of earthquakes, volcanic eruptions, stock market corrections, epileptic seizures, etc. For these and many other examples where we are interested in occurrence or non-occurrence of certain events, we have a binary process $\{Y_t\}$ defined by

$$Y_t = \begin{cases} 1 & \text{if the event occurs} \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

The sequence of random variables $\{Y_t\}$ is, in general, a dependent Bernoulli sequence. It is of interest to model the dependency structure for $\{Y_t\}$. The regression approach seeks to model the conditional distribution of Y_t given the past data $y_{t-1}^* = (y_{t-1}, y_{t-2}, ..., y_1)$ and possible covariates $x_{t-1}^* = (x_{t-1}, x_{t-2}, ..., x_1)$. Generalized linear models have proved very

useful in the regression approach. An alternative approach is to model the joint distribution $p(y_1, y_2, ..., y_n)$ from an underlying 'state' process $\{\pi_t\}$ using the 'state-space' formulation. In this dissertation we present a combination of the two approaches. More specifically, we use observation-driven state processes where the conditional distribution of π_t is specified in terms of the past data on $\{Y_t\}$. For parameter estimation, we use the maximum likelihood method, whenever the likelihood function $p(y_1, y_2, ..., y_n)$ can be constructed easily. When the covariates $\{x_t\}$ are themselves random variables with unknown joint distributions,we use the partial likelihood method for parameter estimation. Certain Bayesian tools are also used in developing state-space models.

Chapter 2 gives a literature review. Standard regression type models, partial likelihood and maximum likelihood estimation methods are reviewed. Binary models obtained by clipping Gaussian autoregressive processes are introduced in Chapter 3. Five alternative methods of estimation are discussed and compared via simulation. A real data application is also given. Chapter 4 contains new state space models for categorical time series. The new observation-driven state space models are used to analyze binary data on IBM and Dow Jones Index data. As an application to multi-category time series, we present DNA data analysis. Finally, Chapter 5 presents results on time series models for dependent contingency tables. Multinomial-logit, conditional exponential family, Markov chain and multinomial-Dirichlet models are discussed and applied to two real data sets.

Chapter 2

Literature Review

In this chapter, we review some standard categorical time series models. Regression type models have been discussed by Kaufmann (1987), Fahrmeir and Kaufmann (1987), Fokianos and Kedem (1998, 2003) among others. Fahrmeir and Tutz (2001) have studied state-space models while Kedem (1980) and Slud & Kedem (1994) have discussed binary models derived from Gaussian processes. Literature on categorical time series appears to be relatively sparse.

2.1 Binary Models: Partial Likelihood Approach

In numerous practical situations, one is interested in the prediction of a future value of a stationary or non-stationary univariate binary time series $\{Y_t\}$, $t = 0, \pm 1, \pm 2, ...$, from past values of $\{Y_t\}$ and past (and sometimes also present) values of covariate variables $\{X_t\}$. That is, $\{Y_t\}$ is predicted from the past either only given the past data which generate the σ-field: $\mathscr{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, ..., x_{t-1}, x_{t-2}, ...)$; or the covariate information at time t is also known before observing y_t, in which case, Y_t is predicted given the σ-field $\mathscr{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, ..., x_t, x_{t-1}, ...)$. In either case, X_t may contain components that are functions of Y_t. The goal is to estimate from past information the one-step conditional probability $p_t = P(Y_t = 1 | \mathscr{F}_{t-1})$.

2.1.1 Partial Likelihood

Partial likelihood (PL) was introduced by Cox (1972,1975) and given more formal definition and theoretical justification by Wong (1986). The general definition given below follows Slud (1992).

Let $\mathscr{F}_k, k = 0, 1, 2...$, be an increasing sequence of σ-fields, and let $Y_1, Y_2, ...,$ be a sequence of random variables on some common probability space, such that Y_k is $\mathscr{F}_k$ measurable. Let $p_k(y_k; \theta)$ be the conditional probability density given $\mathscr{F}_{k-1}$ for Y_k under probability measure P_θ. The partial likelihood function based on the sample $(y_1, ..., y_N)$ is defined as:

$$PL(\theta; y_N) \equiv PL(\theta; y_1, ..., y_N) = \prod_{k=1}^{N} p_k(y_k; \theta).$$

Note that if $\mathscr{F}_k$ is generated by $\{Y_k, Y_{k-1}, ...,\}$ only, the partial likelihood is the same as the ordinary likelihood. If $\mathscr{F}_k$ contains other random variables, say $\{x_k\}$, then we have a partial likelihood.

2.1.2 The Logistic Regression Model

Let $\{Z_t\}, t = 0, \pm1, \pm2, ...,$ be an autoregressive process of order p,

$$Z_t = \beta_1 Z_{t-1} + ... + \beta_p Z_{t-p} + \varepsilon_t$$

where ε_t are i.i.d random variables logistically distributed with density function $e^x/(1+e^x)^2$. Now fix a threshold $r \in (-\infty, \infty)$, and define a binary time series by

$$Y_t \equiv I_{[Z_t \geq r]} = \begin{cases} 1 & if \quad Z_t \geq r \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Then

$$p_t(\beta) \equiv P_\beta(Y_t = 1 | \mathscr{F}_{t-1}) = \frac{1}{1 + \exp[-(-r + \beta_1 Z_{t-1} + ... + \beta_p Z_{t-p})]} \tag{2.2}$$

where $Z_{(t-1)} = (Z_{t-1}, ..., Z_{t-p})$. Since $\{Y_t\}$ is binary, the conditional density of y_t is given by

$$p_t(y_t; \beta) = [p_t(\beta)]^{y_t} [1 - p_t(\beta)]^{1-y_t}.$$

The corresponding partial likelihood is simply the product

$$PL(\beta) = \prod_{t=1}^{N} p_t(y_t; \beta) = \prod_{t=1}^{N} [p_t(\beta)]^{y_t} [1 - p_t(\beta)]^{1-y_t}. \tag{2.3}$$

The value of β which maximizes $PL(\beta)$ is called the Maximum Partial Likelihood Estimator (MPLE) of β.

Knowing that conditionally given $\mathscr{F}_{t-1}$, the binary variable $\{Y_t\}$ has mean $p_t(\beta)$ and variance $p_t(\beta)(1-p_t(\beta))$, we have

$$E_\beta[Z_{(s-1)}Z'_{(t-1)}(Y_s - p_s(\beta))(Y_t - p_t(\beta))|\mathscr{F}_{s-1}] = \begin{cases} 0 & \text{if } s < t \\ Z_{(s-1)}Z'_{(s-1)}p_s(\beta)(1-p_s(\beta)) & \text{if } \quad s = t. \end{cases} \tag{2.4}$$

The score function is defined by:

$$S_N(\beta) \equiv \bigtriangledown logPL(\beta) = \sum_{s=1}^{N} Z_{(s-1)}(Y_s - p_s(\beta)). \tag{2.5}$$

which is easily seen to be a martingale. That is, $E[S_t(\beta)|\mathscr{F}_{t-1}] = S_{t-1}(\beta)$. Clearly, $E[S_t(\beta)] = 0$. Next define

$$I(\beta) \equiv \bigtriangledown \bigtriangledown' (-logPL(\beta)) = \sum_{t=1}^{N} Z_{(t-1)}Z'_{(t-1)}p_t(\beta)(1-p_t(\beta)). \tag{2.6}$$

The quantity $I(\beta)/N$ is *the sample information matrix per observation*. Note that $I(\beta)$ also can be seen as the sum of conditional covariance matrices.

$$I(\beta) = \sum_{s=1}^{N} Var_\beta[Z_{(s-1)}(Y_s - p_s(\beta))|\mathscr{F}_{s-1}]. \tag{2.7}$$

Since $S_t(\beta)$ is a martingale, we also can consider $I(\beta)$ as the cumulative conditional variance-covariance matrix for $S_N(\beta)$.

The large sample properties of the MPLE $\hat{\beta}$ are studied with the aids of $S_t(\beta)$ and $I(\beta)$, see Andersen and Gill (1982), Wong (1986) as well as Arjas and Haara (1987). The approach taken in these references for providing consistency and asymptotic normality of MPLE's is based on the martingale central limit theorem, the almost sure concavity of the random function PL, and the stability of the sample information matrix $I(\beta)/N$.

Theorem: Under regularity conditions, the MPLE $\hat{\beta}$ is almost surely unique for all sufficiently large N, and as $N \to \infty$,

(i) $\hat{\beta} \xrightarrow{p} \beta_0$

(ii) $\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Lambda^{-1}(\beta_0))$,

where $\Lambda(\beta)$ is the probability limit of $I(\beta)/N$.

2.1.3 Probit Link Function

Equation (2.2) implies that the logistic link function is used. Instead, we can use the probit 'link' function Φ where Φ is the standard normal probability distribution function. In this case we obtain what is known as a probit model, $p_t(\beta) \equiv P_\beta(Y_t = 1|\mathscr{F}_{t-1}) = \Phi(\beta' Z_{(t-1)})$. Virtually every aspect of analysis is analogous except that in the AR(p) example here the errors are now Gaussian instead of logistic and the proofs of theorems are very similar. (Slud and Kedem (1994))

2.2 Regression Models for Categorical Time Series

For categorical time series$\{Y_t\}$with m categories for each observation, setting $q = m - 1$, we define the observation vector $Y_t = (Y_{t1}, ..., Y_{tq})'$ by

$$Y_{tj} = \begin{cases} 1 & \text{category j has been observed, j=1,...,q} \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

Correspondingly π_t denotes the q-vector of conditional probabilities

$$\pi_{tj} = P(Y_{tj} = 1|y_{t-1}, ..., y_1), j = 1, ..., q.$$

A general *regression model for categorical time series* is given: $\pi_t = h(Z_t'\beta)$, where β is a vector of unknown parameters. The link function h is a one-to-one mapping from a q-dimensional region $D \subset \Re^q$ to the set $\{(\pi_1, ..., \pi_q)', \pi_j > 0, \sum \pi_j < 1\}$. The predetermined variables now form a matrix Z_t, which is a function of past observations and non-random exogenous variables.

From the first t observations, the parameter β can be estimated by the method of maximum likelihood.

If

$$y_{sm} = 1 - \textstyle\sum_{j=1}^{m-1} y_{sj}, \;\; \pi_{sm} = 1 - \sum_{j=1}^{m-1} \pi_{sj},$$

the log-likelihood based on the observations $\{y_{s,j}\}$, $s = 1, ..., t$ and $j = 1, ..., m$ is

$$l_t(\beta) = \sum_{s=1}^{t} \sum_{j=1}^{m} y_{sj} ln \pi_{sj}.$$

Its first derivative, the score function, is

$$s_t(\beta) = \sum Z_s D_s(\beta) \Sigma_s^{-1}(\beta)(Y_s - \pi_s(\beta)),$$

with $D_s(\beta) = \partial h(\gamma)/\partial \gamma'$, evaluated at $\gamma = Z_s'\beta$, and $\Sigma_s(\beta) = cov_\beta(Y_s|y_{s-1}, ..., y_1)$. The conditional information matrix is given by

$$G_t(\beta) = \sum Z_s D_s(\beta) \Sigma_s^{-1}(\beta) D_s'(\beta) Z_s'.$$

For the widely used logit link function, we have

$$\pi_{si} = \frac{\exp(z_s'\beta si)}{1+\sum_{r=1}^{q} \exp(z_s'\beta_{sr})}, \quad i = 1, ..., q.$$

Now $h = (h(\gamma_1), ..., h(\gamma_q))'$ with $h(\gamma_i) = \pi_{si}$ as above and $\gamma_i = z_s'\beta_{si}$. The design matrix Z_s will be

$$Z_s = \begin{bmatrix} z_s' & 0 & . & . & . & 0 \\ 0 & z_s' & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 0 & 0 & . & . & . & z_s' \end{bmatrix}$$

and $D_s(\beta) = \Sigma_s(\beta)$ below:

$$\Sigma_s(\beta) = \begin{bmatrix} \pi_{s1}(1-\pi_{s1}) & -\pi_{s1}\pi_{s2} & ... & -\pi_{s1}\pi_{sq} \\ . & . & ... & . \\ . & . & ... & . \\ & & & \\ -\pi_{sq}\pi_{s1} & -\pi_{sq}\pi_{s2} & ... & \pi_{sq}(1-\pi_{sq}) \end{bmatrix}.$$

Now,the score function becomes

$$s_t(\beta) = \sum Z_s(Y_s - \pi_s(\beta)),$$

and the conditional information matrix is simplified to:

$$G_t(\beta) = \sum Z_s \Sigma_s(\beta) Z_s'.$$

2.2.1 Asymptotic Theory

Theorem Under regularity conditions, there exists a sequence $\{\hat{\beta}_t\}$ of MLE's which is consistent and asymptotically normal,

$$G_t^{T/2}(\hat{\beta}_t - \beta) \xrightarrow{d} N(0, I)$$

where $G_t^{T/2}(\beta)$ is the right Cholesky square root of $G_t(\beta)$. See (Kaufmannn (1987)) for a proof.

Fokianos & Kedem (1998) extend these large sample results by considering stochastic time-dependent covariates and by dropping the Markovian assumption. They use the concept of partial likelihood which simplifies conditional inference and obviates the Markov assumption.

Chapter 3

Estimation for Binary Models Generated by Gaussian Autoregressive Processes

3.1 Introduction

Kedem (1980), Slud and Kedem (1994), Kedem and Fokianos (2002), and Fokianos and Kedem (1998,2003), among others have discussed models for correlated binary data $\{Y_t\}$, t=1,2,..., where

$$Y_t = \begin{cases} 1 & \text{if the event occurs} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Examples of binary data include level crossings in a Gaussian process, crossing certain temperature levels in global warming studies, exceeding regulated pollution levels in the atmosphere or in water, occurrences of earthquakes, volcanic eruptions, stock market corrections, eplileptic seizures, etc. The binary process $\{Y_t\}$ is, typically, a correlated sequence of Bernolli random variables.

Two related broad classes of models for binary data are: (a) regression type models and (b) partial likelihood models. The regression type approach seeks to model the conditional expectation of Y_t given the past data $(Y_{t-1}, Y_{t-2}, ..., Y_1)$ and possible non-random covariates $(Z_{t-1}, Z_{t-2}, ...Z_1)$. In some situations, the data $\{Y_t\}$ are obtained from another observable random process $\{Z_t\}$ by clipping. For instance, $Y_t = \begin{cases} 1 & \text{if } Z_t > c \\ 0 & \text{otherwise} \end{cases}$ where c is a given threshold. Then the above regression model for $\{Y_t\}$ with random covariates $(Z_{t-1}, Z_{t-2}, ..., Z_1)$ can be used to construct a so-called partial likelihood .

In this chapter, we are concerned with the second model where $\{Y_t\}$ are obtained by clipping a Gaussian autoregressive process. We present five alternative methods of estimation and compare these methods by simulation.

We present the basic model in Section 3.2. The estimation methods are discussed in Section 3.3. Section 3.4 is concerned with the simulation study to compare the estimates. A real data analysis is presented in Section 3.5.

3.2 Binary Models Generated by a Gaussian Autoregressive Process

Suppose $\{Z_t\}$ is a stationary Gaussian process with $E(Z_t) = 0$, $var(Z_t) = \sigma_z^2$ and $cov(Z_t, Z_{t+k}) = \gamma_z(k)$. Note that $\gamma_z(0) = \sigma_z^2$. Also, $\rho_z(k) = corr(Z_t, Z_{t+k}) = \frac{\gamma_z(k)}{\gamma_z(0)}$.

Define a binary process $\{Y_t\}$ by

$$Y_t = \begin{cases} 1 \quad if \quad Z_t > 0, \\ 0 \quad \text{otherwise.} \end{cases} \tag{3.2}$$

We then have $E(Y_t) = P(Y_t = 1) = P(Z_t > 0) = \frac{1}{2}$, $E(Y_t^2) = P(Y_t = 1) = \frac{1}{2}$, and $var(Y_t) = \gamma_y(0) = \frac{1}{4}$; $cov(Y_t, Y_{t+k}) = \gamma_y(k) = E(Y_t Y_{t+k}) - \frac{1}{4}$. Also, $\{Y_t\}$ is a strictly stationary process.

Now, $E(Y_t Y_{t+k}) = P(Z_t > 0, Z_{t+k} > 0) = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} \rho_z(k)$ (Kedem 1980). Hence, $\gamma_y(k) = E(Y_t Y_{t+k}) - \frac{1}{4} = \frac{1}{2\pi} \sin^{-1} \rho_z(k)$, thus $\rho_y(k) = \frac{\gamma_y(k)}{\gamma_y(0)} = \frac{2}{\pi} \sin^{-1} \rho_z(k)$, and we then have the relationship

$$\rho_z(k) = \sin(\frac{\pi}{2} \rho_y(k)) \tag{3.3}$$

Based on the sample $(y_1, ..., y_n)$ only, we can estimate $\rho_z(k)$ by

$$\hat{\rho}_z(k) = \sin(\frac{\pi}{2} \hat{\rho}_y(k)), \tag{3.4}$$

where $\hat{\rho}_y(k) = \frac{\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}, k = 1, 2, ...$

Autoregressive process: AR(p)

Suppose $\{Z_t\}$ is an AR(p) process satisfying

$$Z_t = \beta_1 Z_{t-1} + \beta_2 Z_{t-2} + ... + \beta_p Z_{t-p} + \epsilon_t, \tag{3.5}$$

where $\{\epsilon_t\}$ are i.i.d $N(0,\sigma^2)$ random errors. Assume that all the roots of the equation $1-\phi_1\xi-\phi_2\xi^2-...-\phi_p\xi^p=0$ are larger than 1 in absolute value. Then $\{Z_t\}$ is a stationary Gaussian process with mean zero and autocorrelation function $\rho_z(k)$, $k=1,...,p$ determined by the linear Yule-Walker equation :

$$\begin{pmatrix} \rho_z(1) \\ \rho_z(2) \\ . \\ . \\ \rho_z(p) \end{pmatrix} = \begin{pmatrix} 1 & \rho_z(1) & \rho_z(2) & ... & \rho_z(p-1) \\ \rho_z(1) & 1 & \rho_z(1) & ... & \rho_z(p-2) \\ . & & & & \\ . & & & & \\ \rho_z(p-1) & \rho_z(p-2) & \rho_z(p-3) & ... & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}.$$

This gives the relation

$$\beta = R_z^{-1}\rho_z \tag{3.6}$$

where $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}$, $\rho_z = \begin{pmatrix} \rho_z(1) \\ \rho_z(2) \\ . \\ . \\ \rho_z(p) \end{pmatrix}$, and $R_z = \begin{pmatrix} 1 & \rho_z(1) & \rho_z(2) & ... & \rho_z(p-1) \\ \rho_z(1) & 1 & \rho_z(1) & ... & \rho_z(p-2) \\ . & & & & \\ . & & & & \\ \rho_z(p-1) & \rho_z(p-2) & \rho_z(p-3) & ... & 1 \end{pmatrix}$.

An estimate of β based on $(y_1,...,y_n)$ is given by

$$\hat{\beta}_y = \hat{R_z}^{-1}\hat{\rho}_z \tag{3.7}$$

where $\rho_z(k)$, $k=1,...,p$ in (3.6) are replaced by the estimates $\hat{\rho}_z(k)=\sin(\frac{\pi}{2}\hat{\rho}_y(k))$ with $\hat{\rho}_y(k)=\frac{\sum_{t=1}^{n-k}(y_t-\bar{y})(y_{t+k}-\bar{y})}{\sum_1^n(y_t-\bar{y})^2}$,$k=1,...,p$.

For AR(1), we have $\hat{\beta}_y=\hat{\rho}_z(1)=\sin(\frac{\pi}{2}\hat{\rho}_y(1))$, where $\hat{\rho}_y(1)=\frac{\sum_1^{n-1}(y_t-\bar{y})(y_{t+1}-\bar{y})}{\sum_1^n(y_t-\bar{y})^2}$; while from the original sample $(z_1,...z_n)$, we have $\hat{\beta}_z=\hat{\rho}_z(1)=\frac{\sum_1^{n-1}(z_t-\bar{z})(z_{t+1}-\bar{z})}{\sum_1^n(z_t-\bar{z})^2}$.

3.3 Methods of Estimation

Let $\{Z_t\}$, t=0,$\pm1,\pm2,...,\pm n$, be an autoregressive process of order p defined by (3.5). Here we take $\sigma^2=1$ for simplicity. Also, define $\{Y_t\}$ as in (3.2).

Method 1:Estimation based on $\{Y_t\}$

An estimate of β based on the sample $(y_0, y_1, ..., y_n)$ is given by the Yule-Walker equation $\hat{\beta}_y = \hat{R_z}^{-1}\hat{\rho}_z$, where $\hat{\rho}_z(k) = \sin(\frac{\pi}{2}\hat{\rho}_y(k))$ for $k = 1, ..., p$ and $\hat{\rho}_y(k) = \frac{\sum_1^{n-k}(y_t-\bar{y})(y_{t+k}-\bar{y})}{\sum_1^n(y_t-\bar{y})^2}$.

Method 2: Markov chain assumption

For the AR(p) model mentioned above, we know that $P(Z_t > 0) = P(Z_t \leq 0) = \frac{1}{2}$. Also we have $P(Z_{(k+t)} > 0, Z_t > 0) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho_z(k)$. Therefore, for $k = 1, ..., p$,

$$\begin{aligned} P(Y_{t+k} = 1|Y_t = 1) = P(Z_{t+k} > 0|Z_t > 0) &= \frac{P(Z_{t+k} > 0, Z_t > 0)}{P(Z_t > 0)} \\ &= \frac{\frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho_z(k)}{\frac{1}{2}} = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\rho_z(k). \end{aligned} \tag{3.8}$$

Furthermore, $P(Z_t > 0|Z_s > 0) = P(Z_t \leq 0|Z_s \leq 0)$ by the symmetry property of Gaussian assumption. We define $\theta_k = P(Y_{t+k} = 1|Y_t = 1) = P(Y_{t+k} = 0|Y_t = 0)$. Then suppose $\{Y_t\}$ is a Markov chain with the k-step transition matrix:

$$P_{t,t+k} = \begin{pmatrix} \theta_k & 1-\theta_k \\ 1-\theta_k & \theta_k \end{pmatrix}.$$

The likelihood function conditional on y_0 is given by

$$\begin{aligned} L = \prod_{t=k}^{n}\prod_{k=1}^{p} p(Y_t|Y_{t-k}) &= \prod_{i,j}\prod_{k=1}^{p} p_{i,j}(k)^{n_{i,j}(k)} \\ &= \prod_{k=1}^{p} \theta_k^{(n_{00}(k)+n_{11}(k))}(1-\theta_k)^{(n_{10}(k)+n_{01}(k))}, \end{aligned}$$

where $n = n_{00}(k) + n_{11}(k) + n_{01}(k) + n_{10}(k)$, $p_{i,j} = p(Y_t = j|Y_{t-k} = i)$, $n_{i,j}$=number of transitions "$i \to j$" after k steps in the sample $(y_0, y_1, ..., y_n)$, and $n = n_{00} + n_{11} + n_{01} + n_{10}$. The ML estimator of θ_k is $\hat{\theta}_k = \frac{n_{00}+n_{11}}{n}$ and if we let $\hat{\alpha}_k = 1 - \hat{\theta}_k$, we have $\hat{\alpha}_k = \frac{n_{01}+n_{10}}{n} = \frac{N_k(n)}{n}$, where $N_k(n)$ = number of state changes (0-1 and 1-0) after $k$steps of transition. Then $N_k(n)$ is approximately distributed as $Bin(n, \alpha_k)$. We also can show that $\sqrt{n}(\hat{\alpha}_k - \alpha_k) \xrightarrow{d} N(0, \alpha_k(1-\alpha_k))$.

Now, from the equation $\theta_k = p(Y_{t+k} = 1|Y_t = 1) = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\rho_z(k) \quad k = 1, ..., p$, we have that $\rho_z(k) = \sin(\theta_k\pi - \frac{\pi}{2}) = \cos(\alpha_k\pi)$. Using the estimate $\hat{\alpha}_k$ in place of α_k, we get the estimated $\hat{\rho}_z(k)$ and thus the $\hat{\beta}_y$ through the Yule-Walker equation $\hat{\beta}_y = \hat{R}_z^{-1}\hat{\rho}_z$.

It may be noted that $\{Y_t\}$ is, in general, not a Markov chain. The Markov assumption above is used only as an approximation.

Method 3: Estimation based on $\{Z_t\}$

Here, we directly estimate from observation $\{Z_t\}$ by applying the Yule-Walker equation $\hat{\beta}_z = \hat{R}_z^{-1}\hat{\rho}_z$ where $\hat{\rho}_z(k) = \frac{\sum_1^{n-k}(z_t-\bar{z})(z_{t+k}-\bar{z})}{\sum_1^n(z_t-\bar{z})^2}$. It is known that $\hat{\beta}_z$ is asymptotically equivalent to the ML estimate of β obtained from:

$$L_z(\beta) = \prod_{t=1}^n f_\epsilon(Z_t - \sum_{j=1}^p \beta_t Z_{t-j})$$

where f_ε is the standard normal density function.

Method 4: Partial likelihood

Define $p_t(\beta) = P_\beta(Y_t = 1|Z_{(t-1)})$, where $Z_{(t-1)} = (Z_{t-1}, ..., Z_{t-p})$. Then

$$p_t(\beta) = P(\sum_{j=1}^p \beta_j z_{t-j} + \epsilon_t > 0) = P(\epsilon_t > -\sum_{j=1}^p \beta_j z_{t-j}) = \Phi(\sum_{j=1}^p \beta_j z_{t-j})$$

where Φ is the probability distribution function of a standard normal distribution. We also know for the binary $\{Y_t\}$

$$p_t(y_t;\beta) = P_\beta(Y_t = y_t|Z_{(t-1)}) = [p_t(\beta)]^{y_t}[1 - p_t(\beta)]^{(1-y_t)}.$$

It is easy to see that the Partial Likelihood function is:

$$PL(\beta) = \prod_{t=1}^n(\Phi(\sum_{j=1}^p \beta_j Z_{t-j})^{y_t}(1 - \Phi(\sum_{j=1}^p \beta_j Z_{t-j}))^{(1-y_t)}.$$

The maximum partial likelihood estimator (MPLE) of β is then obtained by maximizing $PL(\beta)$. See ,for instance, Slud and Kedem (1994) and Wong (1986) for details on the concept of partial likelihood.

Method 5: Complete Likelihood

This approach is based on the joint distribution of observations $\{Z_t, Y_t\}$. The joint density function involving both $\{Z_t\}$ and $\{Y_t\}$ can be written as: $p_\beta(Z_t, Y_t) = p_\beta(Y_t|Z_t)p_\beta(Z_t)$. Correspondingly, the full likelihood function is simply

$$\begin{aligned} L(Z, Y; \beta) &= PL(\beta)L_z(\beta) \\ &= \prod_{t=1}^{n}(\Phi(\sum_{j=1}^{p}\beta_j Z_{t-j})^{y_t}(1-\Phi(\sum_{j=1}^{p}\beta_j Z_{t-j}))^{(1-y_t)} f_\epsilon(Z_t - \sum_{j=1}^{p}\beta_t Z_{t-j}), \end{aligned} \tag{3.9}$$

where $PL(\beta)$ and $L_z(\beta)$ are defined as in methods 4 and 3 respectively.

3.4 Simulation Results

In order to evaluate the performance of different methods mentioned above, we applied each of them for the special case AR(1) : $Z_t = \beta Z_{t-1} + \epsilon_t$, where $\{\epsilon_t\}$ are i.i.d N(0,1). Denote $\hat{\beta}(i)$= estimator of β based on method $i, i = 1, 2, 3, 4, 5$.

We simulated n observations of $\{Z_t\}$ from AR(1) time series with parameter β taking different values below:

(i) n=50,100,200 ;

(ii) $\beta = 0, \pm 0.3, \pm 0.5, \pm 0.7$;

In each case, We repeated simulation N=200 times and then computed the means and standard deviation of $\hat{\beta}(i)$. The results are summarized in Table 3.1 below.

For better comparison, we also calculated the relative efficiency of $\hat{\beta}(i)$ i=1,2,3,4 with respect to $\hat{\beta}(5)$ as follows: *Relative Efficiency* of ($\hat{\beta}(i)$,$\hat{\beta}(5)$) =$(\frac{s.d.\hat{\beta}(5)}{s.d.\hat{\beta}(i)})^2$, and the results appear in Table 3.2.

It may be noted that all the five estimators perform reasonably well and the efficiency increases for large n. Not surprisingly, method 3 based on the original autoregressive process $\{Z_t\}$ performs better than the other methods for most choices of β and n. The estimator $\hat{\beta}_5$ based on the complete likelihood is almost as good as $\hat{\beta}_3$. Methods 1 and 2 based only on $\{Y_t\}$ have similar performance but they are not as good as $\hat{\beta}_3$ and $\hat{\beta}_5$ for obvious reasons

Table 3.1: Mean,standard deviation and MSE results

β			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
n=50	$\hat{\beta}_1$	mean	-0.672	-0.466	-0.276	-0.018	0.232	0.376	0.609
		std	0.165	0.196	0.190	0.214	0.218	0.194	0.190
		MSE	0.028	0.039	0.036	0.046	0.052	0.053	0.044
	$\hat{\beta}_2$	mean	-0.636	-0.406	-0.216	0.059	0.291	0.488	0.685
		std	0.173	0.176	0.215	0.212	0.189	0.198	0.167
		MSE	0.034	0.040	0.053	0.048	0.036	0.039	0.028
	$\hat{\beta}_3$	mean	-0.673	-0.478	-0.300	-0.013	0.266	0.420	0.630
		std	0.106	0.127	0.116	0.132	0.129	0.126	0.111
		MSE	0.012	0.016	0.013	0.018	0.018	0.022	0.017
	$\hat{\beta}_4$	mean	-0.675	-0.449	-0.243	0.015	0.270	0.444	0.678
		std	0.178	0.176	0.159	0.178	0.181	0.186	0.158
		MSE	0.032	0.034	0.028	0.032	0.033	0.038	0.025
	$\hat{\beta}_5$	mean	-0.680	-0.468	-0.277	0.011	0.295	0.463	0.690
		std	0.107	0.132	0.122	0.139	0.136	0.132	0.103
		MSE	0.012	0.018	0.015	0.019	0.019	0.019	0.011
n=100	$\hat{\beta}_1$	mean	-0.675	-0.473	-0.271	-0.022	0.269	0.431	0.622
		std	0.114	0.147	0.158	0.147	0.153	0.148	0.133
		MSE	0.015	0.022	0.026	0.022	0.024	0.026	0.024
	$\hat{\beta}_2$	mean	-0.652	-0.428	-0.228	0.017	0.287	0.496	0.705
		std	0.120	0.147	0.146	0.165	0.155	0.144	0.113
		MSE	0.017	0.027	0.026	0.027	0.024	0.021	0.013
	$\hat{\beta}_3$	mean	-0.680	-0.492	-0.299	-0.01	0.287	0.468	0.657
		std	0.067	0.092	0.096	0.095	0.094	0.094	0.075
		MSE	0.005	0.009	0.009	0.009	0.009	0.010	0.007
	$\hat{\beta}_4$	mean	-0.628	-0.438	-0.242	0.008	0.277	0.446	0.650
		std	0.124	0.130	0.122	0.121	0.126	0.132	0.120
		MSE	0.020	0.021	0.018	0.015	0.016	0.020	0.017
	$\hat{\beta}_5$	mean	-0.675	-0.478	-0.282	0.010	0.304	0.490	0.691
		std	0.071	0.097	0.100	0.098	0.098	0.099	0.074
		MSE	0.006	0.010	0.010	0.010	0.010	0.010	0.006
n=200	$\hat{\beta}_1$	mean	-0.677	-0.467	-0.268	-0.018	0.260	0.454	0.667
		std	0.082	0.100	0.101	0.101	0.115	0.106	0.084
		MSE	0.007	0.011	0.011	0.011	0.015	0.013	0.008
	$\hat{\beta}_2$	mean	-0.656	-0.461	-0.236	0.031	0.282	0.479	0.702
		std	0.096	0.105	0.107	0.107	0.110	0.109	0.082
		MSE	0.011	0.012	0.015	0.012	0.012	0.012	0.007
	$\hat{\beta}_3$	mean	-0.691	-0.493	-0.300	-0.009	0.288	0.494	0.685
		std	0.051	0.061	0.064	0.065	0.072	0.063	0.054
		MSE	0.003	0.004	0.004	0.004	0.005	0.004	0.003
	$\hat{\beta}_4$	mean	-0.637	-0.425	-0.237	0.004	0.256	0.449	0.654
		std	0.092	0.085	0.078	0.077	0.087	0.094	0.090
		MSE	0.012	0.013	0.010	0.006	0.010	0.011	0.010
	$\hat{\beta}_5$	mean	-0.684	-0.477	-0.281	0.001	0.293	0.499	0.701
		std	0.054	0.064	0.064	0.067	0.075	0.067	0.054
		MSE	0.003	0.005	0.004	0.004	0.006	0.004	0.003

Table 3.2: Relative efficiency:

β		-0.7	-0.5	-0.3	0	0.3	0.5	0.7
n=50	$\hat{\beta}_1$	0.421	0.454	0.412	0.422	0.389	0.463	0.294
	$\hat{\beta}_2$	0.383	0.563	0.322	0.430	0.518	0.444	0.380
	$\hat{\beta}_3$	1.019	1.080	1.106	1.109	1.111	1.098	0.861
	$\hat{\beta}_4$	0.361	0.563	0.589	0.610	0.656	0.504	0.425
n=100	$\hat{\beta}_1$	0.388	0.435	0.401	0.444	0.410	0.447	0.310
	$\hat{\beta}_2$	0.350	0.435	0.469	0.353	0.400	0.473	0.429
	$\hat{\beta}_3$	1.123	1.112	1.085	1.064	1.087	1.109	0.974
	$\hat{\beta}_4$	0.328	0.557	0.672	0.656	0.605	0.563	0.380
n=200	$\hat{\beta}_1$	0.434	0.410	0.402	0.440	0.425	0.400	0.413
	$\hat{\beta}_2$	0.316	0.372	0.358	0.392	0.465	0.378	0.434
	$\hat{\beta}_3$	1.121	1.101	1.000	1.062	1.085	1.031	1.000
	$\hat{\beta}_4$	0.345	0.567	0.673	0.757	0.743	0.508	0.360

(loss of information). Finally, the partial likelihood estimator $\hat{\beta}_4$ is better than $\hat{\beta}_1$ and $\hat{\beta}_2$, but it is not as good as $\hat{\beta}_3$ and $\hat{\beta}_5$. We can therefore rank in terms of efficiency (preference): ($\hat{\beta}_3$,$\hat{\beta}_5$), $\hat{\beta}_4$,($\hat{\beta}_1$,$\hat{\beta}_2$).

3.5 Data Analysis

We collected IBM stock daily price data from 2005-2006 and denoted the original data as $\{X_t\}$, $t = 1, 2, ..., 288$. Also let $Z_t = X_t - \overline{X}$, where $\overline{X} = \frac{1}{288}\sum_{t=1}^{288} X_t$, that is Z_t is the mean-centered original data.

First, we apply method 3 where the parameter estimation is based on $\{Z_t\}$ for the first 200 observations.

(i) The plot of Z_t vs. t shows that the stationarity is violated with non-constant mean; therefore, the first difference is performed and the new data (defined as C_t) plot satisfies the stationarity.(see Fig.1)

(ii) For the differenced data, the autocorrelation function (ACF) gradually decays after lag

3 while the partial autocorrelation function (PACF) plot cuts off after lag 3. So we apply the model $C_t - \phi C_{t-3} = w_t$ where w_t is $\mathrm{N}(0, \sigma^2)$ for the new data

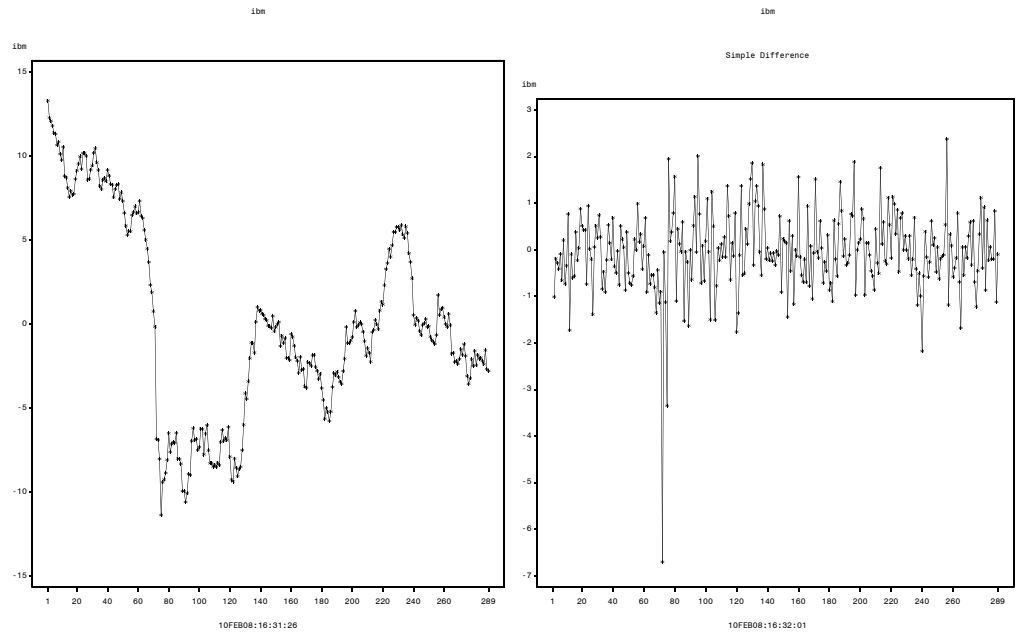

Figure 3.1: Original IBM data plot and after first difference plot

(iii) After fitting the model, we also check the residuals and it shows that the residuals do follow white noise.

All the above steps are performed by SAS procedure PROC ARIMA. The final best model for $\{Z_t\}$ is :

$$Z_t = Z_{t-1} + 0.2075(Z_{t-3} - Z_{t-4}) + \epsilon_t, \tag{3.10}$$

where $\{\epsilon_t\}$ follows *iid* N(0,0.81776).

Now, we do data-clipping below:

$$Y_t = \begin{cases} 1 & if \quad Z_t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Define $p_t(\beta) = P_\beta(Y_t = 1 | Z_{(t-1)})$, where $Z_{(t-1)} = (Z_{t-1}, ..., Z_{t-p})$. Then

$$p_t(\hat{\beta}) = P(\epsilon_t > -[Z_{t-1} + 0.2075(Z_{t-3} - Z_{t-4})]) = \Phi(\tfrac{[Z_{t-1} + 0.2075(Z_{t-3} - Z_{t-4})]}{\sqrt{0.81776}})$$

where Φ is the probability distribution function of a standard normal distribution. We then estimate the prediction probabilities for the last 88 observations. The results are listed in Table 3.3, where * indicates that the predicted and observed are not consistent. It shows that about 83% predictions are correct; one interesting point is that whenever there are state-changes, our prediction results seem often to have one-lag delay.

Table 3.3: The observed price increases(Y_t)and predicted probability of price increases ($\hat{p}_t$)

$\hat{p}_t$	0.07	0.17*	0.56	0.84*	0.49	0.54	0.45	0.51	0.29	0.10	0.01
Y_t	0	1	1	0	0	1	0	0	0	0	0
$\hat{p}_t$	0.03	0.01	0.00	0.25	0.28*	0.76*	0.50	0.40*	0.82	0.94	0.95
Y_t	0	0	0	0	1	0	0	1	1	1	1
$\hat{p}_t$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y_t	1	1	1	1	1	1	1	1	1	1	1
$\hat{p}_t$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.69*	0.38*	0.45*
Y_t	1	1	1	1	1	1	1	1	0	1	1
$\hat{p}_t$	0.53*	0.34	0.19	0.41*	0.49*	0.70*	0.42	0.46	0.15	0.13	0.07
Y_t	0	0	0	1	1	0	0	0	0	0	0
$\hat{p}_t$	0.06	0.20*	0.98	0.79	0.95	0.81	0.71*	0.50	0.36*	0.74*	0.44
Y_t	0	1	1	1	1	1	0	0	1	0	0
$\hat{p}_t$	0.02	0.01	0.00	0.00	0.00	0.01	0.03	0.02	0.09	0.01	0.00
Y_t	0	0	0	0	0	0	0	0	0	0	0
$\hat{p}_t$	0.00	0.00	0.00	0.00	0.05	0.00	0.02	0.00	0.01	0.00	0.00
Y_t	0	0	0	0	0	0	0	0	0	0	0

Chapter 4

State Space Models for Categorical Time Series

4.1 Introduction

Models for categorical time series have been discussed by several authors including Fahrmeir and Kaufmann (1987), Kedem (1980), Kedem and Fokianos (2002), Fokianos and Kedem (1998, 2003), among others. Suppose $\{Y_t\}$, t=1,2..., is a (mx1) vector time series with $Y_t = (Y_{t1}, ..., Y_{tm})'$, and

$$Y_{tj} = \begin{cases} 1 & \text{when jth category is observed} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, ..., m,$$

m denoting the number of categories. We then refer to $\{Y_t\}$ as a categorical time series. Denote the set of past history and possible covariates by the σ-field :

$$\mathscr{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, ..., x_{t-1}, x_{t-2}, ...x_1);$$

where $\{x_{t-1}, ..., x_1\}$ are the covariates which are (usually) non-random. The model for $\{Y_t\}$ is then specified by the probabilities

$$\pi_{tj} = E(Y_{tj}|\mathscr{F}_{t-1}) = p(Y_{tj} = 1|\mathscr{F}_{t-1}).$$

Typically, $\{\pi_{tj}\}$ are assumed to be functions of a parameter vector β which needs to be estimated from the data. The likelihood function based on the data $\{Y_t\}$, $t = 1, ..., N$, is given by

$$L(\beta) = \prod_{t=1}^{N} \prod_{j=1}^{m} \pi_{tj}^{y_{tj}} \tag{4.1}$$

Most of the current literature on modeling categorical time series is based on the regression concept. A generalized linear model for π_{tj} is of the type

$$g(\pi_{tj}) = Z'_{t-1}(j)\beta \tag{4.2}$$

where $g(.)$ is a given link function and $Z_{t-1}(j)$ is a vector of predictor variables which are functions of $\mathscr{F}_{t-1}$. Fokianos and Kedem (2003) give an extensive review of the regression models. See, also Kedem and Fokianos (2002. Fahrmeir and Kaufmann (1987) discuss Markovian regression models where the predictor variables $\{Z_{t-1}(j)\}$ depend only on the past p observations, and covariates, viz,$\{Y_{t-1}, ..., Y_{t-p}, x_{t-1}, ..., x_{t-p}\}$. If, on the other hand, $\{Z_{t-1}(j)\}$ depend on the entire past $\mathscr{F}_{t-1}$, $\{Y_t\}$ is not necessarily a Markov chain. Moreover, if the covariates $\{x_t\}$ are random whose joint distributions depend on the parameter β , the L(β) in (5.1) can be interpreted as a partial likelihood . See Fokianos and Kedem (1998,2003) for further details on partial likelihood. An alternative formulation for modeling the categorical time series $\{Y_t\}$ is via the state space models. The state space formulation typically contains two models: (i) an observation model which is usually specified by the conditional density of Y_t given an unobserved state process (or a latent variable) $\{\beta_t\}$ and (ii) the state model which specifies the probability structure for $\{\beta_t\}$. If the state process is specified by a transition density of β_t given, say, $(\beta_{t-1}, \beta_{t-2}, ..., \beta_{t-p})$, the state space model is said to be a state-driven model. If, on the other hand, the state model for $\{\beta_t\}$ is specified by the conditional density of β_t given the past data $\{Z_{t-1}\}$, the model is referred to as an observation-driven model. Our main goal in this chapter is to present some new observation-driven state space models.

We present the background for the state-driven state space model in section 4.2. our new observation-driven models are discussed in section 4.3. Section 4.4 contains a real data application to the analysis of IBM stock daily prices and Dow Jones Index in the form of binary data. As an application of multi-categorical time series, we present a DNA data analysis in section 4.5.

4.2 State Space Models: state-driven models

State space models relate time series observations $\{Y_t\}$ to unobserved states $\{\beta_t\}$by an observation model for Y_t given β_t and a transition model for $\{\beta_t\}$. The *observation model* is specified by conditional densities

$$p(y_t|\beta_t^*, y_{t-1}^*, x_t^*),\ t=1,2,...$$

the *transition model* is defined by transition densities

$$p(\beta_t|\beta_{t-1}^*, y_{t-1}^*, x_t^*)$$

where $y_t^* = (y_1, ..., y_t)'$, $x_t^* = (x_1, ..., x_t)'$, $\beta_t^* = (\beta_0, \beta_1, ..., \beta_t)'$ denote histories of responses, covariates and a sequence of unobserved states or parameter vector up to time t and an initial density $p(\beta_0)$ is assumed known. We also denote $Z_t = Z_t(x_t^*, y_{t-1}^*)$ as the design matrix, which is predetermined since it is known when y_t is observed. More common but less general, transition densities are assumed to be independent of y_{t-1}^*, x_t^*, i.e. they are defined by $p(\beta_t|\beta_{t-1})$. Also, linear transition equation is the most important specification of transition model, such as $\beta_t = T_t\beta_{t-1} + v_t$, t=1,2,... where the error term v_t has a Gaussian or Non-Gaussian density $f(v)$, with $E(v) = 0$, thus the transition density is given by

$$p(\beta_t|\beta_{t-1}) = f(\beta_t - T_t\beta_{t-1}), t = 1, 2, ... \quad .$$

To specify the bivariate process $\{\ Y_t\ ,\ \beta_t\ \}$ completely in terms of genuine joint densities, additional basic assumptions are required:

A1. Conditional on β_t and (y_{t-1}^*, x_t^*), current observation y_t are independent of β_{t-1}^*, i.e.

$$p(y_t|\beta_t^*, y_{t-1}^*, x_t^*) = p(y_t|\beta_t, y_{t-1}^*, x_t^*),\ t=1,2,... \quad .$$

A2. The state process is conditionally Markovian, i.e.

$$p(\beta_t|\beta_{t-1}^*, x_t^*, y_{t-1}^*) = p(\beta_t|\beta_{t-1}),\ t=1,2,... \quad .$$

It will be assumed for simplicity that the covariates $\{x_t\}$ are non-random. The primary goal is to estimate β_t given the observations $y_1, ..., y_T$. This is termed

(a) filtering for $t = T$

(b) smoothing for $t < T$

(c) prediction for $t > T$.

Using A1 and A2, we obtain the posterior density

$$p(\beta_t^*|y_t^*, x_t^*) \propto \prod_{s=1}^{t} p(y_s|\beta_s, y_{s-1}^*, x_s^*) \prod_{s=1}^{t} p(\beta_s|\beta_{s-1}).$$

Maximization of the conditional density is thus equivalent to maximizing the log-posterior

$$PL(\beta_t^*) = \sum_{s=1}^{t} l_s(\beta_s) + a_0(\beta_0) + \sum_{s=1}^{t} a_s(\beta_s, \beta_{s-1}),$$

where $l_s(\beta_s) = \log p(y_t|\beta_t, y_{t-1}^*, x_t^*)$ is the conditional loglikelihood distribution of y_t; and loglikelihood for the transition model after suppressing y_{t-1}^*, x_t^* is denoted by $a_t(\beta_t, \beta_{t-1}) = logp(\beta_t|\beta_{t-1})$, $a_0(\beta_0) = logp(\beta_0)$. The criterion PL can be interpreted as a penalized loglikelihood, with the sum of the log-prior a_s as the roughness penalty. Numerical maximization of the penalized log-likelihood can be achieved by various algorithms.

Fahrmeir (1992a) suggests the generalized extended Kalman filter and smoother as an approximate posterior mode estimator in dynamic generalized linear models. In addition, Fahrmeir & Kaufmann (1991) also show that this method can be considered as a simplifying approximation of the Fisher scoring iterations. Fahrmeir (1992) recommends the following procedure for common categorical response models: First apply the extended Kalman filter and smoother, then use it as the initial solution for the Gaussian-Newton iterations.

For categorical time series$\{Y_t\}$with$m$categories for each observation, setting $q = m - 1$, we define the observation vector $Y_t = (Y_{t1}, ..., Y_{tq})'$ by

$$Y_{tj} = \begin{cases} 1 & \text{category i has been observed,} \\ 0 & \text{otherwise,} \end{cases} \qquad j = 1, ..., q.$$

Correspondingly $\pi_t = (\pi_{t1}, .., \pi_{tq})'$ denotes the q-vector of conditional probabilities $\pi_{tj} = P(y_{tj} = 1|\beta_t, x_t^*, y_{t-1}^*),\ j = 1, ..., q$. The following is the general specification of logistic observation models for individual univariate categorical time series:

$$\pi_{tj} = \frac{\exp(z_t'\beta_{tj})}{1+\sum_{k=1}^{q}\exp(z_t'\beta_{tk})} = h(Z'\beta_t)$$

where $\{z_t\}$ is function of past observation y_{t-1}^* and non-random exogenous variables x_t^*; Additionally, we supplement a Markovian transition model for $\{\beta_t\}$, i.e. by specifying a transition density $p(\beta_t|\beta_{t-1})$. Note also that

$$y_{sm} = 1 - \textstyle\sum_{j=1}^{m-1} y_{sj},\ \ \pi_{sm} = 1 - \sum_{j=1}^{m-1}\pi_{sj}.$$

The conditional log-likelihood contribution of y_t is

$$l_t(\beta_t) = \log p(y_t|\beta_t, y_{t-1}^*, x_t^*) = \textstyle\sum_{s=1}^{t}\sum_{j=1}^{m} y_{sj} log \pi_{sj}.$$

The transition model may be chosen as the autoregressive process $\beta_t = \phi\beta_{t-1} + \varepsilon_t$. The transition density is thus given by $p(\beta_t|\beta_{t-1}) = f(\beta_t - \phi\beta_{t-1})$, where $f(\epsilon)$ is taken as the normal density function. For state estimation, we can use the PL criterion discussed above.

Suppose we are interested in estimating the parameter ϕ in the transition density. The marginal likelihood function is given by

$$L_t(\phi) = \int [\prod_{s=1}^{t}\prod_{j=1}^{m} \pi_{sj}^{y_{sj}}(\beta_s) f(\beta_s - \phi\beta_{s-1})] d\beta_1 ... d\beta_t. \tag{4.3}$$

Since $L_t(\phi)$ is a multiple integral with increasing (with t) dimension, the problem of ML estimation of ϕ becomes unwieldly, if not impossible. We shall, therefore, introduce in the next section, an alternative observation-driven state space model which is much simpler for ML estimation.

4.3 Observation-driven State Space Models

If the transition density of the state parameter depends upon possible covariates and past responses, the model is said to be 'observation-driven'. Let

$$p(\beta_t|\beta^*_{t-1}, y^*_{t-1}, x^*_t) = p(\beta_t|y^*_{t-1}, x^*_t) = p(\beta_t|Z_t) \ ,$$

where Z_t contains past responses and possible covariates.

In the following, we present'observation-driven'state space models for categorical time series.

4.3.1 Binary Time Series (Conjugate State Process)

Consider a binary process $\{Y_t\}$ where Y_t takes value 0 or 1.

Observation model:

$p(y_t|\pi^*_t, y^*_{t-1}, x^*_t) = p(y_t|\pi_t) = \pi_t^{y_t}(1-\pi_t)^{1-y_t}, y_t = 0, 1.$

Transition model (later refer to State model):

$p(\pi_t|\pi^*_t, y^*_{t-1}, x^*_t) = p(\pi_t|y^*_{t-1}, x^*_t) = p(\pi_t|Z_t) = \frac{\pi_t^{\mu_t-1}(1-\pi_t)^{(1-\mu_t)-1}}{B(\mu_t, 1-\mu_t)}$, $0 < \pi_t < 1$,

where Z_t is a (px1) vector whose components are functions of (y^*_{t-1}, x^*_t), β is a (px1) vector of unknown parameters and $logit(\mu_t)(= log(\frac{\mu_t}{1-\mu_t})) = Z_t'\beta$. Notice that the conditional distribution of π_t is $Beta(\mu_t, 1-\mu_t)$ with $E(\pi_t|Z_t) = \mu_t$ and $Var(\pi_t|Z_t) = \mu_t(1-\mu_t)/2$.

It is easy to verify that

$$\begin{aligned} p(y_t|Z_t) &= \int_0^1 p(y_t, \pi_t|Z_t)d\pi_t = \int_0^1 p(y_t|\pi_t)p(\pi_t|Z_t)d\pi_t \\ &= \frac{B(\mu_t + y_t, 1-\mu_t-y_t)}{B(\mu_t, 1-\mu_t)} = \mu_t^{y_t}(1-\mu_t)^{1-y_t}, y_t = 0, 1. \end{aligned} \tag{4.4}$$

We have $E(y_t|Z_t) = \mu_t, Var(y_t|Z_t) = \mu_t(1-\mu_t)$.

The likelihood function is thus given by

$$L(\beta) = \Pi_{t=1}^n \mu_t^{y_t}(1-\mu_t)^{1-y_t}.$$

The score function $S_n(\beta) = \frac{dlogL(\beta)}{d\beta} = \sum_{t=1}^n (\frac{dlogL}{d\mu_t}\frac{d\mu_t}{d\beta}) = -\sum_{t=1}^n Z_t(y_t - \mu_t)$;

and Fisher information $I_n = -\frac{d^2logL(\beta)}{d\beta d\beta'} = \sum_{t=1}^n Z_t Z_t'\mu_t(1-\mu_t)$.

Under regularity conditions, we have the following theorem:

Theorem:

$$(\Sigma_{t=1}^n Z_t Z_t'\mu_t(1-\mu_t))^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{d} N_p(0, I).$$

Under regularity conditions of Fahrmeir and Kaufmann (1987), their proof applied.

4.3.2 Time Series With Multi-Categories(Conjugate State Process)

We extend the model in 3.1 to the case where the categorical time series $\{Y_t\}$ has $m$categories for each observation. Define

$$Y_{t,j} = \begin{cases} 1 & \text{if category j has been observed,} \\ 0 & \text{otherwise,} \end{cases} \qquad j = 1, ..., m. \tag{4.5}$$

Also let $q = m - 1$, and define $Y_t = (Y_{t,1}, ...Y_{t,q})$, $\pi_t = (\pi_{t,1}, ..., \pi_{t,q})$, where $\pi_{t,j} = p(Y_{t,j} = 1|Z_{t,j})$, $Z_{t,j}$ is a (px1) vector whose components are functions of non-random covariates $x_t^* = (x_t, ..., x_1)$ and past data $y_{t-1}^* = (y_{t-1}, ..., y_1)$.

The observation model is defined by:

$p(y_t|\pi_t^*, y_{t-1}^*, x_t^*) = p(y_t|\pi_t) = (\prod_{j=1}^q \pi_{t,j}^{y_{t,j}})(1 - \sum_{j=1}^q \pi_{t,j})^{1-\sum_{j=1}^q y_{t,j}}$.

The state model is assumed to be a Dirichlet distribution:

$p(\pi_t|\pi_{t-1}^*, y_{t-1}^*, x_t^*) = p(\pi_t|Z_t) = \frac{(\prod_{j=1}^q \pi_{t,j}^{\alpha_{t,j}-1})(1-\sum_{j=1}^q \pi_{t,j})^{\alpha_{t,q+1}-1}}{B(\alpha_{t,1},...,\alpha_{t,q+1})}$.

Now choose $\alpha_{t,j} = \mu_{t,j}, j = 1, ..., q$, and $\alpha_{t,q+1} = 1 - \sum_{j=1}^q \mu_{t,j}$, that is, $\sum_{j=1}^{q+1} \alpha_{t,j} = 1$. We also define $\mu_t = (\mu_{t,1}, ..., \mu_{t,q})'$ and $logit(\mu_t) = (logit(\mu_{t,j}), j = 1, ..., q)'$ where $logit(\mu_{t,j}) = \log(\frac{\mu_{t,j}}{1-\sum_{j=1}^q \mu_{t,j}}) = \log(\frac{\mu_{t,j}}{\mu_{t,q+1}}) = Z_{t,j}'\beta$, We can rewrite it as: $logit(\mu_t) = Z_t'\beta$,where $Z_t = (Z_{t,j}, j = 1, ..., q)$ is a pxq matrix whose jth column is given by $Z_{t,j}$.

From the properties of Dirichlet distribution, we have the following: $E(\pi_{t,j}|Z_t) = \frac{\alpha_{t,j}}{\sum_{j=1}^{q+1} \alpha_{t,j}} = \mu_{t,j}$, $j = 1, ..., q$; $Cov(\pi_{t,j}, \pi_{t,j'}|Z_t) = -\mu_{t,j}\mu_{t,j'}/2, j \neq j'$, and $Var(\pi_{t,j}|Z_t) = \frac{1}{2}\mu_{t,j}(1 - \mu_{t,j})$.

The probability density function $p(y_t|z_t)$ is given by:

$$p(y_t|Z_t) = (\prod_{j=1}^q \mu_{t,j}^{y_{t,j}})(1 - \sum_{j=1}^q \mu_{t,j})^{1-\sum_{j=1}^q y_{t,j}}.$$

The likelihood function is:

$$L_n(\beta) = \prod_{t=1}^n p(y_t|Z_t) = \prod_{t-1}^n \prod_{j=1}^{q+1} \mu_{t,j}^{y_{t,j}},$$

with $y_{t,q+1} = 1 - \sum_{j=1}^q y_{t,j}$; $\mu_{t,q+1} = 1 - \sum_{j=1}^q \mu_{t,j}$.

Correspondingly, the score function is

$$S_n(\beta) = \frac{d \log L_n(\beta)}{d\beta} = -\sum_{t=1}^{n} \sum_{j=1}^{q+1} \mu_{t,j}^{y_{t,j}} = \sum_{t=1}^{n} \sum_{j=1}^{q} Z_{t,j}(y_{t,j} - \mu_{t,j}).$$

We have $E(Y_{t,j}|Z_{t,j}) = \mu_{t,j}$, $Cov(Y_{t,j}, Y_{t,j'}|Z_t) = -\mu_{t,j}\mu_{t,j'}$, for $j \neq j'$ and $Var(Y_{t,j}|Z_t) = \mu_{t,j}(1 - \mu_{t,j})$.

Under regularity conditions, we have

Theorem:

$$(\sum_{t=1}^{n} Z_t \Sigma_t Z_t')^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{d} N_p(0, I),$$

where $Z_t = (Z_{t,1}, ..., Z_{t,q})', \Sigma_t = Cov(Y_{t,1}, ..., Y_{t,q})$. Under regularity conditions of Fahrmeir and Kaufmann (1987), their proof applied.

4.3.3 Binary Model(Non-conjugate State Process)

For a binary process $\{Y_t\}$, we have the following setting:

The observation model is the same as before:

$$p(y_t|\pi_t^*, y_{t-1}^*, x_t^*) = p(y_t|\pi_t) = \pi_t^{y_t}(1 - \pi_t)^{1-y_t},\ y_t = 1, 0.$$

The state model now is assumed to have a general state distribution:

$$p(\pi_t|\pi_t^*, y_{t-1}^*, x_t^*) = p_\beta(\pi_t|Z_t),\ 0 < \pi_t < 1.$$

Then the probability density of $\{Y_t\}$ given $\{Z_t\}$ is :

$$\begin{aligned} p(y_t|Z_t) &= \int_0^1 p(y_t, \pi_t|Z_t)d\pi_t = \int_0^1 p(y_t|\pi_t)p(\pi_t|Z_t)d\pi_t \\ &= \int_0^1 \pi_t^{y_t}(1 - \pi_t)^{1-y_t} p_\beta(\pi_t|Z_t)d\pi_t = \overline{\pi}_t^{y_t}(1 - \overline{\pi}_t)^{1-y_t}, \end{aligned} \tag{4.6}$$

where $E(Y_t|Z_t) = p(Y_t = 1|Z_t) = \int_0^1 \pi_t p_\beta(\pi_t|Z_t)d\pi_t = E(\pi_t|Z_t) = \overline{\pi}_t$, and $Var(Y_t|Z_t) = \overline{\pi}_t(1 - \overline{\pi}_t)$.

The likelihood function then is: $\prod_{t=1}^{n} p(y_t|Z_t) = \prod_{t=1}^{n} \overline{\pi}_t^{y_t}(1 - \overline{\pi}_t)^{1-y_t}$. As a special example, let $\mu_t = logit\pi_t = \log \frac{\pi_t}{1-\pi_t}$, while $\mu_t \sim N(\eta_t = Z_t'\beta, \sigma^2)$. Then

$$p_\beta(\pi_t|Z_t) = \frac{1}{\sqrt{\sigma^2 2\pi}}(\frac{1}{\pi_t(1-\pi_t)})exp[-2\sigma^2(logit\pi_t - \eta_t)^2],\ 0 < \pi_t < 1.$$

We can compute $\overline{\pi}_t$ as

$$\overline{\pi}_t = E(\pi_t|Z_t) = \frac{1}{\sqrt{\sigma^2 2\pi}} \int_{-\infty}^{\infty} (\frac{1}{1-\pi_t}) exp[-2\sigma^2 (logit\pi_t - \eta_t)^2] d\pi_t = E(1+exp(-\mu_t))^{-1}. \tag{4.7}$$

4.3.4 Multi-Categorical Model(Non-conjugate State Process)

Consider the m-categorical model defined in section 3.2. Recall that the observation model is given by :

$$p(y_t|\pi_t^*, y_{t-1}^*, x_t^*) = p(y_t|\pi_t) = (\prod_{j=1}^{q} \pi_{t,j}^{y_{t,j}})(1 - \sum_{j=1}^{q} \pi_{t,j})^{1-\sum_{j=1}^{q} y_{t,j}}.$$

We now consider the general state model:

$$p(\pi_t|\pi_{t-1}^*, y_{t-1}^*, x_t^*) = p_\beta(\pi_t|Z_t).$$

The probability density function of $\{Y_t\}$ conditional on $\{Z_t\}$ is calculated as follows:

$$\begin{aligned} p(y_t|Z_t) &= \int_0^1 ... \int_0^1 p(y_t, \pi_t|Z_t) d\pi_{t,1} ... d\pi_{t,q} \\ &= \int_0^1 ... \int_0^1 (\prod_{j=1}^{q} \pi_{t,j}^{y_{t,j}})(1 - \sum_{j=1}^{q} \pi_{t,j})^{1-\sum_{j=1}^{q} y_{t,j}} p_\beta(\pi_t|Z_t) d\pi_{t,1} ... d\pi_{t,q} \\ &= \int_0^1 ... \int_0^1 (\prod_{j=1}^{q} \pi_{t,j}^{y_{t,j}}) p_\beta(\pi_t|Z_t) d\pi_{t,1} ... d\pi_{t,q} = \prod_{j=1}^{q} \overline{\pi}_{t,j}^{y_{t,j}} (1 - \sum_{j=1}^{q} \overline{\pi}_{t,j})^{1-\sum_{j=1}^{q} y_{t,j}}, \end{aligned} \tag{4.8}$$

where $E(y_{t,i}|Z_t) = E(\pi_{t,i}|Z_t) = \overline{\pi}_{t,i}$ and $Var(y_{t,i}|Z_t) = Var(\pi_{t,i}|Z_t) = \overline{\pi}_{t,i}(1 - \overline{\pi}_{t,i})$ and $Cov(y_{t,i}, y_{t,j}|Z_t) = -\overline{\pi}_{t,i} \times \overline{\pi}_{t,j}$, for $i \neq j$.

Let $\mu_{t,j} = logit\pi_{t,j} = \log(\frac{\pi_{t,j}}{1-\sum_{j=1}^{q} \pi_{t,j}})$, $j = 1, ..., q$; while $\mu_{t,j} \overset{indp}{\sim} N(\eta_{t,j} = Z'_{t,j}\beta, \sigma_j^2)$, $j = 1, ..., q$. Then the probability density function for $\{\pi_{t,j}\}$ will be:

$$p_\beta(\pi_{t,j}|Z_t) = \frac{1}{\sqrt{\sigma_j^2 2\pi}} (\frac{1-\sum_{i=1}^{q} \pi_{t,i} + \pi_{t,j}}{\pi_{t,j}(1-\sum_{i=1}^{q} \pi_{t,i})}) exp[-2\sigma_j^2 (logit\pi_{t,j} - \eta_{t,j})^2], 0 < \pi_{t,j} < 1.$$

The expectation of $\pi_{t,j}$ given Z_t is computed as:

$$\begin{aligned} E(\pi_{t,j}|Z_t) &= \frac{1}{\sqrt{\sigma_j^2 2\pi}} \int (\frac{1-\sum_{i=1}^{q} \pi_{t,i} + \pi_{t,j}}{(1-\sum_{i=1}^{q} \pi_{t,i})}) exp[-2\sigma_j^2 (logit\pi_{t,j} - \eta_{t,j})^2] d\pi_{t,j} \\ &= E(\frac{exp(\mu_{t,j})}{1+\sum_{i=1}^{q} exp(\mu_{t,i})}). \end{aligned}$$

The likelihood functions is given by

$$L_n(\beta) = \prod_{t=1}^{n} p(y_t|Z_t). \tag{4.9}$$

4.3.5 Bivariate Binary Models

Let $\{Y_t\}$ denote a bivariate binary response variable where $Y_t = \begin{pmatrix} Y_t(1) \\ Y_t(2) \end{pmatrix}$, $Y_t(j) = 0, 1$ j=1,2. Also denote:

$$p_{00}(t) = P(Y_t(1) = 0, Y_t(2) = 0), p_{01}(t) = P(Y_t(1) = 0, Y_t(2) = 1),$$
$$p_{10}(t) = P(Y_t(1) = 1, Y_t(2) = 0), p_{11}(t) = P(Y_t(1) = 1, Y_t(2) = 1).$$

Note that $p_{00}(t) + p_{01}(t) + p_{10}(t) + p_{11}(t) = 1$.

The probability distribution of $\{Y_t\}$ given $\{Z_t\}$, where $\{Z_t\}$ is a covariate vector defined by $\{Z_t\} = (y^*_{t-1}(1), y^*_{t-1}(2), x^*_{t-1})$ and $\{x_t\}$ is possibly a random sequence, is given below:

$$\begin{aligned} p(y_t|z_t) &= P(Y_t(1) = y_t(1), Y_t(2) = y_t(2)|Z(t)) \\ &= (p_{00}(t)^{(1-y_t(1))(1-y_t(2))} p_{01}(t)^{(1-y_t(1))y_t(2)} p_{10}(t)^{y_t(1)(1-y_t(2))} p_{11}(t)^{(y_t(1)y_t(2)} \\ &= p_{00}(t)(\frac{p_{10}(t)}{p_{00}(t)})^{y_t(1)}(\frac{p_{01}(t)}{p_{00}(t)})^{y_t(2)}(\frac{p_{11}(t)p_{00}(t)}{p_{01}(t)p_{10}(t)})^{y_t(1)y_t(2)}, \end{aligned}$$

where $p_{ij}(t)$, $i, j = 0, 1$ are now the conditional probabilities given Z_t.

We then are able to compute the expectations as follows:

$E(Y_t(1)|Z_t) = P(Y_t(1) = 1|Z_t) = p_{10}(t) + p_{11}(t) = \mu_t(1)$;

$E(Y_t(2)|Z_t) = P(Y_t(2) = 1|Z_t) = p_{01}(t) + p_{11}(t) = \mu_t(2)$;

$E(Y_t(1)Y_t(2)|Z_t) = P(Y_t(1) = 1, Y_t(2) = 1|Z_t) = p_{11}(t) = \mu_t(1, 2)$;

Suppose we have a zero mean bivariate autoregressive process$\{X_t\}$defined by:

$$X_t = \begin{pmatrix} X_t(1) \\ X_t(2) \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{t-1}(1) \\ X_{t-1}(2) \end{pmatrix} + \begin{pmatrix} \epsilon_t(1) \\ \epsilon_t(2) \end{pmatrix}.$$

We create a bivariate binary process by clipping:

$$Y_t(j) = \begin{cases} 1 & if \quad X_t(j) \geq 0 \\ 0 & \text{otherwise.} \end{cases} \quad where \quad j = 1, 2,$$

Suppose $\varepsilon = \begin{pmatrix} \epsilon_t(1) \\ \epsilon_t(2) \end{pmatrix}$ t=1,2,... are independent and identically distributed having bivariate normal density, $N_2(0, \Sigma)$. We have:

$$\begin{aligned} \mu_t(j) = E(Y_t(j)|Z_{t-1}) &= P(Y_t(j) = 1|Z_{t-1}) \\ &= P(X_t(j) \geq 0|Z_{t-1}) \\ &= P(\epsilon_t(j) \geq -(\phi_{j1}X_{t-1}(1) + \phi_{j2}X_{t-1}(2))) \\ &= F_{\epsilon_j}(\phi_{j1}X_{t-1}(1) + \phi_{j2}X_{t-1}(2)) \quad j = 1, 2, \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} \mu_t(1,2) &= P(X_t(1) \geq 0, X_t(2) \geq 0|Z_{t-1}) \\ &= P(\epsilon_t(1) > -(\phi_{11}X_{t-1}(1) + \phi_{12}X_{t-1}(2)), \epsilon_t(2) > -(\phi_{21}X_{t-1}(1) + \phi_{22}X_{t-1}(2))) \\ &= F_{\epsilon_1,\epsilon_2}((\phi_{11}X_{t-1}(1) + \phi_{12}X_{t-1}(2)), (\phi_{21}X_{t-1}(1) + \phi_{22}X_{t-1}(2))), \end{aligned} \tag{4.11}$$

where $F_{\epsilon_j}, F_{\epsilon_1,\epsilon_2}$ stand for the cdf of zero-mean normal distribution and bivariate normal distribution respectively.

Two methods of estimation are proposed for this model: (a) Maximum likelihood based on $\{X_t\}$; (b) Partial likelihood based on both $\{Y_t, X_t\}$.

(1) **Maximum likelihood:**

The likelihood function based on $\{x_t\}$ is given by

$$\begin{aligned} L(\Phi, \sigma_{\epsilon_j}; x_t) &= \Pi_{t=1}^n p(x_t|x_{t-1}) = \Pi_{t=1}^n f(\epsilon_{t,1}, \epsilon_{t,2}) \\ &= \Pi_{t=1}^n \frac{1}{2\pi\sigma_{\epsilon_1}\sigma_{\epsilon_2}\sqrt{1-\rho^2}} exp\{-\frac{[(\frac{\epsilon_{t,1}}{\sigma_{\epsilon_1}})^2 + (\frac{\epsilon_{t,2}}{\sigma_{\epsilon_2}})^2 - 2\rho\frac{(\epsilon_{t,1})(\epsilon_{t,2})}{\sigma_{\epsilon_1}\sigma_{\epsilon_2}}]}{2(1-\rho^2)}\}. \end{aligned}$$

Where $\epsilon_{t,1} = x_{t,1} - \phi_{11}x_{t-1,1} - \phi_{12}x_{t-1,2}$, and $\epsilon_{t,2} = x_{t,2} - \phi_{21}x_{t-1,1} - \phi_{22}x_{t-1,2}$.

Alternatively, we can estimate the coefficients via the Yule-Walker equation.

If $X_t = \begin{pmatrix} X_t(1) \\ X_t(2) \end{pmatrix}$, $\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix}$ and $\varepsilon_t = \begin{pmatrix} \epsilon_{t1} \\ \epsilon_{t2} \end{pmatrix}$

the model can be written as :

$$X_t = \Phi X_{t-1} + \varepsilon_t, \quad \{\varepsilon_t\} \sim N_2(0, \Sigma).$$

then by Yule-Walker equation: $\hat{\Phi} = (\hat{\Gamma}^{-1}(0))\hat{\Gamma}(1)$, where

$\hat{\Gamma}(h) = n^{-1}\sum_{t=1}^{n-h}(X_{t+h} - \overline{X_n})(X_{t+h} - \overline{X_n})'$ and

$\hat{\Sigma} = n^{-1}\sum_{t=1}^{n}[(X_t - \overline{X_n}) - \hat{\Phi}(X_t - \overline{X_n})][(X_t - \overline{X_n}) - \hat{\Phi}(X_t - \overline{X_n})]'$.

(2)**Partial Likelihood**

The partial likelihood function is given by:

$$\begin{aligned} PL(\Phi, \sigma_{\epsilon_j}; y_t, x_t) &= \Pi_{t=1}^{n} p(y_t|z_t) \\ &= \Pi_{t=1}^{n} p_{00}(t)(\frac{p_{10}(t)}{p_{00}(t)})^{y_t(1)}(\frac{p_{01}(t)}{p_{00}(t)})^{y_t(2)}(\frac{p_{11}(t)p_{00}(t)}{p_{01}(t)p_{10}(t)})^{y_t(1)y_t(2)}. \end{aligned} \tag{4.12}$$

4.4 IBM data analysis

We collected IBM stock daily price data for 2005 as well as the corresponding Dow Jones (DJ) index. Let X_{t1} and X_{t2} denote the mean-centered DJ and IBM stock daily price respectively.

1.Partial Likelihood

Binary Clipping for the IBM data

The binary response variable $\{Y_{t,2}\}$ is created by clipping :

$$Y_{t,2} = \begin{cases} 1 & if \quad X_{t,2} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

Consider the model:

$$X_{t,2} = \phi_1 X_{t-1,1} + \phi_2 X_{t-1,2} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a sequence of independent $N(0, \sigma^2)$ random variables. As an alternative model, we also consider the logistic distribution with density $f_{\epsilon_t}(z) = \frac{exp(-z)}{(1+exp(-z))^2}$. We have

$$\begin{aligned} p_t(\phi) = P_\phi(Y_{t,2} = 1|\mathscr{F}_{t-1}) &= P(\epsilon_t > -(\phi_1 x_{t-1,1} + \phi_2 x_{t-1,2})) \\ &= \begin{cases} \Phi(\phi_1 x_{t-1,1} + \phi_2 x_{t-1,2}): & \text{Normal} \\ \frac{1}{1+exp(-(\phi_1 x_{t-1,1} + \phi_2 x_{t-1,2}))}: & \text{Logistic.} \end{cases} \end{aligned} \tag{4.14}$$

Since $\{Y_t\}$ is binary, the conditional density of y_t is given by

$$p_t(y_t;\phi) = [p_t(\phi)]^{y_t}[1-p_t(\phi)]^{1-y_t}.$$

The corresponding partial likelihood is simply the product

$$PL(\phi) = \prod_{t=1}^{N} p_t(y_t;\phi) = \prod_{t=1}^{N}[p_t(\phi)]^{y_t}[1-p_t(\phi)]^{1-y_t}. \tag{4.15}$$

We can estimate ϕ's by maximizing the above partial likelihood function. For the first t=250 observations, we apply this method to six models (see Table 4.1) to estimate parameter $\phi's$ and then predict the future chain from t=253 to 288.

Table 4.1: Models with normal/logistic error distributions

model	Linear part
Model 1	$X_{t,2} = \phi X_{t-1,2} + \epsilon_t$where $\epsilon_t \sim N(0,\sigma^2)$
Model 2	$X_{t,2} = \phi X_{t-1,2} + \epsilon_t$ where $\epsilon_t \sim logistic(0,1)$
Model 3	$X_{t,2} = \phi_1 X_{t-1,1} + \phi_2 X_{t-1,2} + \epsilon_t$ where $\epsilon_t \sim N(0,\sigma^2)$
Model 4	$X_{t,2} = \phi_1 X_{t-1,1} + \phi_2 X_{t-1,2} + \epsilon_t$ where $\epsilon_t \sim logistic(0,1)$
Model 5	$X_{t,2} = \phi_1 X_{t-1,1} + \phi_2 X_{t-1,2} + \phi_3 X_{t-2,1} + \phi_4 X_{t-2,2} + \epsilon_t$ where $\epsilon_t \sim N(0,\sigma^2)$
Model 6	$X_{t,2} = \phi_1 X_{t-1,1} + \phi_2 X_{t-1,2} + \phi_3 X_{t-2,1} + \phi_4 X_{t-2,2} + \epsilon_t$ where $\epsilon_t \sim logistic(0,1)$

The likelihood ratio tests show that for both normal/logistic error distribution models, the first-order models with only IBM variable involved, that is, Model1 and Model 2 in Table 4.1 are the best. The prediction results are listed in Table 4.2.

Table 4.2: The predicted probabilities based on normal/logistic error distributions, $\hat{p}_{(t,norm)}, \hat{p}_{(t,log)}$,and observed Y_t

Y_t	$\hat{p}_{(t,log)}$	$\hat{p}_{(t,norm)}$	Y_t	$\hat{p}_{(t,log)}$	$\hat{p}_{(t,norm)}$	Y_t	$\hat{p}_{(t,log)}$	$\hat{p}_{(t,norm)}$
0	0.4356	0.3721	0	0.1968	0.0384	0	0.1471	0.0135
0	0.1175	0.0056	0	0.0980	0.0026	1	0.2277	0.0621
1	0.9601	1.0000	1	0.7308	0.8957	1	0.8333	0.9786
1	0.8552	0.9873	0	0.6688	0.8118	0	0.4953	0.4906
1	0.4175	0.3375	0	0.7521	0.9188	0	0.4584	0.4169
0	0.0364	0.0000	0	0.0398	0.0000	0	0.0148	0.0000
0	0.0165	0.0000	0	0.0119	0.0000	0	0.0201	0.0000
0	0.0577	0.0002	0	0.0327	0.0000	0	0.0963	0.0024
0	0.0289	0.0000	0	0.0031	0.0000	0	0.0014	0.0000
0	0.0025	0.0000	0	0.0194	0.0000	0	0.0095	0.0000
0	0.0493	0.0001	0	0.0103	0.0000	0	0.0322	0.0000
0	0.0213	0.0000	0	0.0237	0.0000	0	0.0165	0.0000
0	0.0114	0.0000	0	0.0510	0.0001	0	0.0067	0.0000

2. Observation-driven models

For the series $\{Y_t\}$ created by bivariate clipping, we also apply our observation-driven state-space models.

The observation model:

$$p(y_t|\pi_t^*, y_{t-1}^*, x_t) = p(y_t|\pi_t) = \pi_t^{y_t}(1-\pi_t)^{1-y_t}, \quad y_t = 1, 0$$

The state model:

$$p(\pi_t|\pi_t^*, y_{t-1}^*, x_t^*) = p_\beta(\pi_t|Z_t), \quad 0 < \pi_t < 1$$

(A) Beta prior:

We define $p_\beta(\pi_t|Z_t) = \frac{\pi_t^{\mu_t - 1}(1-\pi_t)^{(1-\mu_t)-1}}{B(\mu_t, 1-\mu_t)}$, $0 < \pi_t < 1$. And $logit(\mu_t)(= log(\frac{\mu_t}{1-\mu_t})) = Z_t'\beta$. The likelihood function is thus given by

$$L(\beta) = \Pi_{t=1}^n \mu_t^{y_t}(1-\mu_t)^{1-y_t}.$$

(B) Logit-normal prior:

Let $\mu_t = logit\pi_t = log\frac{\pi_t}{1-\pi_t}$, while $\mu_t \sim N(\eta_t = Z_t'\beta = \beta_0 + \beta_1 Y_{t-1,1} + \beta_2 Y_{t-1,2}, \sigma^2)$. The Gaussian model here is a non-conjugate prior.

The likelihood function is:

$$\prod_{t=1}^n p(y_t|Z_t) = \prod_{t=1}^n \overline{\pi}_t^{y_t}(1-\overline{\pi}_t)^{1-y_t} \tag{4.16}$$

where $\overline{\pi}_t = E(\pi_t|Z_t) = E((1+exp(-\mu_t))^{-1})$, which is approximated by Gaussian Hermite Quadrature with 8 pair of nodes and the likelihood function is then maximized by Quasi-Newton optmization method.

We also do model selections for each prior as in Table 4.3below:

Table 4.3: Candidate state space models for IBM data

Model 1	$1 + Y_{t-1}$
Model 2	$1 + Y_{t-1} + Y_{t-2}$
Model 3	$1 + Y_{t-1} + Y_{t-2} + Y_{t-3}$
Model 4	$1 + Y_{t-1} + Y_{t-2} + Y_{t-3} + Y_{t-4}$
Model 5	$1 + Y_{t-1} + Y_{t-2} + Y_{t-3} + Y_{t-4} + Y_{t-5}$

The model selections results are listed in Table 4.4, where the second column lists the number of estimated parameters, the third column reports the deviance of the model: $D = -2lnL(\hat{\beta})$ and the next two columns correspond to AIC and BIC criteria, where $AIC = D + 2d$;

$BIC = D + dlog(N)$, where d is the number of parameters being estimated and N is the total number of observations. The last two columns give the values of the likelihood ratio test statistics together with their p-values for testing the order of the model. We see that all criterions chose Model 3 as the best for Beta-prior, while for logit-normal prior, Model 4 is chosen by AIC criterion and LRT, BIC criterion selected Model 3 is the best. After the estimators β are obtained by maximizing the likelihood function, they are plugged into $\overline{\pi}_t$ to estimate the prediction probabilities, see Table4.5.

Table 4.4: The state space model selection for IBM stock data

Method	Model	parameter	D	AIC	BIC	λ_n	p-value
Beta prior	model 1	2	158.499	164.499	175.438		
	model 2	4	150.244	160.244	178.541	8.255	0.0161
	model 3	6	130.594	**144.594**	**170.186**	19.650	< 0.0001
	model 4	8	127.296	149.296	178.168	3.298	0.192
Logit-normal prior	model 1	3	132.346	140.246	149.335		
	model 2	5	124.900	138.900	153.197	7.446	0.0242
	model 3	7	106.068	120.068	**141.659**	18.832	< 0.0001
	model 4	9	100.000	**118.000**	150.872	6.068	**0.0481**
	Model 5	11	99.500	123.500	162.222	0.500	0.779

Table 4.5: The observed Y_t,predicted probabilities by Beta prior $\hat{p}_t^{(1)}$, and by Logit-normal prior $\hat{p}_t^{(2)}$

Y_t	0	0	1	1	1	1	1
$\hat{p}_t^{(1)}$	0.04147	0.04147	0.04147	0.65894	0.59263	0.94674	0.94674
$\hat{p}_t^{(2)}$	0.04414	0.04414	0.04414	0.63152	0.55559	0.83507	0.96943
Y_t	0	0	1	0	0	0	0
$\hat{p}_t^{(1)}$	0.94674	0.28840	0.34992	0.65894	0.03155	0.34992	0.04147
$\hat{p}_t^{(2)}$	0.96943	0.44850	0.52684	0.90604	0.03114	0.17230	0.21853
Y_t	0	0	0	0	0	0	0
$\hat{p}_t^{(1)}$	0.04147	0.04147	0.04147	0.04147	0.04147	0.04147	0.04147
$\hat{p}_t^{(2)}$	0.04414	0.04414	0.04414	0.04414	0.04414	0.04414	0.04414
Y_t	0	0	0	0	0	0	0
$\hat{p}_t^{(1)}$	0.04147	0.04147	0.04147	0.04147	0.04147	0.04147	0.04147
$\hat{p}_t^{(2)}$	0.04414	0.04414	0.04414	0.04414	0.04414	0.04414	0.04414
Y_t	0	0	0	0	0	0	0
$\hat{p}_t^{(1)}$	0.04147	0.04147	0.04147	0.04147	0.04147	0.04147	0.04147
$\hat{p}_t^{(2)}$	0.04414	0.04414	0.04414	0.04414	0.04414	0.04414	0.04414

From Table 4.5, we see that the two priors produce very similar results on the IBM data, when compared with partial likelihood method which additionally employs $\{x_t\}$ information(see Table 4.2), the observation-driven models which only using the $\{Y_t\}$ information perform very well too.

4.5 Multi-Category: DNA Data Analysis

A DNA sequence consists of four nucleotides differing only in the nitrogenous base, whose order determines the genetic information of each organism. The four nucleotides are given one-letter abbreviations as shorthand as follows:

A for adenine; G for guanine; C for cytosine; T for thymine.

Adenine and guanine are purines–the larger of the two types of bases found in DNA–while cytosine and thymine are pyrimidines.

Thus, a strand of DNA can be represented as a sequence of letters from {A,C,G,T} and can be viewed as a nominal categorical time series with the assignment A=1, C=2,G=3 and T=4. For more information, see Waterman (1995).

We present the data analysis for DNA sequence data of the gene BNRF1 of the Epstein-Barr virus (see Shumway and Stoffer (2000), Section 5.9) considering only the first 1000 observations.

We apply the model proposed in Section 3.4 by fitting a series of models.

The observation model is:

$$p(y_t|\pi_t^*, y_{t-1}^*, x_t^*) = p(y_t|\pi_t) = (\prod_{j=1}^{q} \pi_{t,j}^{y_{t,j}})(1 - \sum_{j=1}^{q} \pi_{t,j})^{1-\sum_{j=1}^{q} y_{t,j}}; \text{ with } q = m - 1 = 3$$

and the state model is:

$$p(\pi_t|\pi_{t-1}^*, y_{t-1}^*, x_t^*) = p_\beta(\pi_t|Z_t),$$

where $logit\pi_{t,j} = \log(\frac{\pi_{t,j}}{1-\sum_{j=1}^{q}\pi_{t,j}}) = \mu_{t,j}$, and $\mu_{t,j} \overset{indp}{\sim} N(\eta_{t,j} = Z'_{t,j}\beta_j, \sigma_j^2)$, j=1,2,3. Here$Z_{t,j}$ refers to the past {Y}s.

Accordingly, the likelihood function:

$$L = \prod_{t=1}^{n}\prod_{j=1}^{q} \overline{\pi}_{t,j}^{y_{t,j}}(1 - \sum_{j=1}^{q} \overline{\pi}_{t,j})^{1-\sum_{j=1}^{q} y_{t,j}} \tag{4.17}$$

where $\overline{\pi}_{t,j} = E(\pi_{t,j}|Z_t) = E(\frac{exp(\mu_{t,j})}{1+\sum_{i=1}^{q} exp(\mu_{t,i})})$, which could be approximated by the High-dimension Gaussian Quadrature. Quasi-Newton optimization methods is then applied to

maximize the likelihood function (4.17), and estimate the parameter β_j—-a d-dimensional vector.

For example, a first-order model is given by

$$\mu_{t,j} = logit\pi_{t,j} = log\frac{\pi_{t,j}}{1-\sum_{i=1}^{q}\pi_{t,i}}, \text{ where}$$
$$\mu_{t,j} \overset{indp}{\sim} N(\eta_{t,j} = \beta_{j,0} + \beta_{j,1}Y_{t-1,1} + \beta_{j,2}Y_{t-1,2} + \beta_{j,3}Y_{t-1,3}, \sigma_j^2), \quad j = 1, 2, 3$$

and is denoted by $1 + Y_{t-1}$. A second-order model is labeled $1 + Y_{t-1} + Y_{t-2}$ and $\eta_{t,j}$ consists of the above plus a linear combination in terms of $Y_{t-2,1}, Y_{t-2,2}, Y_{t-2,3}$. In the same manner we carry out up to the fourth-order model fitting.

Fokianos and Kedem (2003) applied regression-type methods for the same data. They employed the multinomial logit model defined by Agresti (1990, section 9.2) below:

Let the tth observation of m categorical time series be expressed by the vector $Y_t = (Y_{t,1}, ..., Y_{t,q})'$ of length $q = m - 1$, with elements

$$Y_{t,j} = \begin{cases} 1 & \text{if the jth category is observed at time } t, \\ 0 & \text{otherwise.} \end{cases} \tag{4.18}$$

for t=1,...,N and $j = 1, ..., q$. Denote by $\pi_t = (\pi_{t,1}, ..., \pi_{t,q})'$ the vector of conditional probabilities given $\mathscr{F}_{t-1}$, where

$$\pi_{t,j} = E(Y_{t,j}|\mathscr{F}_{t-1}) = p(Y_{t,j} = 1|\mathscr{F}_{t-1}),$$

$j = 1, ..., q$. The σ-field $\mathscr{F}_{t-1}$ is generated by $Z_{t-1} = (y_{t-1}^*, x_t^*)$ and x_t^* stands for the covariates. Through the logit link-type function, we have

$$\pi_{t,j} = \frac{exp(Z_{t-1}'\beta_j)}{1+\sum_{i=1}^{q} exp(Z_{t-1}'\beta_i)}, \quad j = 1, ..., q$$

Here β_j,$j = 1, ..., q$, are d-dimensional regression parameters and Z_{t-1} is a corresponding d-dimensional vector of stochastic time-dependent covariates independent of j.

For example, in this DNA data analysis, the first-order model is given by:

$$\log(\frac{\pi_{t,j}(\beta)}{\pi_{t,4}(\beta)}) = \beta_{j,0} + \beta_{j,1}Y_{t-1,1} + \beta_{j,2}Y_{t-1,2} + \beta_{j,3}Y_{t-1,3} \quad j = 1, 2, 3. \tag{4.19}$$

For the purpose of comparison, Table 4.6 lists the models applied to the DNA sequence data and Table 4.7 reports the inferential results from both state-space and regression methods, where the second column lists the number of estimated parameters, the third column reports the deviance of the model D and the next two columns correspond to AIC and BIC cirteria. The last two columns give the values of the likelihood ratio test statistics together with their p-values for testing the order of the model.

Table 4.6: Candidate models for the gene BNRF1 of the Epstein-Barr virus DNA sequence data

Model 1	$1 + Y_{t-1}$
Model 2	$1 + Y_{t-1} + Y_{t-2}$
Model 3	$1 + Y_{t-1} + Y_{t-2} + Y_{t-3}$
Model 4	$1 + Y_{t-1} + Y_{t-2} + Y_{t-3} + Y_{t-4}$

Table 4.7: Comparison of various-order models for the gen BNRF1 of the Epstein-Barr virus DNA sequence data(N=1000)

Model	Order	p	D	AIC	BIC	λ_n	p-value
State-space	Independent	6	2713.025	2725.025	2754.47		
	Model 1	15	2649.49	2679.49	2753.108	63.53	< 0.000001
	Mode 2	24	2628.96	2676.96	2794.75	20.53	0.0149
	Model 3	33	2606.93	2672.93	2834.88	22.03	0.0088
	Model 4	42	2599.049	2683.049	2889.17	7.88	0.545
Regression	Independent	3	2711.31	2717.31	2732.02		
	Model 1	12	2677.75	2701.75	2760.60	33.56	0.0001
	Mode 2	21	2664.27	2706.27	2809.25	13.48	0.1420
	Model 3	30	2648.41	2708.41	2855.52	15.86	0.0698
	Model 4	39	2639.39	2714.39	2905.63	12.02	0.2121

Based on the AIC and likelihood ratio test, the state-space method chooses 3-order model as the best model, while regression type method also selects 3-order model while only $Y_{t-1}$and Y_{t-3} as the covariates of the linear part.

We also predict the transition probabilities by each model. For the regression model,the transition probability $P(Y_t = i|Y_{t-1} = j, Y_{t-3} = l)$ for i, j, l=1,2,3,4 are estimated by substitution of the maximum partial likelihood estimators into the regression equation of $\pi_{t,i}$ using (4.19). Table 4.8 reports the transition probabilities among the different states where, for example, if $Y_{t-3} = A$ and $Y_{t-1} = T$, then the transition probability to $Y_t = C$ is equal to 0.2592.

Table 4.8: Estimated transition matrix from model $1 + Y_{t-1} + Y_{t-3}$ for the gen BNRF1 of the Epstein-Barr virus DNA sequence data(N=1000)

		Y_t			
Y_{t-3}	Y_{t-1}	A	C	G	T
A	A	0.2004	0.2915	**0.3389**	0.1692
	C	0.2756	**0.3097**	0.2089	0.2058
	G	0.2352	**0.3257**	0.3219	0.1172
	T	0.1583	0.2592	**0.3526**	0.2299
C	A	0.1479	0.2889	**0.3828**	0.1804
	C	0.2107	**0.3179**	0.2443	0.2271
	G	0.1763	0.3279	**0.3692**	0.1266
	T	0.1149	0.2526	**0.3916**	0.2409
G	A	0.2167	0.3069	**0.3342**	0.1422
	C	0.2972	**0.3251**	0.2053	0.1724
	G	0.2511	**0.3384**	0.3135	0.0970
	T	0.1738	0.2770	**0.3531**	0.1961
T	A	0.1643	0.2335	**0.3652**	0.2370
	C	0.2289	0.2513	0.2279	**0.2919**
	G	0.2001	0.2705	**0.3596**	0.1688
	T	0.1249	0.1997	**0.3656**	0.3098

Similarly, for state-space model, plug in the linear coefficients obtained by maximizing the likelihood function (4.17), to approximate $\overline{\pi}_{t,j}$, which is the mean of transition probability $\pi_{t,j}$ given Z_t (see Table 4.9).

Table 4.9: Estimated transition matrix from model $1 + Y_{t-1} + Y_{t-2} + Y_{t-3}$ for the gen BNRF1 of the Epstein-Barr virus DNA sequence data(N=1000)

			Y_t			
Y_{t-1}	Y_{t-2}	Y_{t-3}	A	C	G	T
A	A	A	0.1778251	0.2290601	**0.298148**	0.2949668
		C	0.125896	**0.3350852**	0.2454728	0.293546
		G	0.1857601	0.2433225	**0.3222724**	0.248645
		T	0.1247395	0.3089377	0.2520591	**0.3142637**
	C	A	0.1276695	0.3103511	0.2471709	**0.3148085**
		C	0.0838423	**0.4207186**	0.1884157	0.3070234
		G	0.1339133	**0.3312549**	0.2685522	0.2662796
		T	0.0837368	**0.3910701**	0.1951154	0.3300777
	G	A	0.2217018	**0.3479214**	0.1825425	0.2478343

						Y_t
Y_{t-1}	Y_{t-2}	Y_{t-3}	A	C	G	T
		C	0.1460276	**0.4725163**	0.1392451	0.242211
		G	0.232239	**0.3698909**	0.1972054	0.2006647
		T	0.1475015	**0.4446206**	0.1460377	0.2618402
	T	A	0.2198668	0.2655186	0.2296538	**0.2849608**
		C	0.1524703	**0.3801538**	0.1848987	0.2824772
		G	0.2300745	**0.2824692**	0.2485773	0.238879
		T	0.1522858	**0.3533922**	0.1914619	0.3028601
C	A	A	0.2696437	0.1291132	0.1631762	**0.4380669**
		C	0.2051667	0.2031658	0.1446181	**0.4470494**
		G	0.302782	0.1470468	0.188293	**0.3618782**
		T	0.1945817	0.179912	0.1432068	**0.4822995**
	C	A	0.1980809	0.1797594	0.1396695	**0.4824902**
		C	0.1406626	0.2634615	0.1150578	**0.4808181**
		G	0.2280212	0.2092807	0.1642482	**0.3984499**
		T	0.1335763	0.2339403	0.1144074	**0.518076**
	G	A	0.3428175	0.1991569	0.1008562	**0.3571694**
		C	0.2518036	0.3019434	0.0859406	**0.3603124**
		G	**0.3694143**	0.2184261	0.1125526	0.299607
		T	0.2468405	0.2757763	0.0875697	**0.3898135**
	T	A	0.321375	0.1439127	0.1205135	**0.4141988**
		C	0.2443886	0.2261899	0.1066105	**0.422811**
		G	**0.3566548**	0.162203	0.1378338	0.3433084
		T	0.2340492	0.2021491	0.1064831	**0.4573186**
G	A	A	0.1747397	0.225158	0.299624	**0.3004783**
		C	0.1239013	**0.3298966**	0.2470877	0.2991144
		G	0.1826734	0.2394073	**0.3242196**	0.2536997
		T	0.1225878	0.3037134	0.253355	**0.3203438**
	C	A	0.1254795	0.3051336	0.2484676	**0.3209193**
		C	0.0825822	**0.4145577**	0.1898275	0.3130326
		G	0.1318396	**0.3262773**	0.2704566	0.2714265
		T	0.0823287	**0.3846537**	0.1962389	0.3367787
	G	A	0.2188321	**0.3436233**	0.1843542	0.2531904

						Y_t
Y_{t-1}	Y_{t-2}	Y_{t-3}	A	C	G	T
		C	0.1443034	**0.4672584**	0.1408131	0.2476251
		G	0.2291571	**0.365286**	0.1991811	0.2063758
		T	0.1457092	**0.4394632**	0.1475975	0.2672301
	T	A	0.2165566	0.2616201	0.2313515	**0.2904718**
		C	0.1503757	**0.3750927**	0.1865335	0.2879981
		G	0.2267111	**0.2785109**	0.2506232	0.2441548
		T	0.1500188	**0.3482641**	0.1929188	0.3087983
T	A	A	0.1495064	0.1160819	**0.3861615**	0.3482502
		C	0.114884	0.1845877	0.3461148	**0.3544135**
		G	0.1582605	0.1250562	**0.4233102**	0.2933731
		T	0.1102353	0.1648526	0.3444842	**0.3804279**
	C	A	0.1131971	0.1661743	0.3390087	**0.3816199**
		C	0.0821611	0.2494251	0.2866353	**0.3817785**
		G	0.1219395	0.182062	**0.3775234**	0.3184751
		T	0.0788968	0.2231167	0.2860206	**0.4119659**
	G	A	0.2161584	0.2051858	0.2758648	**0.302791**
		C	0.1566986	**0.3073668**	0.2325749	0.3033597
		G	0.2264731	0.2186897	**0.299252**	0.2555852
		T	0.1546952	0.2823492	0.2379602	**0.3249954**
	T	A	0.1970832	0.1436332	0.3178453	**0.3414383**
		C	0.1490883	0.2247163	0.2801112	**0.3460842**
		G	0.208696	0.1548014	**0.348621**	0.2878816
		T	0.144328	0.2024803	0.2812717	**0.37192**

From Table 4.7, we can see that, compared with the regression model, except the independent model, our observation-driven state-space model reduces AIC and BIC values in every order model although we have more parameters being estimated .

Also, from the probability of prediction Table 4.8, 4.9, if we predict the possible outcomes based on the highest probability indicting by the highlight numbers, we see that the state-space model is able to predict 'A' which has lowest frequencies of appearance in the DNA sequence; Also the state-space model predicts the probabilities much more accurately (86 vs.52 out of 200) than the regression type models.

Chapter 5

Categorical Time Series Models for Contingency tables

5.1 Introduction

Fokianos and Kedem (2003) have discussed regression models for categorical time series. See also Fahrmeir and Kaufmann (1987), Fokianos and Kedem (1998), Kaufmann (1987), and Kedem (1980). Consider a bivariate binary time series $\{Y_t\}$, t=1,2,..., where $Y(t) = (Y_1(t), Y_2(t))'$, and

$$Y_i(t) = \begin{cases} 1 & \text{if event i occur at time t} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, 2.$$

Suppose we observe $Y(t)$ for $n(t)$ independent individuals at time t. One can then construct a sequence of (2x2) contingency tables: for t=1,2,...,

		$Y_2(t)$	
		0	1
$Y_1(t)$	0	$n_{00}(t)$	$n_{01}(t)$
	1	$n_{10}(t)$	$n_{11}(t)$

where $n_{uv}(t)$=number of times $Y_1(t) = u$ and $Y_2(t) = v$ among the $n(t)$ individuals; u, v=0,1. Note that $\sum\sum n_{uv}(t) = n(t)$. For each fixed t , the count vector $N(t) = (n_{00}(t), n_{01}(t), n_{10}(t), n_{11}(t))'$ is assumed to be a multinomial vector with cell probabilities $\{p_{uv}(t)\}$,u, v=0,1, and index(assumed non-random) $n(t)$. If $\{N(t)\}$, t=1,2,..., are independent over time, one can apply standard generalized linear regression modeling techniques for the (marginal) means $\mu_i(t) = E(Y_i(t))$, i=1,2, and the pairwise log-odds ratio (LOR)

$$\lambda_{12}(t) = log(\frac{p_{11}(t)p_{00}(t)}{p_{01}(t)p_{10}(t)}),$$

which is a measure of association between $Y_1(t)$ and $Y_2(t)$. In this chapter, we consider the case when $\{N(t)\}$, t=1,2,..., is a sequence of dependent multinomials.

We present four models for $\{Y(t)\}$ (and hence $\{N(t)\}$) in Section 5.2. Section 5.3 is concerned with parameter estimation for the models. The models are applied to two real data sets and the results are reported in Section 5.4. Some concluding remarks are presented in Section 5.5.

5.2 Model Specification

We present four models for contingency data analysis. Models 1 to 3 are based on regression type models and Model 4 is an observation-driven state space model.

Model 1 Conditional Exponential Family

Suppose that conditional on $\mathscr{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, ...)$ the joint density function of $Y(t) = (Y_1(t), Y_2(t))'$ is given by

$$p(Y_t|\mathscr{F}_{t-1}) \propto exp[\theta_0(t) + \theta_1(t)Y_1(t) + \theta_2(t)Y_2(t) + \theta_3(t)Y_1(t)Y_2(t)], \tag{5.1}$$

where $\theta_0(t) = ln(p_{00}(t))$, $\theta_1(t) = log(\frac{p_{10}(t)}{p_{00}(t)})$, $\theta_2(t) = log(\frac{p_{01}(t)}{p_{10}(t)})$ and $\theta_3(t) = log[\frac{p_{11}(t)p_{00}(t)}{p_{01}(t)p_{10}(t)}]$. Here $p_{uv}(t)$ denotes the conditional probabilities given $\mathscr{F}_{t-1}$. Note that (5.1) is just a reparameterization of

$$p(Y_t|\mathscr{F}_{t-1}) = \prod_{u,v=0,1} (p_{uv}(t))^{\xi_{uv}(t)}, \tag{5.2}$$

where

$$\xi_{uv}(t) = \begin{cases} 1 & \text{if } Y_1(t) = u \text{ and } Y_2(t) = v \\ 0 & \text{otherwise.} \end{cases}$$

Since $p_{00}(t) = 1 - p_{01}(t) - p_{10}(t) - p_{11}(t)$, it suffices to model $\theta_i(t)$, i=1,2,3, as functions of past observations. Using the model (5.1) for $n(t)$i.i.d observations $Y^{(j)}(t)$,$j = 1, ..., n(t)$,

where $Y^{(j)}(t)$ denotes the observation corresponding to the jth individual at time t, we have

$$p(N(t)|\mathscr{F}_{t-1}) \propto exp[\theta_0(t)n(t) + \theta_1(t)(n_{10}(t) + n_{11}(t)) + \theta_2(t)(n_{01}(t) + n_{10}(t)) + \theta_3(t)n_{11}(t)]. \tag{5.3}$$

The log-likelihood function based on the contingency data $\{N(t)\}$, t=1,...,N, is then given by

$$lnL(\theta) = \sum_{t=1}^{N}[\theta_0(t)n(t) + \theta_1(t)(n_{10}(t) + n_{11}(t)) + \theta_2(t)(n_{01}(t) + n_{11}(t)) + \theta_3(t)n_{11}(t)]. \tag{5.4}$$

It is convenient to find the mean parameters via $\theta_i(t)$, i=1,2,3,:

$\mu_1(t) = E(Y_1(t)|\mathscr{F}_{t-1})$, $\mu_2(t) = E(Y_2(t)|\mathscr{F}_{t-1})$, and $\mu_3(t) = E(Y_1(t)Y_2(t)|\mathscr{F}_{t-1})$. Note that $\mu_i(t) = -\frac{\partial\theta_0(t)}{\partial\theta_i(t)}$, where $\theta_0(t) = log(p_{00})$ is a function of $\{\theta_i(t), i = 1, 2, 3\}$. We also have $\mu_1(t) = p_{10}(t) + p_{11}(t)$, $\mu_2(t) = p_{01}(t) + p_{11}(t)$, and $\mu_3(t) = p_{11}(t)$.

Consider the model

$$\theta_i(t) = \beta_{i0} + \beta_{i1}log(\frac{n_{01}(t-1)}{n_{00}(t-1)}) + \beta_{i2}log(\frac{n_{10}(t-1)}{n_{00}(t-1)}) + \beta_{i3}log(\frac{n_{11}(t-1)n_{00}(t-1)}{n_{01}(t-1)n_{10}(t-1)}), \tag{5.5}$$

i=1,2,3. The parameters $\theta_i(t)$ and hence $\mu_i(t)$ have natural interpretation for the contingency data. More specifically, if $\{\theta_i(t)\}$ are of main interest, Model 1 would be a useful model to consider.

Model 2 Multinomial-Logit Model

For notational simplicity, we now onwards label $n_{00}(t), n_{01}(t), n_{10}(t)$ and $n_{11}(t)$ as $n_1(t), n_2(t), n_3(t)$ and $n_4(t)$ respectively. Denote, as before, $N(t) = (n_1(t), n_2(t), n_3(t), n_4(t))'$. Suppose

$$p(N(t)|\mathscr{F}_{t-1}) \propto \prod_{i=1}^{4}(p_i(t))^{n_i(t)} \tag{5.6}$$

where $p_i(t) = P(Y(t) = i|\mathscr{F}_{t-1})$. Thus, conditional on $\mathscr{F}_{t-1}$, N(t) is assumed to have a multinomial distribution as in Model 1.

We now proceed to model $p_i(t)$ via the logit transformation. Let

$$\phi_i(t) = logit(p_i(t)) = log(\frac{p_i(t)}{p_4(t)}), \quad i = 1, 2, 3, \tag{5.7}$$

where $\phi_i(t) = \alpha_{i0} + \sum_{j=1}^{3} \alpha_{ij} U_j(t-1), i = 1, 2, 3$, and $U_j(t-1) = logit(\frac{n_i(t-1)}{n(t-1)}) = log(\frac{n_i(t-1)}{n_4(t)})$.

The likelihood function is given by

$$L(\alpha) = \prod_{t=1}^{N} \prod_{i=1}^{4} (p_i(t))^{n_i(t)}, \tag{5.8}$$

where $\{p_i(t)\}$ are modeled by (5.7). Note that Models 1 and 2 are both multinomial regression models with different parameterizations. If the probabilities $\{p_i(t)\}$ are of direct interest, Model 2 would be more appropriate.

Model 3 Markov Chain Model

Suppose $\{Y(t)\}$ is a Markov chain with states 00,01,10 and 11 labeled as 1,2,3,4, and homogeneous transition probabilities $\{p_{ij}(t)\}$ defined by $p_{ij} = p(Y_t = j | Y_{t-1} = i)$, i, j=1,2,3,4.

Assuming stationarity, we have

$$\pi_j(t) = p(Y_t = j) = \sum_{i=1}^{4} \pi_i(t-1) p_{ij}, \tag{5.9}$$

j=1,...,4. Given $\mathscr{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, ...)$, we may approximate the stationary probabilities $\pi_j(t)$ by

$$\pi_j^*(t) = \sum_{i=1}^{4} (\frac{n_i(t-1)}{n(t-1)}) p_{ij}. \tag{5.10}$$

Note that $\pi_j^*(t)$ is obtained from $\pi_j(t)$ by replacing $\pi_i(t-1)$ by the corresponding sample proportions $\frac{n_i(t-1)}{n(t-1)}$. Our assumed model now is

$$p(N(t) | \mathscr{F}_{t-1}) = \prod_{j=1}^{4} (\pi_j^*(t))^{n_j(t)}, \tag{5.11}$$

where $\{\pi_j^*(t)\}$ are defined by (5.10)which depend on the transition probabilities $\{p_{ij}\}$. Define

$$\theta_{ij} = logit p_{ij} = log(\frac{p_{ij}}{p_{i4}}), \quad i = 1, ..., 4, \ \text{J} = 1, 2, 3. \tag{5.12}$$

The likelihood function is then given by

$$\begin{aligned} L(\theta) &= \prod_{t=1}^{N} \prod_{j=1}^{4} (\pi_j^*(t))^{n_j(t)} \\ &= \prod_{t=1}^{N} \prod_{j=1}^{4} \sum_{i=1}^{4} (\frac{n_i(t-1)}{n(t-1)}) p_{ij}(\theta), \end{aligned} \tag{5.13}$$

where $p_{ij}(\theta) = (1 + \sum_{j=1}^{3} e^{\theta_{ij}})^{-1} e^{\theta_{ij}}$, i=1,...,4, j=1,2,3 and $p_{i4} = 1 - \sum_{j=1}^{3} p_{ij}(\theta) = (1 + \sum_{j=1}^{3} e^{\theta_{ij}})^{-1}$. If the one-step transition probabilities are of primary interest, Model 3 should be considered.

Model 4 A State Space Model

We now present a multinomial-Dirichlet state-space model for the count process $\{N(t)\}$. The observation densities are assumed to be conditionally independent multinomials:

$$p(N(t)|\pi(t)) \propto \prod_{i=1}^{4} (\pi_i(t))^{n_i(t)}, \quad t = 1, ..., N. \tag{5.14}$$

The state densities conditional on past data are assumed to be Dirichlet:

$$p(\pi(t)|\mathscr{F}_{t-1}) \propto \prod_{i=1}^{4} (\pi_i(t))^{r_i(t)-1} \tag{5.15}$$

where $r_i(t)$ are functions of the past data $\mathscr{F}_{t-1}$ and $\sum_{i=1}^{4} r_i(t) = n(t)$. From (5.14) and (5.15) the forecast density of N(t) given $\mathscr{F}_{t-1}$ is given by the multinomial-Dirichlet distribution:

$$\begin{aligned} p(N(t)|\mathscr{F}_{t-1}) &\propto \int_0^1 \int_0^1 \int_0^1 \int_0^1 \prod_{i=1}^{4} (\pi_i(t))^{n_i(t)+r_i(t)-1} d\pi_1(t) d\pi_2(t) d\pi_3(t) d\pi_4(t) \\ &\propto \prod_{i=1}^{4} \frac{\Gamma(n_i(t) + r_i(t))}{\Gamma(r_i(t))} \\ &\propto \prod_{i=1}^{4} \prod_{j=1}^{n_i(t)} (r_i(t) + j - 1). \end{aligned} \tag{5.16}$$

Note that $E(\pi_i(t)|\mathscr{F}_{t-1}) = \frac{r_i(t)}{n(t)} = \mu_i(t)$, say. We then have $E(n_i(t)|\mathscr{F}_{t-1}) = n(t)\mu_i(t) = r_i(t)$. We now model $\mu_i(t)$ via the logit link function, viz,

$$logit \mu_i(t) = U'(t-1)\beta_i, i = 1, 2, 3, \tag{5.17}$$

with $\mu_4(t) = 1 - \sum_{i=1}^{3} \mu_i(t)$, $U'(t-1) = (1, logit \frac{n_i(t-1)}{n(t-1)}, i = 1, 2, 3)$, and $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2}, \beta_{i3})'$. The likelihood function is given by

$$L(\beta) = \prod_{t=1}^{N} p(N(t)|\mathscr{F}_{t-1}) = \prod_{t=1}^{N} \prod_{i=1}^{4} \prod_{j=1}^{n_i(t)} (r_i(t) + j - 1). \tag{5.18}$$

In Model 4, the cell probabilities $\{\pi_i(t)\}$ are considered as random variables and $E(\pi_i(t)|\mathscr{F}_{t-1}) = \mu_i(t)$ are of primary interest.

5.3 Parameter Estimation

A common feature of all the four models introduced in Section 2 is that the count process $\{N(t)\}$ is a (vector) Markov process of first order having the property

$$p(N(t)|\mathscr{F}_{t-1}) = p(N(t)|N(t-1)). \tag{5.19}$$

The parameters of interest in Models 1 to 4 are different. For the purpose of parameter estimation, we may exploit the Markovity in (5.19) to study the asymptotic properties of ML estimators. The generic likelihood function for Models 1 to 4 is

$$L(\theta) = \prod_{t=1}^{N} p_\theta(N(t)|N(t-1)) \tag{5.20}$$

where θ is the vector of parameters in the model. The main regularity conditions for the consistency and asymptotic normality of the ML estimator $\hat{\theta}$ can be formulated in terms of the score vector $S_n(\theta) = \frac{\partial lnL(\theta)}{\partial \theta}$ and the sample information matrix $J_n(\theta) = -\frac{\partial^2 lnL(\theta)}{\partial \theta \partial \theta'}$, given below:

(c.1) $n^{-1}J_n(\theta) \xrightarrow{p} F(\theta)$, where $F(\theta)$ denotes the limiting Fisher information matrix which is assumed to be non-singular.

(c.2) $n^{-\frac{1}{2}}S_n(\theta) \xrightarrow{d} N(0, F(\theta))$.

(c.3) $n^{-1}[J_n(\theta_n) - J_n(\theta)] \xrightarrow{p} 0$, for any $\theta_n \xrightarrow{p} \theta$.

Condition (c.1) can be verified by using a law of large number(or checking the ergodicity of the Markov process $\{N(t)\}$, see, for instance, Tweedie (1975)). A martingale central limit theorem yields(c.2). See Hall and Heyde (1980). Various sufficient conditions are available in the literature to verify (c.3). See, for instance, Billingsley (1961). Under (c.1),(c.2) and (c.3) we have

Theorem 3.1 There exists a consistant solution $\hat{\theta}$ of the equation$S_n(\theta) = 0$ with probability tending to 1, and

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, F^{-1}(\theta). \tag{5.21}$$

See Billingsley(1961) for a proof.

It is to be noted that covariates, say $\{x_t\}$, can be included in Model 1,2,4 by simply adding a term x_t^T to the linear component of each model (5.5), (5.7) and (5.17). One can then use regularity conditions such as those of Kaufmann(1987) for the asymptotic theory. Since in Model 3, the transition probabilities $\{p_{uv}\}$ are themselves of interest and they are assumed to be independent of time, it would not be useful to include covariates in Model 3. In this paper we do not consider models with covariates.

5.4 Data Analysis

To illustrate the application of the above models, two sets of data are analyzed. The first example is from Baltimore Eye Survey Study and the second is about IBM stock price and Dow Jones Index data.

Example 4.1: Baltimore Eye Survey Study

In this example, a bivariate binary response is recorded for each subject indicating whether or not an eye (left and right) was visually impaired (VI) (vision less than 20/60) as defined by State of Maryland driving regulations. The explanatory variables include age in years grouped into 4 categories: 40-50, 51-60, 61-70, 70+; race (white, black) and eye (right, left). The scientific interest is to characterize the prevalence of VI in terms of age and race. See Liang, Zeger and Qaqish (1992). At each time t,where $t = \{40-50, 51-60, 61-70, 70+\}$, we create the contingency table for black and white separately as below:

Table 5.1: Contingency Table at time t, t={40-50,51-60, 61-70, 70+}

		Eye	
		Left	Right
Prevalence VI	Yes	$n_1(t)$	$n_2(t)$
	No	$n_3(t)$	$n_4(t)$

Table 5.2 is the time-dependent contingency table we create based on Table5.1

Table (5.3)lists the observed probabilities which are calculated by $p_i(t) = \frac{n_i(t)}{n(t)}$, i=1,2,3,4; t=2,3,4.

Table 5.2: Prevalence of VI left and right eyes for age and race combination in the Baltimore Eye Survey Study

Race	t	Age	$n_1(t)$	$n_2(t)$	$n_3(t)$	$n_4(t)$	n(t)
White	1	40-50	15	19	617	613	1264
	2	51-60	24	25	557	556	1162
	3	61-70	42	48	789	783	1662
	4	70+	139	146	673	666	1624
Black	1	40-50	29	31	750	748	1558
	2	51-60	38	37	574	575	1224
	3	61-70	50	49	473	474	1046
	4	70+	85	93	344	336	858

Table 5.3: The observed probabilities in Baltimore eye survey analysis

Race	Age	Observed probabilities p_i			
White	51-60	0.0190	0.0215	0.4793	0.4802
	60-71	0.0253	0.0289	0.4747	0.4711
	71+	0.0845	0.0900	0.4149	0.4106
Black	51-60	0.0310	0.0302	0.4690	0.4698
	61-70	0.0478	0.0468	0.4522	0.4532
	71+	0.0991	0.1084	0.4009	0.3916

We fit the data and do model selection for each type of proposed models. After that, we use the selected models to do probabilities predictions and the results of conditional exponential model, multimomial-logit model, Markov chain model and multinomial-Dirichlet model are listed in the following tables separately (Table (5.4), Table (5.5), Table (5.6) and Table (5.9)).

Table 5.4: Baltimore eye survey analysis results by mulitnomial-logit type model

Race	Age	Predicted probabilities ($\hat{p}_i$)			
White	51-60	0.021	0.029	0.4891	0.4669
	60-71	0.0250	0.0303	0.4794	0.4653
	71+	0.0829	0.0892	0.4104	0.4175
Black	51-60	0.0294	0.0293	0.4735	0.4678
	61-70	0.0499	0.0482	0.4636	0.4383
	71+	0.0624	0.0711	0.6006	0.2659

Table 5.5: Baltimore eye survey analysis results by conditional exponential model

Race	Age	Predicted probabilities($\hat{p}_i$)			
White	51-60	0.0216555	0.020848	0.4765396	0.480956
	61-70	0.0273647	0.027	0.472587	0.473047
	70+	0.079254	0.09064	0.4110337	0.4190715
Black	51-60	0.0304	0.02955	0.4703	0.4698
	61-70	0.04973	0.05019	0.4523	0.4478
	70+	0.09489	0.1068	0.3990	0.3993

Table 5.6: Baltimore eye survey analysis results by Markov chain model

Race	Age	Predicted probabilities($\hat{p}_i$)			
White	51-60	0.0397	0.0422	0.4600	0.4581
	60-71	0.0455	0.0509	0.4538	0.4499
	71+	0.0519	0.0556	0.4469	0.4455
Black	51-60	0.0426	0.0469	0.4552	0.4553
	61-70	0.0545	0.0566	0.4447	0.4442
	71+	0.0704	0.0723	0.4289	0.4284

For the Markov chain method, in addition to the predicted probabilities, we can also estimate one-step transition probabilities. These transition probabilities provide further insight into one-step changes in the contingency tables.

Table 5.7: One-step transition matrix for white people resulting from Markov chain model

Meaning of state	state	1	2	3	4
Left eye impaired	1	0.0316	**0.9668**	**0.0000**	0.0016
Right eye impaired	2	**0.9053**	0.0935	**0.0002**	0.0010
Left eye good	3	0.0037	0.0152	0.0427	**0.9384**
Right eye good	4	0.0493	0.0451	**0.9056**	**0.0000**

Now state 1 stands for the case that left eye is impaired; state 2 for right eye is impaired; state 3 is for left eye is not impaired; and state 4 is for right eye is not impaired. In Table (5.7), for white people, $p_{12} = 0.9668$ and $p_{21} = 0.9053$ meaning that if the patient's one eye is impaired, the other eye will be impaired with high probability later; while $p_{34} = 0.9384$

Table 5.8: One-step transition matrix for black people resulting from Markov chain model

Meaning of state	state	1	2	3	4
Left eye impaired	1	**1.0000**	**0.0000**	**0.0000**	0.0000
Right eye impaired	2	0.0005	**0.9995**	**0.0000**	0.0000
Left eye good	3	0.0263	**0.0000**	0.3406	0.6331
Right eye good	4	0.0235	0.0562	0.6067	0.3136

and $p_{43} = 0.9056$ indicating that if the patient's one eye is not impaired, the other eye won't be impaired either. On the other hand, in Table (5.8) for the black people, $p_{11} = 1.0000$ and $p_{22} = 0.9995$ means that one eye does not affect later the other eye. These are interesting observations from the estimated transition probabilities in Tables (5.7, 5.8).

Table 5.9: Baltimore eye survey analysis results by mutinomial-Dirichlet model

Race	Age	Predicted probabilities($\hat{p}_i$)			
White	51-60	0.0233	0.0224	0.500	0.4543
	60-71	0.0339	0.0416	0.4800	0.4445
	71+	0.0704	0.0755	0.4067	0.4475
Black	51-60	0.0313	0.0279	0.4552	0.4553
	61-70	0.05563	0.0542	0.4398	0.4497
	71+	0.0875	0.1015	0.4066	0.4044

From Tables (5.4, 5.5, 5.6, 5.9), we see that different methods perform differently when comparing the predicted probabilities with the observed probabilities. In general, conditional exponential model, multinomial-logit and multinomial- Dirichlet models outperform Markov chain model, but Markov chain model provides more information in term of the one-step transition probabilities; Conditional exponential model works very well for both white and black people, multinomial-logit model predicts better for white people while multinomial-Dirichlet model gives more precise predictions for black people.

Example 4.2: IBM and Dow Jones Index data

We collected IBM stock price and Dow Jones Index data from 2004 to 2006. For each quarter, we calculated the number of days they both increase, represented by n_{11}, they both drop, represented by n_{00}, number of days IBM stock prince dropped while Dow Jones Index went up,represented by n_{01} and IBM stock price went up while Dow Jones Index dropped,represented by n_{10}. Thus we created the contingency Table (5.10). The corresponding observed probabilities are listed in Table (5.11).

Table 5.10: Contingency Table for IBM & DowJone Data

Quarter	n_{00}	n_{01}	n_{10}	n_{11}	n(t)
1	28	5	11	18	62
2	27	13	6	16	62
3	26	8	11	19	64
4	20	6	12	26	64
5	27	9	11	14	61
6	28	12	8	16	64
7	26	13	5	20	64
8	25	9	11	18	63
9	24	13	9	16	62
10	30	8	6	19	63
11	25	7	10	21	63
12	19	9	15	20	63

Table 5.11: IBM and DJ data observed probabilities during 12 quarters

Quarter	Observed probabilities			
2	0.4355	0.2097	0.0968	0.2581
3	0.4063	0.1250	0.1719	0.2969
4	0.3125	0.0938	0.1875	0.4063
5	0.4426	0.1475	0.1803	0.2295
6	0.4375	0.1875	0.1250	0.2500
7	0.4063	0.2031	0.0781	0.3125
8	0.3968	0.1429	0.1746	0.2857
9	0.3871	0.2097	0.1452	0.2581
10	0.4762	0.1270	0.0952	0.3016
11	0.3968	0.1111	0.1587	0.3333
12	0.3016	0.1429	0.2381	0.3175

For each model (Model1 to 4), we fit the data and do model selection, the selected models are used to predict the probabilities and the results of four different models are listed in the Table (5.12, 5.13, 5.14, 5.15) below.

It must be noted that stock prices are notoriously difficult to predict from past data due to the random walk nature of the data. Nevertheless, the three models do provide some good estimates of the prediction probabilities.

Table 5.12: IBM and DJ data analysis results by conditional exponential model

Quarter	Predicted probabilities ($\hat{p}_i$)			
2	0.4003	0.1598	0.1489	0.2909
3	0.3973	0.1432	0.1469	0.3127
4	0.4007	0.1621	0.1491	0.2882
5	0.4017	0.1731	0.1501	0.2752
6	0.4005	0.1609	0.1490	0.2896
7	0.3988	0.1504	0.1478	0.3030
8	0.3963	0.1391	0.1463	0.3183
9	0.4008	0.1633	0.1492	0.2867
10	0.4002	0.1584	0.1487	0.2927
11	0.3966	0.1402	0.1465	0.3168
12	0.4004	0.1604	0.1489	0.2903

Table 5.13: IBM and DJ data analysis results by Markov chain model

Quarter	Predicted probabilities($\hat{p}_i$)			
2	0.3691	0.1749	0.1505	0.3054
3	0.4157	0.1442	0.1350	0.3050
4	0.3876	0.1707	0.1520	0.2896
5	0.3775	0.1618	0.1997	0.2610
6	0.3970	0.1834	0.1220	0.2976
7	0.4089	0.1572	0.1313	0.3025
8	0.4119	0.0861	0.1594	0.3427
9	0.3953	0.1730	0.1465	0.2851
10	0.4205	0.1637	0.1334	0.2824
11	0.3809	0.1391	0.1566	0.3233
12	0.3809	0.1604	0.1687	0.2900

Table 5.14: IBM and DJ data analysis results by multinomial-logit type model

Quarter	Predicted probabilities($\hat{p}_i$)			
2	0.3805	0.1984	0.1391	0.2820
3	0.4084	0.1410	0.1396	0.3110
4	0.3993	0.1650	0.1458	0.2899
5	0.3794	0.1379	0.2140	0.2688
6	0.4112	0.1891	0.1067	0.2929
7	0.4116	0.1577	0.1267	0.3040
8	0.3926	0.1170	0.1816	0.3088
9	0.4070	0.1654	0.1387	0.2889
10	0.4222	0.1556	0.1282	0.2940
11	0.3794	0.1406	0.1672	0.3128
12	0.3882	0.1527	0.1701	0.2891

Table 5.15: IBM and DJ data analysis results by mutinomial-Dirichlet model

Quarter	Predicted probabilities($\hat{p}_i$)			
2	0.3796	0.1783	0.1435	0.2985
3	0.4053	0.1441	0.1419	0.3088
4	0.3978	0.1662	0.1487	0.2873
5	0.3750	0.1415	0.2167	0.2667
6	0.4131	0.1890	0.1107	0.2872
7	0.3863	0.2089	0.1224	0.2824
8	0.3868	0.1211	0.1816	0.3105
9	0.4060	0.1676	0.1430	0.2835
10	0.4212	0.1601	0.1343	0.2844
11	0.3747	0.1405	0.1633	0.3215
12	0.3852	0.1541	0.1710	0.2897

5.5 Concluding Remarks

In this chapter, we have presented four related models for time series of (2x2) contingency tables. The models exploit the basic multinomial structure of the data and introduce time-dependence using a regression type approach. A state space model is also introduced using an observation-driven state process. The models are applied to two real data sets. Maximum likelihood estimation and model selection are discussed for the data applications, and the prediction probabilities are computed for the four models.

Bibliography

Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

Andersen, P.K.and Gill, R.D. (1982). Cox's regression model for counting processes:A large sample study. *Ann.Statist.* 10, 1100-1120.

Arjas, E.and Haara, P. (1987). A logistic regression model for hazard:Asymptotic results. *Scand. J.Statist.* 14, 1-18.

Billingsley, P. (1961). *Statistical Inference for Markov Processes.* Chicago Univ. Press.

Cox, D.R. (1972). Regression models and life tables.*J. Roy. Statist. Assoc.* 39, 357-365.

Cox, D.R. (1975). Partial likelihood. *Biometrika* 62, 69-76.

Fahrmeir, L. (1992). State space modeling and conditional mode estimation for categorical time series. In IMA volume *New Directions in Time Series Analysis* (Edited by D.Brillinger et al.), 87-109, Springer-Verlag, New York.

Fahrmeir, L.(1992a). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *J of the Amer. Statist. Assoc.* 87, 501-509.

Fahrmeir, L.and Kaufmann, H. (1987). Regression models for non-stationary categorical time series.*J.Time Series Anal.* 8,147-160.

Fahrmeir, L.and Kaufmann,H.(1991). On Kalman filtering, posterior mode estimation and fisher scoring in dynamic exponential family regression. *Metrika* 38, 37-60.

Fahrmeir, L.and Tutz, G. (2001). *Multivariate Statistical Modeling Based on Generalized Linear Models.* 2nd ed. Springer, New York.

Fokianos, K. and Kedem, B . (1998). Prediction and classification of non-stationary categorical time series. *J. of Multivariate Anal.* 67, 277-296.

Fokianos, K. and Kedem, B. (2003). Regression Theory for Categorical Time Series. *Statistical Science.* 18, 357-376,

Hall, P.G. and Heyde, C.C. (1980). *Martingale Limit Theory and Its Applications.* Academic Press, New York.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series:Asymptotic estimation theory. *Ann. Statist.* 15,79-98.

Kedem, B. (1980). *Binary Time Series.* Dekker, New York.

Kedem, B. and Fokianos, L. (2002). *Regression Models for Time Series Analysis.* Wiley, New Jersey.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data. *J.R.Statist. Soc.B* 54, No.1, 3-40.

Sage, A.and Melsa, J. (1971). *Estimation Theory, with Applications to Communications and Control.* New York: McGraw-Hill.

Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis and Its Applications.* Springer, New York.

Slud, E. (1992). Partial likelihood for continuous-time stochastic processes. *Scand. J.Statist.* 19, 97-109.

Slud, E.and Kedem, B. (1994). Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica.* 89-106.

Tweedie, R.L. (1975). Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stoch. Proc. & Applns.* 3,385-403.

Waterman, M.S. (1995). *Introduction to Computational Biology: Maps,Sequences and Genomes.* Chapman and Hall, New York.

Wong, W.H.(1986). Theory of partial likelihood. *Ann.Statist.* 14,88-123.