THESIS REPORT

Ph.D.

Categorical Time Series: Prediction and Control

by K. Fokianos
Advisor: B. Kedem

Ph.D. 96-7

INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Abstract

Title of Dissertation: Categorical Time Series: Prediction and Control

Konstantinos Fokianos, Doctor of Philosophy, 1996

Dissertation directed by: Professor Benjamin Kedem
Department of Mathematics
Statistics Program

We study regression models for nonstationary categorical time series and their applications, and address the issues of prediction, estimation and control. Genaralized Linear Models and Partial Likelihood are the basic tools in the present study. The models link the probabilities of each category to a covariate process through a vector of time invariant parameters. Under mild regularity conditions, asymptotic properties of the estimators are established by appealing to martingale theory, and certain diagnostic tools are presented for checking the model adequacy. The methodology is demonstrated using real rainfall data. Subsequently we discuss a new recursive estimation method for time series following generalized linear models, motivated by the logistic regression model in conjunction with binary time series. The estimation procedure, suitably modified, gives rise to a stochastic approximation scheme used here to illustrate a connection between control theory and generalized linear models.

Categorical Time Series: Prediction and Control

by

Konstantinos Fokianos

Dissertation submitted to the Faculty of the Graduate School
of The University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1996

Advisory Committee:

Professor Benjamin Kedem, Chairman/Advisor
Professor Eric Slud
Professor Abram Kagan
Associate Professor Paul J. Smith
Assistant Professor Athanassios Dimas

Dedication

To my family

Acknowledgements

I would like to express my deep appreciation and thankfulness to my mentor and guide Professor Benjamin Kedem. He has contributed to this dissertation in countless ways. His knowledge, insight and enthusiasm in statistics has been inspiring, and his guidance, invaluable.

I am also indebted to my mother, my sister and my grandmother, and to my good friend Carol. They all offered me unlimited support and constant encouragement when I needed it. This work would not have been possible without their understanding.

Finally, I would like to express my gratitude to Professor Eric Slud and Associate Professor Paul Smith. They both advised me and encouraged me during my graduate studies. Their constructive remarks and suggestions improved this dissertation. I also want to thank Professor Abram Kagan and Assistant Professor Thanassis Dimas for their useful suggestions.

This research was partially supported by a scholarship of Pateras Foundation, Pireas, Greece and by NASA grant NAG52783.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Categorical Time Series

A time series is a collection of random variables, say $\{Y_t\}$, $t = 1, \ldots, N$, ordered in time. The seminal texts [17], [19], [50] [79], to name a few, provide an introduction to the subject while discussing analysis, forecasting and control. The assumptions of stationarity, Gaussianity and linearity dominate the results in this area. Recently, however, various attempts have been made to weaken these assumptions; see [80]. Stationarity, in its strongest form means that ([50, p. 67])

$$P(Y_{t_1} \leq y_1, \ldots, Y_{t_n} \leq y_n) = P(Y_{t_1+h} \leq y_1, \ldots, Y_{t_n+h} \leq y_n)$$

for all n, $t_1, \ldots, t_n$, $y_1, \ldots, y_n$ and h where $\{Y_t\}$ is a sequence of random variables.

In addition to the aforementioned assumptions, current time series methods are usually applicable for continuous valued time series, especially Gaussian time series. However, categorical time series arise in numerous practical applications, many of which are reported in the recent books [28], [33], [50, Ch. 9], and [61]. Examples of categorical time series include signals quantized at several levels, clipped binary time series, and any multi-response longitudinal data observed on ordinal or nominal scales. For instance, [65] examines data on repeated

choices between discrete alternatives. Another example is the amount of daily rainfall divided into three categories, low/medium/high. A linear model may not fit such data but a categorical prediction model can accommodate variations of dynamic structure (see [92]). And just as with "ordinary" time series, the problem of forecasting or prediction in categorical series is of importance, except that usually it concerns the estimation of a future transition probability given past data and auxiliary information. In this regard, the prediction problem is essentially synonymous with the problem of classification of a future value into one of several categories, given the past.

Methods for statistical inference concerning such data are less well developed. Some notable exceptions are [45], [46], [49], [53] and [64]. In [45] and [46] the authors approach the subject by appealing to ARMA (Autoregressive Moving Average) methodology. They replace the linear combinations of the continuous-valued case with a probabilistic mixture and call their models discrete autoregressive moving average models (DARMA). In [49] and [64] the authors model binary time series by clipping an underlying Gaussian process. In [53], the author assumes that an underlying time series (unobservable) of continuous-valued data generates the time series of discrete valued data. Another approach is to assume the process is a homogeneous Markov chain. Then the inference problem can be attacked using the methods of [13].

Recent advances in categorical time series owe a great deal to the introduction of generalized linear models and link functions as described in [67]. Generalized Linear Models (GLM)—introduced in [73]—are regression models for a response measured along with some covariates. Under the GLM formulation, the response has a distribution which is member of the exponential family. In addition some

monotonic differentiable function of the expected value of the response–called the link function–is expressed as linear combination of covariates. We present a brief description of these models in Section 2.4.

In the context of categorical time series, one can parametrize the one step transition probability conveniently via the link, and this goes along well with conditional inference, allowing for some form of non-stationarity. The idea goes back to Cox [25].

The work that has already been done in this area is based, implicitly or explicitly, on the use of conditional likelihood by utilizing the Markov assumption. Examples can be found in [16], [32], [47], [54], [60], [71] and [94]. For the case of count data see [102], and for a more general setup see [59] , [104] and [106]. The general idea in all these works is to link the transition probabilities, via a suitable function, to the past responses and covariate information. Then a conditional likelihood can be calculated explicitly and methods from independent data carry over to the dependent data. However, the conditional likelihood approach implies that the covariates are fixed during the period of observation.

The first part of this thesis studies regression models for nonstationary categorical time series with random time dependent covariates, without any Markov assumption. It generalizes the work in [92], where only logistic regression for binary time series is considered. We perform conditional inference using partial likelihood, a concept introduced in [23], and extended and ramified in [91] and [100]. Partial likelihood simplifies conditional inference—for example, it obviates the Markov assumption—and is particularly useful for time series, where the dependence is unknown, let alone knowledge of joint distributions. We prove under some assumptions (Theorem 2.6.1) that the maximum partial likelihood

estimator exists and is unique. In addition, it is consistent and asymptotically normally distributed.

1.2 Partial Likelihood

This is a brief description of the partial likelihood concept. Assume that we observe a stochastic process, say (x_t, y_t), $t = 1, \ldots, N$. In principle, we can write down the joint distribution of all the observations up to time N, by employing the law of total probability; that is ([100])

$$f(x_1, y_1, x_2, y_2, \ldots, x_N, y_N) = [\prod_{t=1}^{N} f(y_t \mid d_t)][\prod_{t=1}^{N} f(x_t \mid c_t)] \tag{1.1}$$

where $d_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1})$ and $c_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1}, y_t)$.

The second product on the right hand side of (1.1) was defined in [23] as the Partial Likelihood. It is helpful to note that the σ-field generated by c_{t-1} is contained in the one generated by c_t. This is a key feature which motivates our definition (see [91] and [92]):

Definition 1.2.1 Let $\mathcal{F}_t$, $t = 0, 1, \ldots$ be an increasing sequence of σ-fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \ldots$, and let $X_1, X_2, \ldots$ be a sequence of random variables on some common probability space such that X_t is $\mathcal{F}_t$ measurable. Denote the density of X_t, given $\mathcal{F}_{t-1}$, by $f_t(x_t; \beta)$, where $\beta \in R^p$ is a parameter. The partial likelihood (PL) function relative to β, $\mathcal{F}_t$, and the data $X_1, X_2, \ldots, X_N$, is given by the product

$$PL(\beta; X_1, \ldots, X_N) = \prod_{t=1}^{N} f_t(x_t; \beta) \tag{1.2}$$

This definition generalizes both likelihood and conditional likelihood. Unlike (full) likelihood, partial likelihood does not require complete knowledge of the

joint distribution of the covariates. Unlike conditional likelihood, complete covariate information need not be known throughout the period of observation. Partial likelihood takes into account only what is known to the observer up to the time of actual observation.

The vector β that maximizes (1.2) is called the maximum partial likelihood estimator (MPLE). Its asymptotic distribution has been studied by several authors (see [92] or [100]). In the context of survival analysis and counting processes see [4] or [7], for example. The key point is that the gradient of the logarithm of (1.2) is a martingale with respect to the nested sequence of histories $\mathcal{F}_t$.

1.3 Motivation

The original motivation for this study is due to an important problem from the field of meteorology, namely, the prediction of rainfall from spaceborne precipitation radar. Let us be more specific. Due to various technological constraints, the effective dynamic range of a spaceborne precipitation radar (PR) flying at an altitude of 350 km is limited at present to intermediate values of rain rate. In particular at high rain rates–the source of most of the rainfall volume–there is a degraded signal to noise ratio due to large attenuation (see [69]). Basically this means the spaceborne PR saturates at some intermediate value (roughly 10-15 mm/hr) so that high rain rates are indistinguishable from lower rates. It is therefore useful to construct methods that can help a PR discern instantaneously high rain rates using covariate information.

On the other hand, instantaneous tropical rain rate snapshots obtained from a radar over a given large area show that the instantaneous fraction of the area (FA) where rain rate exceeds a given threshold and the instantaneous area

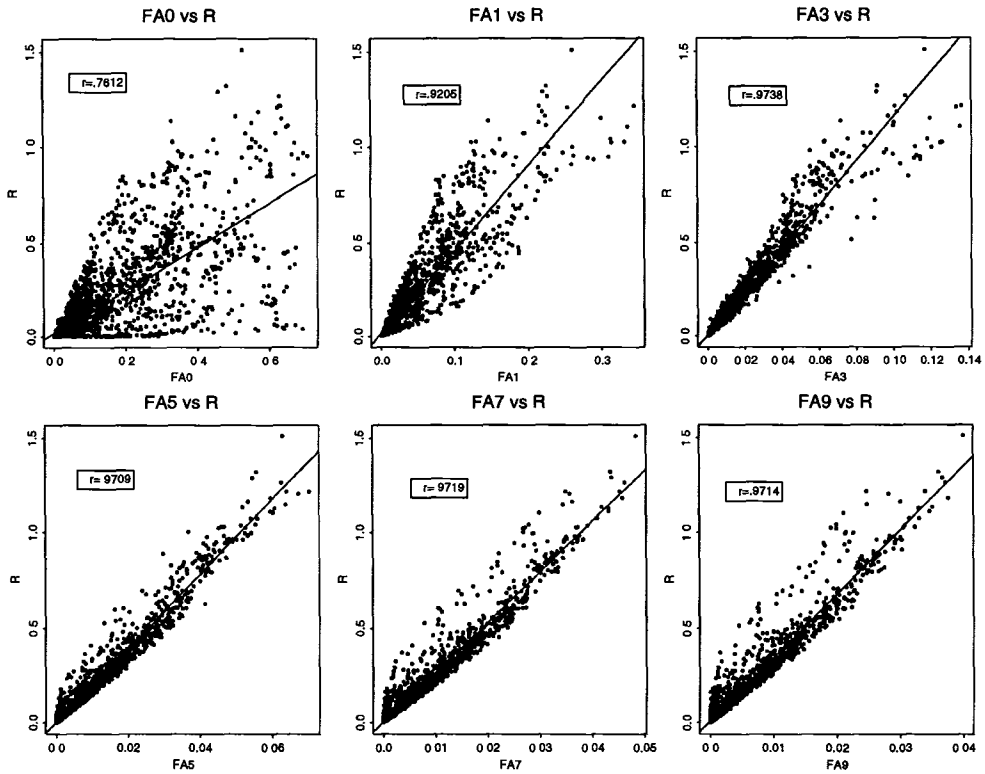

Figure 1.1: Area average rain rate versus fractional area for TOGA/COARE Data.

average (R) are highly correlated quantities. An example of this fact is shown in figure 1.1. for 3400 pairs (FA,R) from TOGA/COARE data (see Appendix A). We see for a threshold in the range of 3 to 9 mm/hr that the correlation is quite high, exceeding 97%. The same has been observed time and again using different radar data sets of tropical rainfall from different parts of the globe, yielding correlations well above 95% and at times reaching 99%. In [88] the authors have observed the same high degree of linearity from rain-gauge data. This seemingly puzzling phenomenon has attracted a fair amount of attention in the scientific community and several attempts to explain it have been suggested including [51], [52], [70] and [84].

We will illustrate in Chapter 3 that the fractional area is a useful covariate in categorical prediction of the area average rain rate. This has an application in rainfall measurement from imprecise satellite-borne instruments, for it is still possible to classify rain rate reliably as being above or below a threshold and to determine accordingly the fractional area and from it the area average. Moreover, in Section 2.7 we wil give a diagnostic tool for assessing the fit of these models.

So, the present work was originally motivated by the need to classify or predict rain rate in several categories given the values of some covariates. However, we address the more general problem of predicting categorical time series.

1.4 Controlling a Probability

Partial likelihood–as explained in Section 1.2–is an off-line estimation method. However, it can be modified to accommodate sequential processes. There are many occasions when data are collected sequentially in time, and one needs to update an estimator–in our context the estimator of the regression coefficients–every time a new observation becomes available. An illustration of this phenomenon is apparent in control theory employed by engineers and scientists in modeling and controlling physical systems. By controlling a system, we mean choosing optimally an input to the system so that the output is close to some predetermined value. There is a well developed theory for the so called ARMAX (Autoregressive Moving Average with External Input Models) models. In this theory, the system is specified linearly in both the parameters and observations. Then, the control input is chosen so that a certain cost function, for example, the variance of the difference between output and target, attains its optimal value. We give a very brief review of the subject in Section 4.2.

Since our methods generalize the results from the classical theory of time series, it is quite natural to ask how and in what sense one can control categorical time series.

The second part of this thesis studies the adaptive control of binary time series. We use a very simple model, the so called logistic model, to control the transition probabilities of the system. Actually, this gives a method for the adaptive control of a nonstationary Markov chain when the action space is uncountable. We first give a new set of recursion formulas for the update of the regression coefficients and the control law. We show that these recursions possess crucial geometric properties that can be utilized for proving not only self-optimality of the control but self-tuning as well.

Chapter 2

Regression Models for Nonstationary Categorical Time Series: Partial Likelihood Inference

2.1 Introduction

In this chapter, a general regression model for nonstationary categorical time series is proposed. The model links the probabilities of each category to a covariate process through a vector of time invariant unknown parameters. This extends ideas from generalized linear models (see [67]) and provides a convenient framework for the analysis of categorical time series. Partial likelihood, as introduced in Section 1.2, is used for parameter estimation. We establish good asymptotic properties of the estimator, under mild regularity conditions, based on the fact that the partial score is a martingale. We emphasize that no Markovian assumption is necessary. Finally, we propose goodness of fit statistics to examine the quality of the fit.

2.2 The General Model

Suppose that we observe a nonstationary categorical time series, say $\{y_t,\ t = 0, 1, \ldots, N\}$. Let m denote the number of possible categories and assume that

the tth observation is given by a vector $y_t = (y_{t1}, \ldots, y_{tq})'$ of length $q = m - 1$, where

$$y_{tj} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ category is observed at time } t \\ 0 & \text{otherwise} \end{cases}$$

Let $p_t = (p_{t1}, \ldots p_{tq})'$ denote the corresponding vector of conditional probabilities given $\mathcal{F}_{t-1}$. In other words $p_{tj} = P(y_{tj} = 1 \parallel \mathcal{F}_{t-1})$ for $j = 1, \ldots, q$. The σ-algebra $\mathcal{F}_{t-1}$ represents all information available to the observer up to and including time t. It follows that

$$y_{tm} = 1 - \sum_{j=1}^{m-1} y_{tj} \tag{2.1}$$

and

$$p_{tm} = 1 - \sum_{j=1}^{m-1} p_{tj} \tag{2.2}$$

Equation (2.1) is due to the fact that an observation belongs to one and only one category at each time t. Equation (2.2) follows either by just taking conditional expectations in (2.1) or upon noting that $\sum_{j=1}^{m} p_{tj} = 1$.

Assume that $\mathbf{Z}_{t-1}$ is a $p \times q$ matrix that represents a covariate process. In other words each response y_{tj} corresponds to a vector of random time dependent covariates, say $z_{(t-1)j}$, which is the jth column of $\mathbf{Z}_{t-1}$. The covariate matrix usually consists of lagged values of the response process and/or exogenous variables that evolve in time simultaneously with the response variable. Moreover, lagged values of the exogenous variables are allowed as well as any interactions between the response and the covariates. This will become clear in the next chapter where specific examples are discussed.

The aim of this chapter is to develop an asymptotic theory for a flexible and parsimonious class of models that link the probability of the jth category with

the covariate process in a certain way. This leads to an attractive parametrization, which extends ideas from the generalized linear models (GLM) and the autoregressive moving average models (ARMA) ([17]).

Define (see [32], [47] and [92])

$$p_t(\beta) = h(\mathbf{Z}'_{t-1}\beta) \tag{2.3}$$

Here β denotes a p dimensional vector of time invariant unknown parameters which belongs to an open set $B \subseteq R^p$. The function h is called the link function. We assume that the link function maps a subset $H \subseteq R^q$ bijectively (one-to-one) onto $\{(w_1, \ldots, w_q)' : \ w_j > 0, \ j = 1, \ldots, q, \ \sum_{j=1}^{q} w_j < 1\}$.

The above setup leads to some basic results that we will use throughout this chapter. These are the main subject of the following section.

2.3 Some Simple Properties

Before proving any result, we would like to emphasize that expectation and variance are calculated under a family of probabilities measures indexed by β, where $\beta \in B$ for some admissible subset B of R^p. We state these properties in the form of lemmas.

Lemma 2.3.1 We have

$$E[y_t \parallel \mathcal{F}_{t-1}] = p_t(\beta).$$

Proof: It follows from the definition of conditional expectation. □

Lemma 2.3.2 We have

$$E[\mathbf{Z}_{t-1}(y_t - p_t(\beta)) \parallel \mathcal{F}_{t-1}] = 0,$$

where 0 is a $p \times 1$ vector, with each element identical to 0.

Proof: This follows from Lemma 2.3.1 and from the fact that $\mathbf{Z}_{t-1}$ is $\mathcal{F}_{t-1}$ measurable. □

Lemma 2.3.3 If $s < t$, then

$$E[\mathbf{Z}_{s-1}(y_s - p_s(\beta))(y_t - p_t(\beta))'\mathbf{Z}'_{t-1} \parallel \mathcal{F}_{s-1}] = \mathbf{0},$$

where $\mathbf{0}$ is the $p \times p$ null matrix.

Proof: Upon noting that for $s < t$ we have that $\mathcal{F}_{s-1} \subset \mathcal{F}_{t-1}$, the definition of conditional expectation and its properties lead to

$$\begin{aligned}
& E[\mathbf{Z}_{s-1}(y_s - p_s(\beta))(y_t - p_t(\beta))'\mathbf{Z}'_{t-1} \parallel \mathcal{F}_{s-1}] \\
&= E[E[\mathbf{Z}_{s-1}(y_s - p_s(\beta))(y_t - p_t(\beta))'\mathbf{Z}'_{t-1}] \parallel \mathcal{F}_{t-1}] \parallel \mathcal{F}_{s-1}] \\
&= E[\mathbf{Z}_{s-1}(y_s - p_s(\beta))E[(y_t - p_t(\beta))'\mathbf{Z}'_{t-1} \parallel \mathcal{F}_{t-1}] \parallel \mathcal{F}_{s-1}] \\
&= \mathbf{0}
\end{aligned}$$

The last equality follows from Lemma 2.3.2. □

Lemma 2.3.4 We have that

$$E[\mathbf{Z}_{t-1}(y_t - p_t(\beta))(y_t - p_t(\beta))'\mathbf{Z}'_{t-1}] = E[\mathbf{Z}_{t-1}\mathbf{\Sigma}_t(\beta)\mathbf{Z}'_{t-1}],$$

where $\mathbf{\Sigma}_t(\beta)$ is the conditional covariance matrix of y_t with generic element

$$\sigma_t^{(ij)}(\beta) = \begin{cases} -p_{ti}(\beta)p_{tj}(\beta) & \text{if } i \neq j \\ p_{ti}(\beta)(1 - p_{ti}(\beta)) & \text{if } i = j \end{cases}$$

for $i, j = 1 \ldots, q$.

Proof: The above expression follows again from properties of conditional expectation:

$$\begin{aligned} & E[\mathbf{Z}_{t-1}(y_t - p_t(\beta))(y_t - p_t(\beta))'\mathbf{Z}'_{t-1}] \\ = \; & E[E[\mathbf{Z}_{t-1}(y_t - p_t(\beta))(y_t - p_t(\beta))'\mathbf{Z}'_{t-1}] \parallel \mathcal{F}_{t-1}], \end{aligned}$$

and then by noticing that

$$E[(y_t - p_t(\beta))(y_t - p_t(\beta))' \parallel \mathcal{F}_{t-1}] = \mathbf{\Sigma}_t(\beta)$$

with $\mathbf{\Sigma}_t(\beta)$ as above. □

2.4 Generalized Linear Models

We are going to give a brief overview of generalized linear models. This will help us to compute useful quantities involved in asymptotics under our setup. A comprehensive account of these models can be found in [67]. The original definition is in [73].

Remark 2.4.1 The notation in the next section is general, and has no connection with the previous notation.

2.4.1 Definition and Maximum Likelihood Inference

Let y be a q-dimensional random variable distributed according to a natural exponential family. That means that its density is, with respect to a σ-finite measure ν,

$$f(y;\theta) = c(y)\exp(\theta' y - b(\theta))$$

where $\theta \in R^q$ is a parameter and $c \geq 0$ is measurable. We call $\Theta = \{\theta : 0 < \int c(y)\exp(\theta' y)d\nu(y) < \infty\}$ the natural parameter space. It is well known that Θ

is convex, and in the interior $\overset{\circ}{\Theta}$ of Θ all derivatives of $b(\theta)$ and all moments of y exist (assuming that $\overset{\circ}{\Theta} \neq \emptyset$) (see [24], [58]). Actually, $E_\theta(y) = \mu(\theta) = \partial b(\theta)/\partial\theta$ and $Cov_\theta(y) = \partial^2 b(\theta)/\partial\theta\partial\theta' = \mathbf{\Sigma}(\theta)$, say. The covariance matrix is assumed to be positive definite in the interior of Θ, implying that the restriction of $\mu(\theta) = E_\theta(y)$ to $\overset{\circ}{\Theta}$ is injective. Denote by M the image $\mu(\overset{\circ}{\Theta})$ of $\overset{\circ}{\Theta}$.

Generalized linear models are defined by the following three components:

1. **The Random Component**

   The $\{y_n\}$ are independent realizations of y with densities belonging to the exponential family

$$f(y_n; \theta_n) = c(y_n)\exp(\theta_n' y_n - b(\theta_n)), \quad \theta_n \in \overset{\circ}{\Theta}$$

2. **The Systematic Component**

   The matrix $\mathbf{Z}_n$ affects y_n through the linear form

$$\gamma_n = \mathbf{Z}_n'\beta,$$

   where $\mathbf{Z}_n$ is a $p \times q$ matrix of fixed covariates and β is a p-dimensional unknown vector of parameters.

3. **The Link Function**

   The systematic component is linked to the random component by a smooth invertible function g with $g : M \to R^q$ in the following way:

$$\gamma_n = g(\mu(\theta_n)).$$

Remark 2.4.2 If $g = \mu^{-1}$ then $\gamma_n = \theta_n$. In such a case, g is called the canonical link.

Maximum likelihood is the estimation method that is used for fitting these models. We can see that the log-likelihood of a sample of independent observations $y_1, \ldots, y_n$ is given by

$$l_n(\beta) = \sum_{i=1}^{n}(\theta_i' y_i - b(\theta_i)) - C, \quad \theta_i = u(\mathbf{Z}_i'\beta), \; i = 1, \ldots, n \tag{2.4}$$

where $u = (g \circ \mu)^{-1}$ and C does not depend on β. Let $\mu_n(\beta) = \mu(u(\mathbf{Z}_n'\beta))$, $\mathbf{\Sigma}_n(\beta) = \mathbf{\Sigma}(u(\mathbf{Z}_n'\beta))$, $\mathbf{U}_n(\beta) = [\partial u(\mathbf{Z}_n'\beta)/\partial\gamma_n]'$. Differentiating (2.4), we get the score function $s_n(\beta)$ and the information matrix $\mathbf{I}_n(\beta)$, respectively

$$s_n(\beta) = \frac{\partial l_n(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{U}_i(\beta)(y_i - \mu_i(\beta)), \tag{2.5}$$

$$\mathbf{I}_n(\beta) = cov_\beta(s_n(\beta)) = \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{U}_i(\beta) \mathbf{\Sigma}_i \mathbf{U}_i'(\beta) \mathbf{Z}_i'. \tag{2.6}$$

Further differentiation shows that

$$\mathbf{H}_n(\beta) = -\frac{\partial^2 l_n(\beta)}{\partial\beta\partial\beta'} = \mathbf{I}_n(\beta) - \mathbf{R}_n(\beta), \tag{2.7}$$

where

$$\mathbf{R}_n(\beta) = \sum_{i=1}^{n}\sum_{r=1}^{q} \mathbf{Z}_i \mathbf{W}_{ir}(\beta) \mathbf{Z}_i'(y_{ir} - \mu_{ir}(\beta)),$$

$y_i = (y_{i1}, \ldots, y_{iq})'$ and $\mathbf{W}_{ir}(\beta) = [\partial^2 u_r(\mathbf{Z}_i'\beta)/\partial\gamma_i\partial\gamma_i']$. For natural link functions the above formulas simplify to

$$s_n(\beta) = \sum_{i=1}^{n} \mathbf{Z}_i(y_i - \mu_i(\beta)), \quad \mathbf{I}_n(\beta) = \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{\Sigma}_i(\beta) \mathbf{Z}_i', \quad \mathbf{H}_n(\beta) = \mathbf{I}_n(\beta). \tag{2.8}$$

The maximum likelihood estimator $\hat{\beta}$ is given as the solution of the equations $s_n(\beta) = 0$. Generally speaking this solution does not guarantee a global maximum. However, for many important applications the log-likelihood is concave so that $\hat{\beta}$ corresponds to a global maximum. Furthermore, if the likelihood is

strictly concave, the maximum likelihood estimator is unique. Existence and uniqueness of the maximum likelihood estimator has been studied by several authors. Among them are [2], [42], [48], [90], [99]. Large sample properties of the estimator have been studied extensively in [31].

2.4.2 Iterative Reweighted Least Squares

We saw that the MLE is given as solution to the following system of non-linear equations

$$s_n(\beta) = 0.$$

An iterative procedure can be used to locate the estimator. The most widely used method is Fisher scoring. This updates the new estimator by using the Fisher information matrix rather than the second derivative of the log-likelihood in the Newton-Raphson iterations. In other words the recursions are given by

$$\hat{\beta}_{(k+1)} = \hat{\beta}_{(k)} + \mathbf{I}_n^{-1}(\hat{\beta}_{(k)})s(\hat{\beta}_{(k)}) \;\; k = 0, 1, \ldots, \tag{2.9}$$

where $\hat{\beta}_{(0)}$ is an initial guess. The algorithm terminates if a certain convergence criterion is satisfied, for example, if the absolute difference of the results of two iterations is less than a small positive number. Notice that in the case of the canonical link, the Fisher information coincides with the negative second derivative of the log-likelihood. Then the iterations (2.9) reduce to the Newton-Raphson method.

One can give a nice interpretation to these recursions. Define the observations (see [1], [67])

$$\tilde{y}(\beta) = (\tilde{y}_1(\beta), \ldots, \tilde{y}_n(\beta))$$

$$\tilde{y}_i(\beta) = \mathbf{Z}_i'\beta + \mathbf{U}_i^{-1}(y_i - \mu_i(\beta))$$

Then, it can be shown that the above recursions can be written as

$$\hat{\beta}_{(k+1)} = (\mathbf{Z}'\mathbf{W}_{(k)}\mathbf{Z})^{-1}\mathbf{Z}\mathbf{W}_{(k)}\tilde{y}_k \tag{2.10}$$

with

$$\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_n]$$

$$\mathbf{W} = \text{diag}(\mathbf{W}_i)$$

$$\mathbf{W}_i = \mathbf{U}_i'\mathbf{\Sigma}_i\mathbf{U}_i$$

and $\mathbf{W}_{(k)}$, $\tilde{y}_k$ mean evaluation of $\mathbf{W}$ and $\tilde{y}$ at $\beta = \hat{\beta}_{(k)}$. This shows that at each iteration the updated estimator is a solution to a weighted least squares problem. However, note that we update the weights at each iteration, whence the name iterative reweighted least squares.

2.5 Partial Likelihood Estimation

Now we compute the standard quantities that are involved in asymptotics for partial likelihood estimation, based on the results of Section 2.4. More specifically we will calculate the partial score function, the conditional covariance matrix, and the second derivative of the log-likelihood, establishing some notation. But first let us give the following definition:

Definition 2.5.1 Let $x = (x_1, \ldots x_q)'$ belong to R^q with $0 < x_i < 1$ for every $i = 1, \ldots, q$ and $\sum_{i=1}^q x_i < 1$. The logit function of x is defined as

$$l(x) = \left[\log(\frac{x_1}{1 - \sum_{i=1}^q x_i}), \ldots \log(\frac{x_q}{1 - \sum_{i=1}^q x_i})\right]$$

We pay special attention to this function since it is the canonical link function for the multinomial distribution.

Assume now that the general model, which was described in Section 2.2 holds for $\beta \in B \subset R^p$. Then, the partial likelihood (Section 1.2) based on the observed process $y_1, \ldots, y_N$ is easily obtained by means of the multinomial distribution. Since each component of $\{y_t\}$ can take either 0 or 1 as possible values, we have that

$$f(y_t \parallel \mathcal{F}_{t-1}) = \prod_{j=1}^{m} p_{tj}(\beta)^{y_{tj}}.$$

Consequently, the corresponding partial likelihood is:

$$\begin{aligned} PL_N(\beta) &= \prod_{t=1}^{N} f(y_t \parallel \mathcal{F}_{t-1}) \\ &= \prod_{t=1}^{N} \prod_{j=1}^{m} p_{tj}(\beta)^{y_{tj}}. \end{aligned}$$

It follows that the partial log-likelihood is given by

$$pl_N(\beta) = \sum_{t=1}^{N} \sum_{j=1}^{m} y_{tj} \log p_{tj}(\beta),$$

where $p_{tj}(\beta) = h_j(\mathbf{Z}'_{t-1}\beta)$ for $j = 1, \ldots, q$. Writing the partial likelihood in the form of a natural exponential family, we get that

$$\begin{aligned} PL_N(\beta) &= \prod_{t=1}^{N} \exp(y_{t1} \log p_{t1}(\beta) + \cdots + y_{tm} \log p_{tm}(\beta)) \\ &= \prod_{t=1}^{N} \exp(l(p_t)' y_t - \log p_{tm}(\beta)), \end{aligned}$$

where l is the logit function from Definition 2.5.1. So the partial log-likelihood can be written as

$$pl_N(\beta) = \sum_{t=1}^{N} (l(p_t)' y_t - \log p_{tm}(\beta)). \tag{2.11}$$

Comparing (2.11) and (2.4) we can identify θ_i as $l(p_t)$ and $b(\theta_i)$ as $1/(1 + \sum_{j=1}^{q} \exp(l_j(p_t)))$ and by letting

$$d = l \circ h = \left[\log(\frac{h_1}{1 - \sum_{i=1}^{q} h_i}), \ldots, \log(\frac{h_q}{1 - \sum_{i=1}^{q} h_i})\right],$$

we get the partial score function

$$ps_N(\beta) = \frac{\partial pl_N(\beta)}{\partial \beta} = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_{t-1}(\beta)(y_t - p_t(\beta)) \tag{2.12}$$

where

$$\mathbf{D}_{t-1}(\beta) = [\partial d(\gamma_{t-1})/\partial \gamma_{t-1}] = \begin{bmatrix} \partial d_1(\gamma_{t-1})/\partial \gamma_{t-1} \\ \vdots \\ \partial d_q(\gamma_{t-1})/\partial \gamma_{t-1} \end{bmatrix}$$

with $\gamma_{t-1} = \mathbf{Z}'_{t-1}\beta$. Equation 2.12 follows from the chain rule

$$\frac{\partial pl_N(\beta)}{\partial \beta} = \sum_{t=1}^{N} \frac{\partial(l(p_t)'y_t - \log p_{tm}(\beta))}{\partial l} \frac{\partial l \circ h}{\partial \gamma_{t-1}} \frac{\partial \gamma_{t-1}}{\partial \beta}$$

Furthermore, the conditional information matrix is given by

$$\begin{aligned} \mathbf{G}_N(\beta) &= \sum_{t=1}^{N} Var[\mathbf{Z}_{t-1}\mathbf{D}_{t-1}(\beta)(y_t - p_t(\beta)) \mid \mathcal{F}_{t-1}] \\ &= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_{t-1}(\beta)\mathbf{\Sigma}_t(\beta)\mathbf{D}'_{t-1}(\beta)\mathbf{Z}'_{t-1} \end{aligned} \tag{2.13}$$

utilizing Lemmas 2.3.3 and 2.3.4. The unconditional information matrix is

$$\mathbf{F}_N(\beta) = E_\beta[\mathbf{G}_N(\beta)] \tag{2.14}$$

and plays a significant role in asymptotic considerations. Finally, the second derivative of the partial log likelihood, multiplied by -1, is

$$\mathbf{H}_N(\beta) = -\frac{\partial^2 pl_N(\beta)}{\partial \beta \partial \beta'} = \mathbf{G}_N(\beta) - \mathbf{R}_N(\beta) \tag{2.15}$$

where

$$\mathbf{R}_N(\beta) = \sum_{t=1}^{N}\sum_{r=1}^{q} \mathbf{Z}_{t-1}\mathbf{W}_{(t-1)r}(\beta)\mathbf{Z}'_{t-1}(y_{tr} - p_{tr}(\beta))$$

with $\mathbf{W}_{(t-1)r}(\beta) = [\partial^2 d_r(\mathbf{Z}'_{t-1}\beta)/\partial\gamma_{t-1}\partial\gamma'_{t-1}]$.

The solution of the non-linear equations $ps_N(\beta) = 0$ is the maximum partial likelihood estimator (MPLE). The discussion about existence and uniqueness at the end of Section 2.4 applies here as well. Asymptotics concerning the behavior of MPLE have been studied under different setting by several authors; see [4], [7], [91], [92].

2.6 Large Sample Theory

Now we prove existence, consistency and asymptotic normality of the maximum partial likelihood estimator (MPLE).

First we state the assumptions that we are going to use for the proof of our main theorems. These are clearly regularity conditions and have been used previously in the literature for establishing good asymptotic properties of the maximum likelihood estimator in problems involving regression.

A.1 The parameter β belongs to an open set $B \subseteq R^p$.

A.2 The covariate matrix $\mathbf{Z}_{t-1}$ almost surely lies in a nonrandom compact subset Γ of $R^{p\times q}$ such that $P[\sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} > \mathbf{0}] = 1$. Furthermore we assume that $\mathbf{Z}'_{t-1}\beta$ lies almost surely in the domain H of h for all $\mathbf{Z}_{t-1} \in \Gamma$ and $\beta \in B$.

A.3 The probability measure P which governs $\{y_t, \mathbf{Z}_{t-1}\}$, $t = 1, \ldots, N$ obeys (2.3) with $\beta = \beta_0$.

A.4 The link function h is twice continuously differentiable and $\det[\partial h(\gamma)/\partial\gamma] \neq 0$.

A.5 There is a probability measure μ on $R^{p\times q}$ such that $\int_{R^{p\times q}} \mathbf{Z}\mathbf{Z}'\mu(d\mathbf{Z})$ is positive definite and such that under (2.3) with $\beta = \beta_0$, for Borel sets $A \subset R^{p\times q}$, we have

$$\frac{1}{N}\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A]} \xrightarrow{p} \mu(A), \text{as } N \to \infty.$$

Assumptions A.1 and A.4 guarantee that the second derivative of the partial log-likelihood is a continuous function of β. The condition $\det[h(\gamma)/\partial\gamma] \neq 0$ implies in particular that $\mathbf{D}_{t-1}$ is not singular, so from A.2 the conditional information matrix is positive definite with probability one. To prove it, note that for any vector $\lambda \in R^p$

$$\begin{aligned}
\lambda'\mathbf{G}_N\lambda &= \lambda'(\sum_{t=1}^{N}\mathbf{Z}_{t-1}\mathbf{D}_{t-1}\mathbf{\Sigma}_t\mathbf{D}'_{t-1}\mathbf{Z}'_{t-1})\lambda \\
&\geq \min_t \lambda_{\min}(\mathbf{D}_{t-1}\mathbf{\Sigma}_t\mathbf{D}'_{t-1})\lambda'(\sum_{t=1}^{N}\mathbf{Z}_{t-1}\mathbf{Z}_{t-1})\lambda' \\
&> 0
\end{aligned}$$

with $\lambda_{\min}(.)$ denoting the minimum eigenvalue. Indeed, since the variance-covariance matrix is positive definite and the matrix of derivatives is not singular we have that the minimum eigenvalue is positive almost everywhere. It follows that the unconditional information matrix is positive definite as well. The last part of assumption A.2 assures that we have a well defined model. The compactness assumption will be useful in deriving bounds for the asymptotics. Assumption A.5 comes from the ergodic theorem of Birkhoff in the setting of independent ergodic stationary processes. It simply states that the empirical

measure of the set $\{\mathbf{Z}_{t-1},\ t = 1, \ldots, N\}$ converges weakly almost surely to a nonrandom measure μ. In conclusion we can say that, if g is any continuous and bounded function on Γ taking values on $R^{p\times q}$, then we have

$$\frac{\sum_{t=1}^{N} g(\mathbf{Z}_{t-1})}{N} \xrightarrow{p} \int_{R^{p\times q}} g(\mathbf{Z})\mu(d\mathbf{Z})$$

by using the definition of weak convergence. Thus the conditional information matrix $\mathbf{G}_N(\beta)$ has a nonrandom limit

$$\frac{\mathbf{G}_N(\beta)}{N} \xrightarrow{p} \int_{R^{p\times q}} \mathbf{Z}\mathbf{D}(\beta)\mathbf{\Sigma}(\beta)\mathbf{D}'(\beta)\mathbf{Z}'\mu(d\mathbf{Z}) = \mathbf{G}(\beta), \tag{2.16}$$

where $\mathbf{D}(\beta) = [\partial h(\mathbf{Z}'\beta)/\partial(\mathbf{Z}'\beta)]$ and $\mathbf{\Sigma}$ has generic element

$$\sigma^{(ij)}(\beta) = \begin{cases} -h_i(\mathbf{Z}'\beta)h_j(\mathbf{Z}'\beta) & \text{if } i \neq j \\ h_i(\mathbf{Z}'\beta)(1 - h_i(\mathbf{Z}'\beta)) & \text{if } i = j \end{cases}$$

for $i, j = 1 \ldots, q$. Note that integration with respect to a matrix means that we integrate with respect to each element of the matrix. From A.4, $\mathbf{G}(\beta)$ is a positive definite matrix at the true value and therefore its inverse exists. We would like to emphasize at this point that no Markovian property was necessary (compare with [47]).

Our proof of consistency and asymptotic normality is based on the classical approach of Cramér. Namely, ee first exhibit a solution of the score equations and then prove that it is consistent and asymptotically normally distributed.

For this end, we will need some helpful lemmas. The following lemmas show that the partial score process is a square integrable martingale which has mean zero and satisfies the conditions for an application of a martingale central limit theorem.

Remark 2.6.1 We use the following convention throughout the chapter. Denote by $\mathbf{B}^{\frac{1}{2}}$ ($\mathbf{B}^{\frac{t}{2}}$) the left (right) square root of a positive definite matrix $\mathbf{B}$, that

is $\mathbf{B} = \mathbf{B}^{\frac{1}{2}}\mathbf{B}^{\frac{t}{2}}$. Furthermore, let $(\mathbf{B}^{-\frac{t}{2}}) = (\mathbf{B}^{\frac{t}{2}})^{-1}$. Square roots of a positive definite matrix are unique up to an orthogonal transformation. Unique square roots of a positive definite matrix are given by the Cholesky decomposition. The right Cholesky square root $\mathbf{B}^{\frac{t}{2}}$ is defined as the unique upper triangular matrix with positive elements such that $\mathbf{B} = (\mathbf{B}^{\frac{t}{2}})'(\mathbf{B}^{\frac{t}{2}})$ where $(\mathbf{B}^{\frac{t}{2}})' = \mathbf{B}^{\frac{1}{2}}$.

Remark 2.6.2 Our proofs will make use of the notion of the norm of a matrix. There are several norms in the vector space of $p \times q$ matrices (see [40]). An example is:

$$\|\mathbf{A}\| = (\sum_{i,j} |a_{ij}|^2)^{1/2}$$

where a_{ij} is the (i,j) element of $\mathbf{A}$. However, since we are dealing with a finite dimensional vector space, all the norms are equivalent.

From now on, we suppress the notation which involves β_0, the true value of β. It is understood, however, that all the calculations are being made under the true probability model.

Lemma 2.6.1 Under A.1-A.4 the partial score function $\{ps_N\}$ evaluated at β_0 is a mean zero square integrable martingale with respect to $\{\mathcal{F}_N\}$.

Proof: Obviously $\mathcal{F}_N \subset \mathcal{F}_{N+1}$, since at time $N+1$ there is more information than at time N. That ps_N is $\mathcal{F}_N$-measurable follows from (2.12). We now show that its expectation is 0.

$$\begin{aligned} E[ps_N] &= E[\sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_{t-1}(y_t - p_t)] \\ &= \sum_{t=1}^{N} E[\mathbf{Z}_{t-1}\mathbf{D}_{t-1}(y_t - p_t)] \\ &= \sum_{t=1}^{N} E[E[\mathbf{Z}_{t-1}\mathbf{D}_{t-1}(y_t - p_t) \parallel \mathcal{F}_{t-1}]] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=1}^{N} E[\mathbf{Z}_{t-1}\mathbf{D}_{t-1}E[(y_t - p_t) \parallel \mathcal{F}_{t-1}]] \\
&= 0.
\end{aligned}
$$

Furthermore, by decomposing the sum,

$$
\begin{aligned}
E[ps_N \parallel \mathcal{F}_{N-1}] &= E[ps_{N-1} + \mathbf{Z}_{N-1}\mathbf{D}_{N-1}(y_N - p_N) \parallel \mathcal{F}_{N-1}] \\
&= ps_{N-1} + E[\mathbf{Z}_{N-1}\mathbf{D}_{N-1}(y_N - p_N) \parallel \mathcal{F}_{N-1}] \\
&= ps_{N-1}
\end{aligned}
$$

utilizing Lemma 2.3.2. Thus the partial score function evaluated at the true parameter is a mean zero martingale. In order to show that it is square integrable it is sufficient to prove that the increments are square integrable. But the increments are

$$a_N = ps_N - ps_{N-1}$$

which turns out to be

$$\mathbf{Z}_{N-1}\mathbf{D}_{N-1}(y_N - p_N)$$

It follows that

$$E[a_N a_N' \parallel \mathcal{F}_{N-1}] = \mathbf{Z}_{N-1}\mathbf{D}_{N-1}\mathbf{\Sigma}_N\mathbf{D}_{N-1}'\mathbf{Z}_{N-1}' \tag{2.17}$$

We first need to observe that $\mathbf{D}_{N-1}$ is a matrix of derivatives of continuous functions. This follows from A.4 and the fact that $d = l \circ h$ is twice continuously differentiable. So the whole expression in (2.17) is a continuous function which is supported in a compact set. We conclude that $\|E[a_N a_N' \parallel \mathcal{F}_{N-1}]\| \leq M$, where M is a constant. Hence square integrability of the increments follows. □

The following two lemmas will give sufficient conditions for the application of a multivariate central limit theorem for martingales. The basic idea is to use

the Cramér-Wold device (see [14]). That is, a vector X_n of random variables converges in distribution to a random variable X if and only if the univariate random variable $\lambda' X_n$ converges in distribution to $\lambda' X$, for every vector λ. This technique is useful in dealing with multivariate convergence theorems. It will be demonstrated below.

Lemma 2.6.2 Under A1-A5 the following is true:

$$\mathbf{F}_N^{-\frac{1}{2}} \mathbf{G}_N \mathbf{F}_N^{-\frac{t}{2}} \xrightarrow{p} \mathbf{I}$$

as $N \to \infty$.

Proof: Let $\lambda \in R^p$ with $\lambda \neq 0$ and assume, without loss of generality, that its Euclidean norm is one; that is, $\| \lambda \| = 1$. We are going to show the equivalent result that

$$\frac{\lambda' \mathbf{G}_N \lambda}{\lambda' \mathbf{F}_N \lambda} \xrightarrow{p} 1.$$

Then the required results follows. Indeed, if the above relations holds for any $\lambda \in R^p$, then there exists a subsequence which has a further subsequence such that

$$\frac{\lambda' \mathbf{G}_{N_{k_l}} \lambda}{\lambda' \mathbf{F}_{N_{k_l}} \lambda} \to 1 \text{ a.s.}$$

But this implies that

$$\mathbf{F}_{N_{k_l}}^{-1} \mathbf{G}_{N_{k_l}} \to \mathbf{I} \text{ a.s.}$$

and therefore the required relation. Consider $\zeta_N = \lambda' ps_N$. Then clearly ζ_N is a martingale with respect to $\{\mathcal{F}_N\}$. Its conditional covariance matrix is $\lambda' \mathbf{G}_N \lambda$ which, suitably normalized, has a limit, due to A.5 and (2.16). That is

$$\frac{\lambda' \mathbf{G}_N \lambda}{N} \xrightarrow{p} \lambda' \mathbf{G} \lambda,$$

where $\mathbf{G} = \int_{R^{p\times q}} \mathbf{ZD\Sigma D'Z'}\mu(d\mathbf{Z})$. On the other hand, its unconditional covariance matrix is $\lambda'\mathbf{F}_N\lambda = \lambda' E[\mathbf{G}_N]\lambda$. We know, however, that if $\{X_n\}$ is a sequence of random variables with $X_n \xrightarrow{p} X$ and $|X_n| \leq c$, where c is a constant, then $E[X_n] \to E[X]$. In our case $\lambda'\mathbf{G}_N\lambda/N \xrightarrow{p} \lambda'\mathbf{G}\lambda$ and $\|\lambda'\mathbf{G}_N\lambda/N\| \leq NM/N = M$ from the proof of Lemma 2.6.1. So

$$\lambda'\mathbf{F}_N\lambda = \lambda' E[\mathbf{G}_N]\lambda \xrightarrow{p} \lambda' E[\mathbf{G}]\lambda = \lambda'\mathbf{G}\lambda.$$

It follows that

$$\frac{\lambda'\mathbf{F}_N\lambda}{\lambda'\mathbf{G}_N\lambda} = \frac{\lambda'\mathbf{F}_N\lambda/N}{\lambda'\mathbf{G}_N\lambda/N} \xrightarrow{p} 1$$

from Slutsky's theorem, which was to be proved. □

We now demonstrate that ps_N satisfies the Lindeberg's condition.

Lemma 2.6.3 Let $\zeta_N = \lambda' ps_N$, as in the proof of Lemma 2.6.2. Then under A1-A5 Lindeberg's condition holds for ζ_N. That is, for every $\epsilon > 0$, we have

$$\frac{1}{\lambda'\mathbf{F}_N\lambda}\sum_{t=1}^{N} E[|\lambda' a_t|^2 I_{Nt}(\epsilon) \parallel \mathcal{F}_{t-1}] \xrightarrow{p} 0$$

as $N \to \infty$, where $I_{Nt}(\epsilon)$ is the indicator of the set $\{|\lambda' a_t| \geq (\lambda'\mathbf{F}_N\lambda)^{\frac{1}{2}}\epsilon\}$ and $a_t = ps_t - ps_{t-1}$.

Proof: We have that

$$\frac{1}{\lambda'\mathbf{F}_N\lambda}\sum_{t=1}^{N} E[|\lambda' a_t|^2 I_{Nt}(\epsilon) \parallel \mathcal{F}_{t-1}] = \frac{1}{\lambda'\mathbf{F}_N\lambda}\sum_{t=1}^{N}\int_{|\lambda' a_t|\geq(\lambda'\mathbf{F}_N\lambda)^{\frac{1}{2}}\epsilon} |\lambda' a_t|^2 dQ_{t,N}$$

where $Q_{t,N}$ is the conditional probability measure given $\mathcal{F}_{t-1}$. But $|\lambda' a_t| \geq (\lambda'\mathbf{F}_N\lambda)^{\frac{1}{2}}\epsilon$ implies that $|\lambda' a_t|/(\lambda'\mathbf{F}_N\lambda)^{\frac{1}{2}}\epsilon \geq 1$ and therefore we have:

$$\begin{aligned}\frac{1}{\lambda'\mathbf{F}_N\lambda}\sum_{t=1}^{N}\int_{|\lambda' a_t|\geq(\lambda'\mathbf{F}_N\lambda)^{\frac{1}{2}}\epsilon} |\lambda' a_t|^2 dQ_{t,N} &\leq \frac{1}{(\lambda'\mathbf{F}_N\lambda)^{\frac{3}{2}}\epsilon}\sum_{t=1}^{N} E[|\lambda' a_t|^3 \parallel \mathcal{F}_{t-1}] \\ &\leq \frac{NM_1}{(\lambda'\mathbf{F}_N\lambda)^{\frac{3}{2}}\epsilon},\end{aligned}$$

where M_1 is a bound. Such a bound exists from A.2. It follows that

$$\frac{N}{(\lambda'\mathbf{F}_N\lambda)^{\frac{3}{2}}}\frac{1}{\epsilon} = \frac{1}{\sqrt{N}(\frac{\lambda'\mathbf{F}_N\lambda}{N})^{\frac{3}{2}}}\frac{1}{\epsilon} \to 0.$$

Thus, Lindeberg's condition is established for ζ_N and consequently for ps_N. □

The next lemma parallels a well known result from linear models ([57])

Lemma 2.6.4 Under A1-A5 we have that

$$\lambda_{\min}(\mathbf{F}_N) \to \infty$$

as $N \to \infty$, where $\lambda_{\min}(\mathbf{F}_N)$ is the minimum eigenvalue of the unconditional information matrix.

Proof: We have that if $\mathbf{A}$ and $\mathbf{B}$ are both positive definite matrices,

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq c\|\mathbf{A} - \mathbf{B}\|, \tag{2.18}$$

where the positive constant c depends only on the norm of the matrix. Then, by the proof of Lemma 2.6.2 we get

$$|\lambda_{\min}(\frac{\mathbf{F}_N}{N}) - \lambda_{\min}(\mathbf{G})| \to 0.$$

It follows that $\lambda_{\min}(\mathbf{F}_N) = \mathrm{O}(N)$ and the claim is true. □

We now prove a continuity condition. Namely, we would like to have the matrix of second derivatives as close as possible to the information matrix. This is a technical lemma and the proof is along the lines of [47].

Lemma 2.6.5 Under A.1-A.5 the following continuity condition holds

$$\sup_{\tilde{\beta}\in O_N(\delta)} \|\mathbf{F}_N^{-\frac{1}{2}}(\mathbf{H}_N(\tilde{\beta}) - \mathbf{G}_N)\mathbf{F}_N^{-\frac{t}{2}}\| \xrightarrow{p} 0$$

with $O_N(\delta) = \{\tilde{\beta} : \|\mathbf{F}_N^{\frac{t}{2}}(\tilde{\beta} - \beta_0)\| \leq \delta\}$, holds for any $\delta > 0$.

Proof: Let $\lambda \in R^p$ with $\lambda \neq 0$, and assume without loss of generality that $\|\lambda\| = 1$. We will show the equivalent condition (see the discussion at the beginning of the proof of lemma 2.6.2) that, for any $\delta > 0$,

$$\sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-\frac{1}{2}} (\mathbf{H}_N(\tilde{\beta}) - \mathbf{G}_N) \mathbf{F}_N^{-\frac{t}{2}} \lambda \xrightarrow{p} 0, \tag{2.19}$$

using the Cramér-Wold device. By decomposing $\mathbf{H}_N(\tilde{\beta}) = \mathbf{G}_N(\tilde{\beta}) - \mathbf{R}_N(\tilde{\beta})$, we only need to show that

$$g_N = \sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-\frac{1}{2}} (\mathbf{G}_N(\tilde{\beta}) - \mathbf{G}_N) \mathbf{F}_N^{-\frac{t}{2}} \lambda \xrightarrow{p} 0 \tag{2.20}$$

and

$$\sup_{\tilde{\beta} \in O_N(\delta)} \lambda' \mathbf{F}_N^{-\frac{1}{2}} \mathbf{R}_N(\tilde{\beta}) \mathbf{F}_N^{-\frac{t}{2}} \lambda \xrightarrow{p} 0 \tag{2.21}$$

hold simultaneously. Define the vectors $w'_{(t-1)N} = \lambda' \mathbf{F}_N^{-\frac{1}{2}} \mathbf{Z}_{t-1}$, for $1 \leq t \leq N$, and $w_N = \sum_{t=1}^N w'_{(t-1)N} w_{(t-1)N}$. Then we have that

$$g_N = \sup_{\tilde{\beta} \in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N} (\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}) w_{(t-1)N},$$

where $\mathbf{L}_{t-1}(\beta) = \mathbf{D}_{t-1}(\beta) \mathbf{\Sigma}_t(\beta) \mathbf{D}'_{t-1}(\beta)$ for $t = 1, \ldots, N$. It follows that

$$g_N \leq w_N \sup_{\tilde{\beta} \in O_N(\delta), t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\|.$$

Using A.2, $\sup_{1 \leq t \leq N} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\|$ can be estimated from above by a continuous function of $\tilde{\beta}$ with a zero at $\tilde{\beta} = \beta$. Notice that $\{O_N(\delta)\}$ shrinks to β. Hence

$$\sup_{\tilde{\beta} \in O_N(\delta), t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \to 0$$

By applying Markov's inequality we have that

$$P[|g_N| \geq \epsilon] \leq \frac{E[g_N]}{\epsilon}$$

$$
\begin{aligned}
&\leq E[w_N] \sup_{\tilde{\beta}\in O_N(\delta),t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \\
&= \lambda'[\frac{\mathbf{F}_N}{N}]^{-\frac{1}{2}} \frac{\sum_{t=1}^N E[\mathbf{Z}_{t-1}\mathbf{Z}'_{t-1}]}{N} [\frac{\mathbf{F}_N}{N}]^{-\frac{t}{2}} \lambda \\
&\quad \times \sup_{\tilde{\beta}\in O_N(\delta),t} \|\mathbf{L}_{t-1}(\tilde{\beta}) - \mathbf{L}_{t-1}\| \to 0
\end{aligned}
$$

since the other terms converge to a limit by the continuity of the square root and the assumption A.5 (compare with the proof of Lemma 2.6.2). By further decomposition we obtain that

$$
\sup_{\tilde{\beta}\in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N}(\mathbf{W}_{(t-1)j}(\tilde{\beta}) - \mathbf{W}_{(t-1)j}) w_{(t-1)N}(y_{tj} - p_{tj}) \xrightarrow{p} 0, \tag{2.22}
$$

$$
\sup_{\tilde{\beta}\in O_N(\delta)} \sum_{t=1}^N w'_{(t-1)N}\mathbf{W}_{(t-1)j}(\tilde{\beta}) w_{(t-1)N}(p_{tj} - p_{tj}(\tilde{\beta})) \xrightarrow{p} 0, \tag{2.23}
$$

and

$$
\sum_{t=1}^N w'_{(t-1)N}\mathbf{W}_{(t-1)j} w_{(t-1)N}(y_{tj} - p_{tj}) \xrightarrow{p} 0 \tag{2.24}
$$

for any j, $1 \leq j \leq q$, are sufficient for (2.21). The proofs of (2.22), (2.23), are similar to the proof of (2.20). We prove (2.24). Consider the increments of (2.24), that is,

$$
u_{(t-1)N} = w'_{(t-1)N}\mathbf{W}_{(t-1)j} w_{(t-1)N}(y_{tj} - p_{tj})
$$

Then we see that

$$
E[u_{(t-1)N} \parallel \mathcal{F}_{t-1}] = 0
$$

and

$$
\begin{aligned}
Var[u_{(t-1)N} \parallel \mathcal{F}_{t-1}] &= w'_{(t-1)N}\mathbf{W}_{(t-1)j} w_{(t-1)N} \\
&\quad Var[y_{tj} - p_{tj} \parallel \mathcal{F}_{t-1}] w'_{(t-1)N}\mathbf{W}'_{(t-1)j} w_{(t-1)N} \\
&\leq K(w'_{(t-1)N} w_{(t-1)N})^2,
\end{aligned}
$$

where K is a bound on $\|\mathbf{W}_{(t-1)j}\|^2 Var[y_{tj} - p_{tj} \parallel \mathcal{F}_{t-1}]$. From A.2 and A.4 and the boundedness of y_{tj}, such a bound exists. This follows from the relation $x'\mathbf{A}x \leq x'x\|\mathbf{A}\|$. Actually, the above two relations make clear that $\{u_{(t-1)N},\ t = 1, \ldots, N\}$ are the orthogonal increments of a square integrable mean zero martingale. It follows that

$$E(\sum_{t=1}^{N} u_{(t-1)N}) = 0$$

and

$$\begin{aligned} Var[\sum_{t=1}^{N} u_{(t-1)N}] &\leq K \sum_{t=1}^{N} E[(w'_{(t-1)N} w_{(t-1)N})^2] \\ &\leq K \sup_t w'_{(t-1)N} w_{(t-1)N} E[w_N]. \end{aligned}$$

However

$$\begin{aligned} \sup_t w'_{(t-1)N} w_{(t-1)N} &= \sup_t \lambda' \mathbf{F}_N^{-\frac{1}{2}} \mathbf{Z}_{t-1} \mathbf{Z}'_{t-1} \mathbf{F}_N^{-\frac{t}{2}} \lambda \\ &\leq \lambda' \mathbf{F}_N^{-1} \lambda \sup_{\mathbf{Z}_{t-1} \in \Gamma} \|\mathbf{Z}_{t-1}\|^2 \\ &\leq \frac{\sup_{\mathbf{Z}_{t-1} \in \Gamma} \|\mathbf{Z}_{t-1}\|^2}{\lambda_{\min}(\mathbf{F}_N)} \to 0 \end{aligned}$$

due to Lemma 2.6.4. Since $E[w_N]$ is bounded, from its convergence, relation (2.24) holds and therefore the continuity condition is established. □

Now we aprove the main result of this chapter. We prove existence and consistency of the sequence $\hat{\beta}_N$ together, and then we use the asymptotic normality of the partial score to establish the asymptotic distribution (see also [9]).

Theorem 2.6.1 Under A.1-A.5, the probability that a locally unique maximum partial likelihood estimator exists converges to one. Moreover there exists a sequence of maximum partial likelihood estimators $\hat{\beta}_N$ which is consistent and

asymptotically normal. That is,

$$\sqrt{N}(\hat{\beta}_N - \beta_0) \xrightarrow{D} \mathcal{N}(0, \mathbf{G}^{-1}(\beta_0)).$$

Proof: From Lemmas 2.6.2 and 2.6.3 (see [43]) we get that

$$(\mathbf{F}_N^{-\frac{1}{2}} ps_N,\ \mathbf{F}_N^{-\frac{1}{2}} \mathbf{G}_N \mathbf{F}_N^{-\frac{1}{2}}) \xrightarrow{D} (\mathcal{N}, \mathbf{I}),$$

where $\mathcal{N}$ is a standard normal vector . Choosing $\mathbf{G}_N^{\frac{1}{2}}$ such that $\mathbf{F}_N^{-\frac{1}{2}} \mathbf{G}_N^{\frac{1}{2}}$ is the Cholesky square root of $\mathbf{F}_N^{-\frac{1}{2}} \mathbf{G}_N \mathbf{F}_N^{-\frac{1}{2}}$, we have from the continuity of the square root that

$$(\mathbf{F}_N^{-\frac{1}{2}} ps_N,\ \mathbf{F}_N^{-\frac{1}{2}} \mathbf{G}_N^{\frac{1}{2}}) \xrightarrow{D} (\mathcal{N}, \mathbf{I})$$

where $\mathcal{N}$ is as above.

We first prove asymptotic existence and consistency. By Taylor expansion we have that

$$pl_N(\tilde{\beta}) = pl_N(\beta_0) + (\tilde{\beta} - \beta_0)' ps_N - \frac{1}{2}(\tilde{\beta} - \beta_0)' \mathbf{H}_N(\tilde{\tilde{\beta}})(\tilde{\beta} - \beta_0),$$

where $\tilde{\tilde{\beta}}$ lies between $\tilde{\beta}$ and β_0. Equivalently

$$pl_N(\tilde{\beta}) - pl_N(\beta_0) = (\tilde{\beta} - \beta_0)' ps_N - \frac{1}{2}(\tilde{\beta} - \beta_0)' \mathbf{H}_N(\tilde{\tilde{\beta}})(\tilde{\beta} - \beta_0). \tag{2.25}$$

Now let $\tilde{\lambda} = \mathbf{F}_N^{\frac{t}{2}}(\tilde{\beta} - \beta_0)/\delta$. Then it follows that $(\tilde{\beta} - \beta_0)' = \tilde{\lambda}' \mathbf{F}_N^{-\frac{1}{2}} \delta$. Substituting into (2.25), we have

$$pl_N(\tilde{\beta}) - pl_N(\beta_0) = \delta \tilde{\lambda}' \mathbf{F}_N^{-\frac{1}{2}} ps_N - \frac{\delta^2}{2} \tilde{\lambda}' \mathbf{F}_N^{-\frac{1}{2}} \mathbf{H}_N(\tilde{\tilde{\beta}}) \mathbf{F}_N^{-\frac{t}{2}} \tilde{\lambda}. \tag{2.26}$$

We are going to prove that for every $\eta > 0$ there exists N and δ such that

$$P[pl_N(\tilde{\beta}) - pl_N(\beta_0) < 0 \ \ \forall \tilde{\beta} \in \partial O_N(\delta)] \geq 1 - \eta. \tag{2.27}$$

This shows that, with probability tending to one, there exists a local maximum inside $O_N(\delta)$. This follows from the definition of the local maximum. If we are able to show that the maximum is not attained at the boundary, then necessarily it is attained in the interior from the almost sure continuity of the partial likelihood. Demonstration of (2.27) is based on the proof of Theorem 1 in [31]. It provides a weaker condition of asymptotic existence. Note, however, that we do not assume any conditions for the third derivatives of the partial log-likelihood. From (2.26), we recognize that it is sufficient to show

$$P\left[\|\mathbf{F}_N^{-\frac{1}{2}} ps_N\|^2 \leq \delta^2 \frac{\lambda_{\min}^2(\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_\mathbf{N}(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})}{4}\right] \geq 1-\eta. \tag{2.28}$$

This is so because of the inequality

$$\tilde{\lambda}\mathbf{F}_N^{-\frac{1}{2}} ps_N - \frac{\delta}{2}\tilde{\lambda}\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}}\tilde{\lambda} \leq \|\mathbf{F}_N^{-\frac{1}{2}} ps_N\| - \frac{\delta}{2}\lambda_{\min}(\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}}).$$

The last inequality follows from Cauchy-Schwartz and the fact that $\tilde{\lambda}'\tilde{\lambda} = 1$. Consequently we have that

$$P\left[\|\mathbf{F}_N^{-\frac{1}{2}} ps_N\|^2 \leq \delta^2 \frac{\lambda_{\min}^2(\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})}{4}\right] \geq 1 - \frac{4E[\|\mathbf{F}_N^{-\frac{1}{2}} ps_N\|^2]}{\delta^2\lambda_{\min}^2(\mathbf{F}_N^{-\frac{1}{2}}\mathbf{H}_N(\tilde{\tilde{\beta}})\mathbf{F}_N^{-\frac{t}{2}})}$$

Since $E[\|\mathbf{F}_N^{-\frac{1}{2}} ps_N\|^2]$ is bounded and the denominator is bounded as well, the above expression can become arbitrarily small. The last claim follows from (2.18) and Lemma 2.6.5. Asymptotic existence therefore is established. More specifically, we have that there exists a sequence $\{\hat{\beta}_N\}$ of MPLE's such that for any $\eta > 0$, there is a δ, N_1 with

$$P[\hat{\beta}_N \in O_N(\delta)] \geq 1-\eta \ \ \forall N \geq N_1. \tag{2.29}$$

From Lemmas 2.6.4 and 2.6.5, we obtain that $\mathbf{H}_N(\beta)$ is positive definite throughout $O_N(\delta)$ with probability converging to 1. Therefore the MPLE $\hat{\beta}_N$ is

also locally unique. Consistency was established as well, upon noting

$$\begin{aligned} 1-\eta &\leq P[\|\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N-\beta_0)\| \leq \delta] \\ &\leq P[\|\hat{\beta}_N-\beta_0\| \leq \frac{\delta}{\sqrt{\lambda_{\min}(\mathbf{F}_N)}}], \end{aligned}$$

and using lemma 2.6.4. One can determine such a sequence of estimators which does not depend on δ (see [58, Theorem 6.2.2]).

We prove now asymptotic normality. By Taylor expansion around $\hat{\beta}_N$, and using the mean value theorem for multivariate functions we obtain

$$ps_N = \tilde{\mathbf{H}}_N(\hat{\beta}_N-\beta_0) \tag{2.30}$$

where $\tilde{\mathbf{H}}_N = \int_0^1 \mathbf{H}_N(\beta_0+s(\hat{\beta}_N-\beta_0)ds$ and the integration is taken elementwise. We need to show that

$$\mathbf{F}_N^{-\frac{1}{2}}\tilde{\mathbf{H}}_N\mathbf{F}_N^{-\frac{t}{2}} \xrightarrow{p} \mathbf{I}. \tag{2.31}$$

But

$$\begin{aligned} \mathbf{F}_N^{-\frac{1}{2}}\tilde{\mathbf{H}}_N\mathbf{F}_N^{-\frac{t}{2}} &= \mathbf{F}_N^{-\frac{1}{2}}(\tilde{\mathbf{H}}_N-\mathbf{G}_N)\mathbf{F}_N^{-\frac{t}{2}} + \mathbf{F}_N^{-\frac{1}{2}}\mathbf{G}_N\mathbf{F}_N^{-\frac{t}{2}} \\ &\xrightarrow{p} \mathbf{0}+\mathbf{I}=\mathbf{I}. \end{aligned}$$

The quantity $\mathbf{F}_N^{-\frac{1}{2}}(\tilde{\mathbf{H}}_N-\mathbf{G}_N)\mathbf{F}_N^{-\frac{t}{2}}$ converges to $\mathbf{0}$ in probability from Lemma 2.6.5 and from the first part of the proof (see equation 2.29). The rest follows from Lemma 2.6.2. Therefore, from (2.30)

$$\mathbf{F}_N^{-\frac{1}{2}}ps_N = (\mathbf{F}_N^{-\frac{1}{2}}\tilde{\mathbf{H}}_N\mathbf{F}_N^{-\frac{t}{2}})(\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N-\beta_0)).$$

Thus

$$\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N-\beta_0) \xrightarrow{D} \mathcal{N}.$$

But

$$\mathbf{G}_N^{\frac{t}{2}}(\hat{\beta}_N-\beta_0) = \mathbf{G}_N^{\frac{t}{2}}\mathbf{F}_N^{-\frac{t}{2}}\mathbf{F}_N^{\frac{t}{2}}(\hat{\beta}_N-\beta_0) \xrightarrow{D} \mathcal{N}$$

since $\mathbf{G}_N^{\frac{t}{2}}\mathbf{F}_N^{-\frac{t}{2}} \xrightarrow{p} \mathbf{I}$. From the continuity of the square root

$$\frac{\mathbf{G}_N^{\frac{t}{2}}}{\sqrt{N}} \xrightarrow{p} \mathbf{G}^{\frac{t}{2}}.$$

An application of Slutsky's theorem yields the conclusion of the theorem. □

Corollary 2.6.1 Under A.1-A.5 we have

$$\sqrt{N}(\hat{\beta}_N - \beta_0) - \frac{1}{N}\mathbf{G}^{-1}ps_N \xrightarrow{p} 0.$$

Proof: Again using Slutsky's theorem and the continuity of the square root, we obtain that

$$\frac{1}{N}\mathbf{G}^{-1}ps_N \xrightarrow{D} \mathcal{N}(0, \mathbf{G}^{-1}).$$

The claim follows. □

Now, assume that each component of the link function is log-concave. That is $\log h_j$ is concave for every $j = 1, \ldots, m$ with $h_m = 1 - \sum_{j=1}^{q} h_j$. It follows that the logarithm is concave, and if the parameter space B is R^p, we obtain the following:

Corollary 2.6.2 Suppose A.1-A.5 holds. Assume further that $\log h_j$ is concave for $j = 1, \ldots, m$. Then the probability that a unique maximum partial likelihood estimator exists converges to one. This sequence is consistent and asymptotically normal as in Theorem 2.6.1.

Remark 2.6.3 If we have k (with k not depending on N) independent processes $\{y_t^i, \mathbf{Z}_t^i\}$ and we model each one as before, the claim of Theorem 2.6.1 is still true provided that $Nk \to \infty$. (see the proof of Theorem 3.1 in [92].)

2.7 Goodness of Fit Statistics

A question which arises naturally after every procedure that involves regression is that of goodness of fit. A clear way to examine goodness of fit is to use the residuals, defined as , the observed minus the predicted values. Our approach will be the following. We classify the responses y_t according to mutually exclusive events in terms of the covariates $\mathbf{Z}_{t-1}$, and then for each category we examine the deviation of the number of positive responses from its conditional expected value (see [87], [92]). Before we proceed to the development of our test statistic, we need to give a couple of definitions, which will be found useful in what follows. These definitions can be found in [40].

Definition 2.7.1 Let $\mathbf{B}_1$,$\mathbf{B}_2$ be matrices of any order. The direct sum of $\mathbf{B}_1$ and $\mathbf{B}_2$, denoted by $\mathbf{B}_1 \oplus \mathbf{B}_2$, is defined as the square matrix

$$\mathbf{B}_1 \oplus \mathbf{B}_2 = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{bmatrix}.$$

The operation of direct sum extends to any finite number of matrices. We have that

$$\oplus_{l=1}^{k} \mathbf{B}_l = \mathbf{B}_1 \oplus \mathbf{B}_2 \oplus \ldots \oplus \mathbf{B}_k = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_k \end{bmatrix}.$$

Definition 2.7.2 Let $\mathbf{A}$ be an $r \times c$ matrix with elements a_{ij} and $\mathbf{B}$ be an $s \times d$ matrix. The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \otimes \mathbf{B}$, is defined as

the following $rs \times cd$ matrix:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1c}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2c}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{r1}\mathbf{B} & a_{r2}\mathbf{B} & \cdots & a_{rc}\mathbf{B} \end{bmatrix}.$$

Now we proceed to develop diagnostic tools for the model at hand.

Suppose that $A_1, \ldots, A_k$ constitute a partition of $R^{p \times q}$. For $l = 1, \ldots, k$ define

$$M_N^l = \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]} y_t$$

and

$$E_N^l(\beta) = \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]} p_t(\beta),$$

where $I_{[\mathbf{Z}_{t-1} \in A_l]}$ is the indicator of the set $\{\mathbf{Z}_{t-1} \in A_l\}$, for $l = 1, \ldots, k$. In addition set the $kq \times 1$ vectors $M_N = (M_N^1, \ldots, M_N^k)$, $E_N(\beta) = (E_N^1(\beta), \ldots, E_N^k(\beta))$. If we let the $k \times 1$ vector $I_{t-1} = (I_{[\mathbf{Z}_{t-1} \in A_1]}, \ldots, I_{[\mathbf{Z}_{t-1} \in A_k]})$, then we can see that the $kq \times 1$ vector

$$d_N(\beta) = M_N - E_N(\beta) = \sum_{t=1}^{N} I_{t-1} \otimes (y_t - p_t(\beta))$$

just by applying Definition 2.7.2.

We are going to study the asymptotic properties of the $kq \times 1$ vector d_N evaluated at the true value. Note that once again we drop the dependence of the various quantities on the true value. Under our assumptions it will be shown that d_N is a mean zero square integrable martingale. Then, we will verify necessary conditions for the application of a central limit theorem, working as before.

Lemma 2.7.1 Under A1-A.3 $\{d_N,\ \mathcal{F}_N\}$ is a zero mean square integrable martingale.

Proof: We have that the sequence of σ-fields is nested and that d_N is $\mathcal{F}_N$-measurable, just by its definition. Now we see that

$$\begin{aligned} E[d_N] &= E[\sum_{t=1}^{N} I_{t-1} \otimes (y_t - p_t)] \\ &= \sum_{t=1}^{N} E[I_{t-1} \otimes (y_t - p_t)] \\ &= \sum_{t=1}^{N} E[E[I_{t-1} \otimes (y_t - p_t) \parallel \mathcal{F}_{t-1}]] \\ &= \sum_{t=1}^{N} E[I_{t-1} \otimes E[(y_t - p_t) \parallel \mathcal{F}_{t-1}]] \\ &= 0. \end{aligned}$$

Moreover

$$\begin{aligned} E[d_N \parallel \mathcal{F}_{N-1}] &= E[d_{N-1} + I_{N-1} \otimes (y_N - p_N) \parallel \mathcal{F}_{N-1}] \\ &= d_{N-1} + I_{N-1} \otimes E[y_N - p_N \parallel \mathcal{F}_{N-1}] \\ &= d_{N-1}. \end{aligned}$$

So, we conclude that d_N is a zero mean martingale with respect to $\mathcal{F}_N$. To prove square integrability we consider the increments of d_N

$$d_N - d_{N-1} = I_{N-1} \otimes (y_N - p_N)$$

We have that

$$E[[I_{N-1} \otimes (y_N - p_N)][(I_{N-1} \otimes (y_N - p_N)]' \parallel \mathcal{F}_{N-1}] = \oplus_{l=1}^{k} \mathbf{C}_{Nl},$$

where

$$\mathbf{C}_{Nl} = I_{[\mathbf{Z}_{t-1}\in A_l]} \begin{bmatrix} p_{N1}(1-p_{N1}) & -p_{N1}p_{N2} & \cdots & -p_{N1}p_{Nq} \\ -p_{N1}p_{N2} & p_{N2}(1-p_{N2}) & \cdots & -p_{N2}p_{Nq} \\ \vdots & \vdots & \vdots & \vdots \\ -p_{Nq}p_{N1} & p_{Nq}p_{N2} & \cdots & p_{Nq}(1-p_{Nq}) \end{bmatrix}.$$

The above equation follows from properties of the multinomial distribution and the fact that $I_{[\mathbf{Z}_{t-1}\in A_l]}$ and $I_{[\mathbf{Z}_{t-1}\in A_r]}$ have zero product for $l \neq r$ for $l, r = 1, \ldots, k$. The boundedness follows by the same reasoning as in Lemma 2.6.1. □

Now we prove the following.

Proposition 2.7.1 Under A.1-A.5 we have that

$$\frac{d_N}{\sqrt{N}} \xrightarrow{p} \mathcal{N}(0, \mathbf{C})$$

where $\mathbf{C} = \oplus_{l=1}^{k} \mathbf{C}_l$ and $\mathbf{C}_l$ is a $q \times q$ symmetric matrix which is given by

$$\mathbf{C}_l = \begin{bmatrix} \int_{A_l} p_1(\beta_0)(1-p_1(\beta_0))\mu(d\mathbf{Z}) & \cdots & -\int_{A_l} p_1(\beta_0)p_q(\beta_0)\mu(d\mathbf{Z}) \\ \vdots & \ddots & \vdots \\ -\int_{A_l} p_1(\beta_0)p_q(\beta_0)\mu(d\mathbf{Z}) & \cdots & \int_{A_l} p_q(\beta_0)(1-p_q(\beta_0))\mu(d\mathbf{Z}) \end{bmatrix}.$$

Proof: Once again we are going to use the Cramér-Wold device. Let $\lambda \in R^{kq}$,$\lambda \neq 0$ and assume again, without loss of generality that $|\lambda| = 1$. Consider $\phi_N = \lambda' d_N$. Clearly ϕ_N is a mean zero square integrable martingale. The conditional covariance matrix of ϕ_N is given by

$$\mathbf{\Lambda}_N = \sum_{t=1}^{N} \oplus_{l=1}^{k} \mathbf{C}_{tl}.$$

Thus the unconditional information matrix is

$$\tilde{\mathbf{\Lambda}}_N = E[\mathbf{\Lambda}_N].$$

Hence, the conditional covariance matrix of ϕ_N is $\lambda'\mathbf{\Lambda}_N\lambda$ and its unconditional covariance matrix is $\lambda'\tilde{\mathbf{\Lambda}}_N\lambda$. Now, following the steps of the proof of Lemma 2.6.2, we have that

$$\frac{\lambda'\mathbf{\Lambda}_N\lambda}{N} \xrightarrow{p} \lambda'\mathbf{C}\lambda$$

and

$$\frac{\lambda'\tilde{\mathbf{\Lambda}}_N\lambda}{N} \xrightarrow{p} \lambda'\mathbf{C}\lambda.$$

It follows from Slutsky's theorem that

$$\frac{\lambda'\mathbf{\Lambda}_N\lambda}{\lambda'\tilde{\mathbf{\Lambda}}_N\lambda} \xrightarrow{p} 1.$$

So the first condition for the application of a multivariate central limit theorem for martingales has been established. To prove Lindeberg's condition for ϕ_N, that is,

$$\frac{1}{\lambda'\tilde{\mathbf{\Lambda}}_N\lambda}\sum_{t=1}^{N} E[|\phi_t|^2 I_{Nt}(\epsilon) \parallel \mathcal{F}_{t-1}] \xrightarrow{p} 0,$$

where $I_{Nt}(\epsilon)$ is the indicator of the set $\{|\phi_t| \geq (\lambda'\tilde{\mathbf{\Lambda}}_N\lambda)^{\frac{1}{2}}\epsilon\}$, we will need to work as in the proof of Lemma 2.6.3. By the same argument we have that

$$\begin{aligned}\frac{1}{\lambda'\tilde{\mathbf{\Lambda}}_N\lambda}\sum_{t=1}^{N} E[|\phi_t|^2 I_{Nt}(\epsilon) \parallel \mathcal{F}_{t-1}] &\leq (\frac{1}{\lambda'\tilde{\mathbf{\Lambda}}_N\lambda})^{\frac{3}{2}}\frac{1}{\epsilon}\sum_{t=1}^{N} E[|\phi_t|^3 \parallel \mathcal{F}_{t-1}] \\ &\leq \frac{NM_2}{(\lambda'\tilde{\mathbf{\Lambda}}_N\lambda)^{\frac{3}{2}}}\frac{1}{\epsilon}.\end{aligned}$$

where M_2 is a bound which exists from A.2. Lindeberg's condition therefore follows and hence the proposition. □

Since we have already established asymptotic normality of the residuals the following proposition is a natural consequence.

Proposition 2.7.2 Suppose that A.1-A.5 hold. As $N \to \infty$, the asymptotic distribution of the statistic

$$\chi^2(\beta_0) = \frac{1}{N}\sum_{l=1}^{k} d_l'\mathbf{C}_l^{-1}d_l$$

is chi-squared with kq degrees of freedom.

Proof Since $d_l'\mathbf{C}_l^{-1}d_l/N$ is distributed as chi-square with q degrees of freedom, from Proposition 2.7.1 we have that $\chi^2(\beta_0)$ follows a chi-square with kq degrees of freedom because $\chi^2(\beta_0)$ is a sum of independent chi-square distributed random variables. The inverse of $\mathbf{C}_l$, $l = 1, \ldots, k$ is guaranteed to exist by assumption A.5. □

We are going to demonstrate now another theorem which gives rise to another goodness of fit statistic.

Theorem 2.7.1 Suppose that A.1-A.5 hold. Let $A_1, \ldots, A_k$ be a partition of $R^{p\times q}$. Then we have, as $N \to \infty$,

1.
$$\sqrt{N}(\frac{d_N'}{N}, (\hat{\beta}_N - \beta_0)) \xrightarrow{D} \mathcal{N}(0, \mathbf{\Gamma}),$$

where $\mathbf{\Gamma}$ is a square matrix of dimension $p + kq$

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{C} & \mathbf{B}' \\ \mathbf{B} & \mathbf{G}^{-1} \end{bmatrix}.$$

Here $\mathbf{C}$ is as in proposition 2.7.1, $\mathbf{G}$ is the limiting $p\times p$ information matrix, and the lth column of $\mathbf{B}$ is given by the matrix

$$\mathbf{G}^{-1}\int_{A_l} \mathbf{ZD\Sigma}\mu d(\mathbf{Z}).$$

2. We also have, as $N \to \infty$, that

$$\frac{E_N(\hat{\beta}_N) - E_N(\beta_0)}{\sqrt{N}} - \sqrt{N}\mathbf{B}'\mathbf{G}(\hat{\beta}_N - \beta_0) \xrightarrow{p} 0.$$

Proof: To prove (1) we only need to observe from Corollary 2.6.1 that, for some integer N greater than N_0, we have that

$$\frac{1}{\sqrt{N}}(d'_N, (\hat{\beta} - \beta_0)) \overset{p}{\approx} \frac{1}{\sqrt{N}}(d'_N, \mathbf{G}^{-1} ps_N). \tag{2.32}$$

Now we know that d_N and ps_N are martingales which obey the conditions for an application of a central limit theorem for martingales. It follows that jointly (using again the Cramér-Wold device) the vector on the right hand side of the above equation converges to normal as $N \to \infty$. We only need to compute the asymptotic covariance matrix of its components. We have

$$\begin{aligned}
&\frac{1}{N}\mathbf{G}^{-1} ps_N \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t) \\
&= \frac{1}{N}\mathbf{G}^{-1} \sum_{s=1}^{N} \mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s) \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t)
\end{aligned}$$

But for $s < t$

$$\begin{aligned}
&E[\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s) I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t)] \\
&= E[\mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s) I_{[\mathbf{Z}_{t-1} \in A_l]} E[(y_t - p_t) \parallel \mathcal{F}_{t-1}]] \\
&= 0
\end{aligned}$$

Therefore, we have from Assumption (A.5) that

$$\begin{aligned}
&E[\frac{1}{N}\mathbf{G}^{-1} ps_N \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t)] \\
&= E[\frac{1}{N}\mathbf{G}^{-1} \sum_{s=1}^{N} \mathbf{Z}_{s-1}\mathbf{D}_{s-1}(y_s - p_s) \sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]}(y_t - p_t)] \\
&= E[\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1} \in A_l]} \mathbf{Z}_{t-1}\mathbf{D}_{t-1}\mathbf{\Sigma}_t] \xrightarrow{p} \mathbf{G}^{-1} \int_{A_l} \mathbf{Z}\mathbf{D}\mathbf{\Sigma}\mu d(\mathbf{Z})
\end{aligned}$$

The first part of the theorem follows. To prove the second part, we have by Taylor's expansion for $l = 1, \ldots, k$,

$$E_N^l(\hat{\beta}_N) \approx E_N^l(\beta_0) + [\frac{\partial E_N^l(\beta)}{\partial \beta}]_{\beta_0}(\hat{\beta}_N - \beta_0) + o_p(\|\hat{\beta}_N - \beta_0\|)$$

$$
\begin{aligned}
&= E_N^l(\beta_0) + [\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial \beta}](\hat{\beta}_N - \beta_0) + o_p(\|\hat{\beta}_N - \beta_0\|) \\
&= E_N^l(\beta_0) + [\sum_{t=1}^{N} \mathbf{Z}_{t-1} I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial \gamma_{t-1}}](\hat{\beta}_N - \beta_0) + o_p \\
&= E_N^l(\beta_0) + [\sum_{t=1}^{N} \mathbf{Z}_{t-1} I_{[\mathbf{Z}_{t-1}\in A_l]}\frac{\partial p_t}{\partial l}\frac{\partial l}{\partial p_t}\frac{\partial p_t}{\partial \gamma_{t-1}}](\hat{\beta}_N - \beta_0) + o_p \\
&= E_N^l(\beta_0) + [\sum_{t=1}^{N} I_{[\mathbf{Z}_{t-1}\in A_l]}\mathbf{Z}_{t-1}\mathbf{D}_{t-1}\mathbf{\Sigma}_t](\hat{\beta}_N - \beta_0) + o_p,
\end{aligned}
$$

where l is the logit function and $\gamma_{t-1} = \mathbf{Z}'_{t-1}\beta$. So the desired result follows. □

Remark 2.7.1 From the second part of Theorem 2.7.1 we obtain that

$$
\begin{aligned}
\frac{1}{N}(M_N - E_N(\hat{\beta}_N)) &= \frac{1}{N}(M_N - E_N(\beta_0) + E_N(\beta_0) - E_N(\hat{\beta}_N) \\
&\approx \frac{1}{N}(M_N - E_N(\beta_0)) - \sqrt{N}\mathbf{B}'\mathbf{G}(\hat{\beta}_N - \beta_0).
\end{aligned}
$$

It follows that the asymptotic covariance matrix of $(M_N - E_N(\hat{\beta}_N))/N$ is given by $\mathbf{C} - \mathbf{B}'\mathbf{G}\mathbf{B}$. So another useful statistic is

$$
\frac{1}{N}(M_N - E_N(\hat{\beta}_N)'[\mathbf{C} - \mathbf{B}'\mathbf{G}\mathbf{B}]^{-1}(M_N - E_N(\hat{\beta}_N)),
$$

where the inverse is a symmetric generalized inverse. The asymptotic distribution of this statistic is again chi-squared but the number of degrees of freedom is less or equal to $kq - 1$.

Chapter 3

Models for Categorical Time Series with Applications

3.1 Introduction

This chapter presents models that can be used for the statistical analysis of nonstationary categorical time series. We briefly first review models for binary time series. Then we discuss models for nominal and ordinal time series. An application to the field of meteorology is discussed in detail. We close the chapter by considering some further generalizations.

3.2 Models for Binary Time Series

Suppose that $\{y_t\}$ is a binary time series. According to our previous discussion, the general model that links the transition probabilities of $\{y_t\}$ with a linear function of some covariate process can be expressed as

$$p_t = P[y_t = 1 \mid \mathcal{F}_{t-1}] = h(\beta' z_{t-1})$$

where z_{t-1} is a vector of random time dependent covariates and β is a vector of unknown parameters. Since p_t is a probability, the function h must take values between zero and one. Models for binary time series have been previously

considered in [7], [53], [60], [71], and [92]. The case of stationary binary time series is discussed in [49] in detail.

The modeling aspect–the choice of the link function–can be resolved by considering models for independent data (see [25]). The most widely used model is the logit model ([11], [35]). It is given by letting $h(x) = 1/(1 + \exp(-x))$, i.e. the cumulative distribution function of a logistically distributed random variable. The probit model ([34]) is another popular choice. We put $h(x) = \Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Finally, for $h(x) = 1 - \exp(-\exp(-x))$, we get the complementary log-log model. We recognize that this link function corresponds to the extreme-value distribution. Various generalization of these models have been proposed in the literature. Among them are [6], [26], [37], [77], [78], and [97].

All these models can be applied to time series data. A justification can be found in [92, pp.92–93]. Although this paper treats the logistic regression model, the theory applies to other models as well by appropriately modifying the distribution of the error terms.

3.3 Models for Nominal Time Series

3.3.1 Multinomial Logits Model

Suppose now that we observe a categorical time series, say $\{y_t\}$, with m categories. Assume further that the categories do not have any specific structure and are totally unordered. Such a response variable is called a nominal variable. Examples of nominal responses are, for instance, the daily choice of transportation or the daily choice of the newspaper, possibly based on some covariate

information.

The most widely used model for the analysis of such data is the multinomial logits model (see [1], for example). In the setting of time series this is given as

$$p_{tj} = P[y_t = j \mid \mathcal{F}_{t-1}] = \frac{\exp(\beta_j' z_{t-1})}{1 + \sum_{i=1}^{q} \exp(\beta_i' z_{t-1})} \quad j = 1, \ldots, q, \tag{3.1}$$

where β_j is a p-dimensional regression parameter and z_{t-1} is a vector of stochastic time dependent covariates of the same dimension.

There are two lines of argument for the derivation of this model. The first one, which extends the logistic model, is to put

$$\log \frac{p_{tj}}{p_{tm}} = \beta_j' z_{t-1}.$$

Now, recall that $\sum_{j=1}^{m} p_{tj} = 1$. Then (3.1) follows.

The second line of argument is by maximizing a random utility, as described in [68]. Accordingly, the main idea is that an individual chooses a category at each time, independently of the past, such that a maximization of a random utility is obtained.

Observe from (3.1) that

$$\log \frac{p_{tj}}{p_{ti}} = (\beta_j' - \beta_i') z_{t-1}.$$

So, we see that the ratio p_{tj}/p_{ti} for the jth and i^{th} category is the same irrespective of the total number of categories m. This property is usually referred to as independence of irrelevant alternatives.

In this section we let β be the pq-vector

$$\beta = (\beta_1', \ldots, \beta_q')'$$

and $\mathbf{Z}_{t-1}$ be the $qp \times q$ matrix

$$\mathbf{Z}_{t-1} = \begin{bmatrix} z_{t-1} & 0 & \cdots & 0 \\ 0 & z_{t-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_{t-1} \end{bmatrix}.$$

The partial score function is

$$ps_N(\beta) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}(y_t - p_t(\beta)). \tag{3.2}$$

It readily follows that

$$\mathbf{G}_N(\beta) = \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{\Sigma}_t(\beta) \mathbf{Z}'_{t-1}$$

in the notation of Chapter 2. Furthermore the negative of the matrix of second partial derivatives of the partial log-likelihood coincides with the conditional information matrix

$$\mathbf{H}_N(\beta) = \mathbf{G}_N(\beta).$$

The last equality is of great advantage since it guarantees uniqueness of the estimator.

3.3.2 Application to TOGA/COARE Data

We apply now the multinomial logits model to TOGA/COARE data set. This data set is described in Appendix A. We partitioned the original data set into a training set and a testing set. The training data set has 711 observations and the remaining 2613 comprise the test data set.

Covariate	$r=0$		$r=1$		$r=3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept	2.91	.24	7.47	.49	16.25	2.45
	2.65	.23	4.46	.36	12.67	2.47
$\mathrm{FA}(r)_t$	-66.31	6.41	-2691.12	321.15	-9193.23	382.87
	-46.43	5.01	-169.18	20.31	-740.76	159.24
$\mathrm{FA}(r)_{t-3}$	43.26	5.57	1305.45	146.6	-2758.81	422.21
	36.07	4.61	75.79	14.57	-63.29	55.52

Table 3.1: Multinomial Logits fit using $r = 0, 1, 3$ for $Y_t^1(3)$.

Let R_t denote the mean rain rate at time t. We first divide R_t in three categories as follows:

$$Y_t^1(3) = \begin{cases} 1 & \text{if } 0 \leq R_t < 0.004, \\ 2 & \text{if } 0.004 \leq R_t < 0.2, \\ 3 & \text{if } R_t \geq 0.2. \end{cases} \tag{3.3}$$

Let $\mathrm{FA}(r)_t$ denote the fractional area at threshold r, at time t. It is the instantaneous fraction of the area where rain rate exceeds r. In our case $r = 0, 1, 3, 5, 7, 9$. We fitted a series of models by using $\mathrm{FA}(r)_t$, $\mathrm{FA}(r)_{t-3}$ and an intercept as covariates to the training data set. Tables 3.1 and 3.2 give the resulting estimators for $Y_t^1(3)$. We observe that the estimators become larger with higher variability as the threshold increases. This is due to the presence of a lot of zeroes in the training data set, which is implied by the fact that as the threshold increases the fractional area decreases.

Table 3.3 summarizes some diagnostics for the model at hand. The first column contains the threshold levels. The second column is $-2\log pl$. This should be as small as possible. The third column gives the χ^2 test statistic that

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept	11.84	1.26	9.61	.91	7.55	.73
	8.65	1.26	7.13	.89	6.45	.72
$\mathrm{FA}(r)_t$	-13351.47	446.62	-17404.71	229.15	-28674.21	93.24
	-622.25	116.29	-616.28	110.81	-722.87	124.56
$\mathrm{FA}(r)_{t-3}$	-5008.23	483.19	-6040.53	234.65	-7793.65	95.85
	-262.81	71.53	-356.23	86.57	-425.24	102.67

Table 3.2: Multinomial Logits fit using $r = 5, 7, 9$ for $Y_t^1(3)$.

r	$-2\log pl$	χ_4^2	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1040.26	140.61	6.9%	39.1%	52.4%	30.8%
1	408.27	13.99	3.8%	4.4%	29.8%	8.9%
3	197.99	1.97	5.1%	5.7%	6.1%	5.2%
5	242.28	13.31	3.8%	8%	3.9%	5.8%
7	308.98	3.45	3.9%	7.8%	4.5%	5.9%
9	392.43	7.28	6.4%	12.1%	5.5%	9.2%

Table 3.3: Multinomial Logits Model diagnostics for $Y_t^1(3)$.

r	$-2\log pl$	χ_4^2	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1101.73	87.67	30.9%	9.5%	45.5%	26.1%
1	360.98	18.23	5%	8.7%	10.4%	8.1%
3	238.10	10.44	6.7%	10.6%	1.4%	6.8%
5	314.16	17.10	5.1%	11.8%	3.3%	7.4%
7	376.22	23.75	4.8%	13.9%	4.1%	8.4%
9	438.19	25.94	6.3%	14.9%	5.5%	9.7%

Table 3.4: Multinomial Logits Model diagnostics for $Y_t^2(3)$.

was developed in Section 2.7 (see Proposition 2.7.2). This was calculated from the training data set. We used a fixed partition of the covariate space throughout the data analysis. This partition had five cells: [0,0.00001), [0.00001,0.0009), [0.0009,0.005), (0.005,0.01) and [0.1,∞). The degrees of freedom were estimated by subtracting the dimension of the regression parameter from the theoretical degrees of freedom. More specifically, in this case, $2 \times 5 - 6 = 4$. This is only an approximation to the actual degrees of freedom. The appropriate way is described in [92]. The last column of Table 3.3 illustrates the probabilities of misclassification by category and total. The total probability of misclassification is defined as the ratio of the number of misclassified observations to the total number of observations in the test data set. The predicted probabilities were calculated using the estimators obtained from the training data set.

We classify an observation, into the ith category, if $\hat{p}_{ti} = \max(\hat{p}_{t1}, \ldots, \hat{p}_{tm})$. It is evident that the fractional area corresponding to level $r = 3$ appears to give accurate prediction of mean rain rate. In summary, we have an independent

verification of the phenomenon first observed by [52].

The above procedure was repeated. We now use

$$Y_t^2(3) = \begin{cases} 1 & \text{if } 0 \leq R_t < 0.002, \\ 2 & \text{if } 0.002 \leq R_t < 0.1, \\ 3 & \text{if } R_t \geq 0.1. \end{cases} \tag{3.4}$$

Table 3.4 illustrates the same phenomenon as before.

We now categorize R_t in four categories as follows:

$$Y_t^1(4) = \begin{cases} 1 & \text{if } 0 \leq R_t < 0.002, \\ 2 & \text{if } 0.002 \leq R_t < 0.03, \\ 3 & \text{if } 0.03 \leq R_t < 0.1, \\ 4 & \text{if } R_t \geq 0.1. \end{cases} \tag{3.5}$$

Tables 3.5 and 3.6 give the estimators and their standard errors. Table 3.7 displays the same phenomenon as in the case of three categories. The fractional area with rain rate exceeding level 3 seems to give the most accurate prediction of mean rain rate.

Table 3.8 displays the summary diagnostics for

$$Y_t^2(4) = \begin{cases} 1 & \text{if } 0 \leq R_t < 0.004, \\ 2 & \text{if } 0.004 \leq R_t < 0.03, \\ 3 & \text{if } 0.03 \leq R_t < 0.2, \\ 4 & \text{if } R_t \geq 0.2. \end{cases} \tag{3.6}$$

The same phenomenon appears again. Finally, we would like to mention that we also fitted a series of models with five categories. The results were in line with the models with three and four categories.

For an illustration we plot the predicted versus the observed rate for $Y_t^2(4)$ using the test data set (see Figure 3.1). It is interesting to note that the pictures are very close. We see that the multinomial logits model does not classify the

Covariate	$r = 0$		$r = 1$		$r = 3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept	1.75	.18	10.31	.62	17.98	1.81
	1.31	.17	7.66	.63	15.59	1.84
	.64	.18	3.76	.48	10.38	1.75
$FA(r)_t$	-58.05	6.44	-5524.81	26.33	-9196.91	386.31
	-35.11	4.83	-1001.51	113.12	-2896.41	333.91
	-32.73	4.58	-287.9	40.28	-1035.67	185.3
$FA(r)_{t-3}$	32.07	5.09	714.11	29.27	-2885.31	429.4
	22.51	4.31	284.34	52.12	-944.3	205.2
	24.03	4.03	105.82	21.41	-347.8	98.59

Table 3.5: Multinomial Logits fit using $r = 0, 1, 3$ for $Y_t^1(4)$.

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept	12.45	.99	10.38	.79	8.67	.64
	10.76	1.01	9.01	.81	7.62	.65
	6.84	.92	5.54	.71	4.51	.56
$FA(r)_t$	-17749.81	63.87	-22008.31	71.85	-25121.88	92.46
	-2982.61	368.54	-3448.92	427.99	-4041.11	511.21
	-854.25	149.81	-919.73	166.54	-1012.99	185.61
$FA(r)_{t-3}$	-5185.71	64.86	-7179.92	73.38	-9215.11	95.80
	-1617.43	261.82	-2061.35	310.51	-2466.61	387.75
	-599.51	131.67	-688.05	153.52	-725.72	166.92

Table 3.6: Multinomial Logits fit using $r = 5, 7, 9$ for $Y_t^1(4)$.

r	$-2\log pl$	χ^2_6	Probabilities of Misclassification				
			1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Total
0	1486.13	280.24	4.2%	91.2%	93.3%	31.6%	50.5%
1	498.58	43.17	5.2%	15.0%	19.5%	9.3%	11.4%
3	243.50	6.21	1.5%	19.8%	14.4%	1.4%	7.1%
5	434.86	6.93	5.9%	26.1%	15.5%	3.3%	11.5%
7	509.71	31.09	4.8%	31.1%	16.5%	4.1%	12.8%
9	585.73	38.01	6.3%	31.9%	17.6%	5.5%	14.1%

Table 3.7: Multinomial Logits Model diagnostics for $Y_t^1(4)$.

r	$-2\log pl$	χ^2_6	Probabilities of Misclassification				
			1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Total
0	1492.85	258.36	1.5%	93.6%	53.9%	49.8%	43.7%
1	547.75	23.69	3.7%	14.6%	9.6%	29.4%	12.2%
3	269.28	1.09	4.6%	20.7%	5.2%	6.1%	7.8%
5	367.81	13.58	3.4%	30.6%	8.3%	3.9%	9.75%
7	448.24	15.49	3.7%	31.0%	8.8%	4.5%	10.1%
9	496.14	16.13	4.6%	37.8%	11.2%	5.5%	12.5%

Table 3.8: Multinomial Logits Model diagnostics for $Y_t^2(4)$.

observations abnormally. In other words, the worst that can happen for an observation that belongs to the first category is to be classified in the second one. Figure 3.2 gives a plot of the predicted probabilities for model $Y_t^2(4)$. We see that the prediction is very good, in the sense that when one category is predicted with a high probability the other probabilities tend to be close to zero.

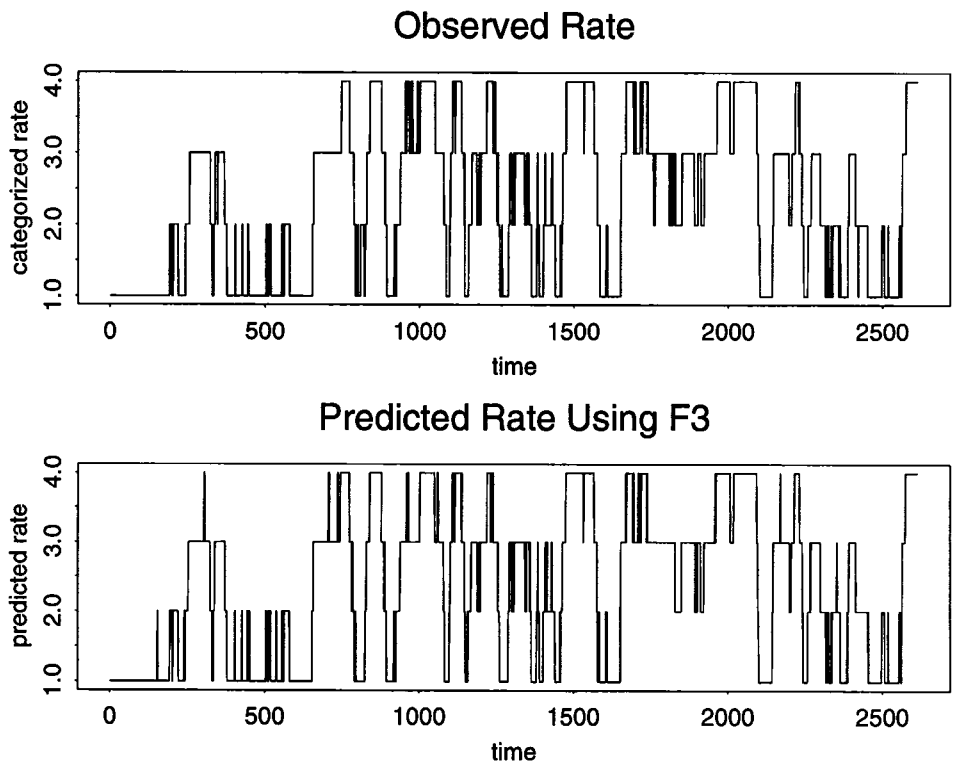

Figure 3.1: Plot of the predicted versus the observed rate using F3=FA(3) as covariate for $Y_t^2(4)$ for the Multinomial Logits Model. Graph based on test data set only.

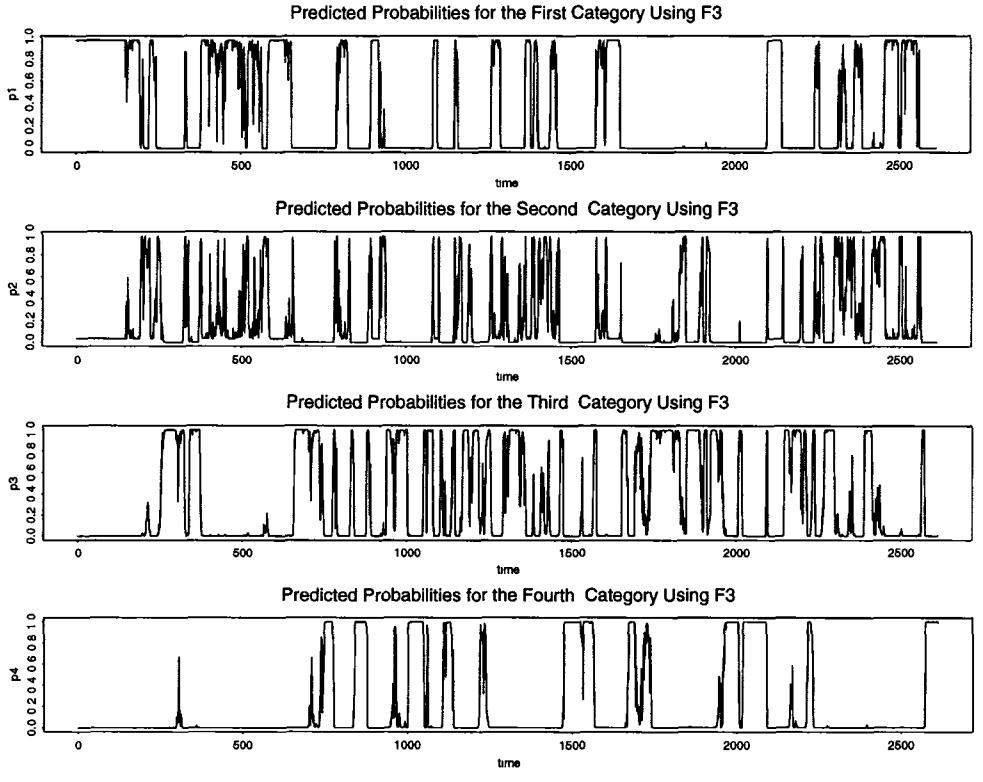

Figure 3.2: Time series plot of the predicted probabilities using F3=FA(3) as covariate for $Y_t^2(4)$ for the Multinomial Logits Model. Graph based on test data set only.

3.4 Models for Ordinal Time Series

3.4.1 Cumulative Odds Model

Assume once again that we observe a categorical time series. Suppose that its categories are ordered. An example is a consumer's opinion about a new product (not satisfactory/good/satisfactory). Such data is called ordinal. The most widely used models for the analysis of ordinal data are the cumulative odds models ([66] ; [93]). We show how one can derive these models by using the method of a latent variable.

Assume that $x_t = -\gamma' z_{t-1} + e_t$, where e_t is a sequence of i.i.d random variables, with cumulative distribution F, γ is a vector of parameters and z_{t-1} is a covariate vector of the same dimension. Suppose that we observe

$$Y_t = j \Longleftrightarrow y_{tj} = 1 \Longleftrightarrow \theta_{j-1} \leq x_t < \theta_j.$$

for $j = 1, \ldots, m$, where $-\infty = \theta_0 < \theta_1 < \ldots < \theta_m = \infty$ are the so called threshold parameters. It follows that

$$\begin{aligned} P(y_t = j \mid \mathcal{F}_{t-1}) &= P(\theta_{j-1} \leq x_t < \theta_j \mid \mathcal{F}_{t-1}) \\ &= F(\theta_j + \gamma' z_{t-1}) - F(\theta_{j-1} + \gamma' z_{t-1}). \end{aligned}$$

The model can be formulated somewhat more compactly by the equation:

$$P(y_t \leq j \mid \mathcal{F}_{t-1}) = F(\theta_j + \gamma' z_{t-1}) \quad j = 1, \ldots, q. \tag{3.7}$$

Since the set of cumulative probabilities corresponds one to one to the set of the response probabilities, estimating the former enables estimation of the latter. Various choices for F can arise. For example, the logistic distribution gives the

so called proportional odds model. In principle, any link that is used for binary time series can be used here as well.

In this section we let β denote the $q+p$ vector

$$\beta = (\theta_1, \ldots, \theta_q, \gamma')'$$

and

$$\mathbf{Z}'_{t-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & z_{t-1} \\ 0 & 1 & \cdots & 0 & z_{t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & z_{t-1} \end{bmatrix},$$

$$h = (h_1, \ldots, h_q),$$

with

$$\begin{aligned} h_1(\beta) &= F(\eta_{(t-1)1}\beta), \\ h_j(\beta) &= F(\eta_{(t-1)j}\beta) - F(\eta_{(t-1)(j-1)}\beta), \quad j = 2, \ldots, q, \end{aligned}$$

and $h_m = 1 - \sum_{j=1}^{q} h_j$, where

$$\eta_{t-1} = (\eta_{(t-1)1}, \ldots, \eta_{(t-1)q}) = \mathbf{Z}_{t-1}\beta.$$

With this notation, the partial score function for the cumulative odds model becomes

$$ps_N(\beta) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_{t-1}(\beta)(y_t - p_t(\beta)) \tag{3.8}$$

where $\mathbf{U}_{t-1}(\beta) = [\partial(l \circ h)/\partial\eta_{t-1}]$, and l is the logit function. It follows that the conditional information matrix is

$$\mathbf{G}_N(\beta) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_{t-1}(\beta)\mathbf{\Sigma}_t(\beta)\mathbf{U}'_{t-1}(\beta)\mathbf{Z}'_{t-1}$$

where $\mathbf{\Sigma}_t(\beta)$ is as before (see Chapter 2). The negativeof the matrix of partial second derivatives is

$$\mathbf{H}_N(\beta) = \mathbf{G}_N(\beta) - \mathbf{R}_N(\beta) \tag{3.9}$$

where $\mathbf{R}_N(\beta) = \sum_{t=1}^{N} \sum_{r=1}^{q} \mathbf{Z}_{t-1} \mathbf{W}_{(t-1)r}(\beta) \mathbf{Z}'_{t-1}(y_{tr} - p_{tr}(\beta))$ and $\mathbf{W}_{(t-1)r}(\beta) = [\partial^2(l \circ h)/\partial\eta_{t-1}\partial\eta'_{t-1}]$. Uniqueness of the estimator can be established in the case that F is log-concave, e.g. if F the logistic cumulative distribution function (see the discussion after [66]).

3.4.2 Revisiting TOGA/COARE Data

We continue our investigation using the TOGA/COARE data set by again fitting a series of models. Tables 3.9 and 3.10 give the estimators and the corresponding standard errors of a proportional odds model fit for $Y_t^1(3)$ (see eq. (3.7)). Recall that the proportional odds model means that F is the logistic distribution function. Table 3.11 displays the diagnostics for each model. The degrees of freedom used in the second column were calculated as $2 \times 5 - 4 = 6$. We see once again that the fractional area exceeding level three seems to be the most useful covariate in the prediction of rain rate. Although a direct comparison of the multinomial logit model and a proportional odds model is not possible, we would like to mention (see Table 3.3) that the former gave a better prediction. We also used $Y_t^2(3)$ as before (see (3.4)). The summary diagnostics are illustrated in Table 3.12.

Our investigation continues by considering $Y_t^1(4)$ (eq. (3.5)) and $Y_t^2(4)$ (eq. 3.6), i.e., models with four categories. Tables 3.13 and 3.14 display the estimators for $Y_t^1(4)$. Note that the estimators become larger as the threshold increases as in the case of the multinomial logits model. This is true for the three category

Covariate	$r = 0$		$r = 1$		$r = 3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	.01	.10	.33	.10	1.69	.17
Intercept2	3.30	.19	6.92	.59	55.59	6.43
FA$(r)_t$	-36.29	3.32	-226.2	23.31	-3186.13	387.82
FA$(r)_{t-3}$	23.55	2.95	73.72	13.50	-435.85	136.83

Table 3.9: Proportional Odds Model fit using $r = 0, 1, 3$ for $Y_t^1(3)$.

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	1.05	.13	.74	.12	.03	.11
Intercept2	26.37	3.20	16.64	1.89	11.23	1.29
FA$(r)_t$	-1962.76	258.43	-1528.03	207.02	-1324.61	197.72
FA$(r)_{t-3}$	-852.12	149.35	867.51	137.63	-785.2	143.21

Table 3.10: Proportional Odds Model fit using $r = 5, 7, 9$ for $Y_t^1(3)$.

r	$-2 \log pl$	χ_6^2	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1079.89	160.31	16.2%	20.5%	53.3%	25.1%
1	708.01	147.01	.5%	15.1%	27.1%	12.5%
3	341.53	20.42	1.3%	8.2%	4.9%	5.3%
5	457.61	86.61	.6%	15.7%	3.1%	7.9%
7	530.01	115.99	.3%	18.7%	3.3%	9.6%
9	591.92	127.55	13.2%	11.4%	4.5%	10.3%

Table 3.11: Proportional Odds Model diagnostics for $Y_t^1(3)$.

r	$-2\log pl$	χ^2_6	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1114.46	222.34	96.9%	5.9%	47.6%	43.9%
1	612.48	160.80	6.8%	11.6%	8.2%	9.2%
3	335.58	45.69	4.6%	13.4%	1.3%	7.3%
5	482.19	78.24	3.0%	19.5%	2.8%	9.9%
7	540.62	82.25	3.6%	19.3%	3.9%	10.3%
9	587.68	120.82	3.8%	21.4%	4.9%	11.5%

Table 3.12: Proportional Odds Model diagnostics for $Y_t^2(3)$.

Covariate	$r=0$		$r=1$		$r=3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-.34	.10	.29	.11	1.10	.15
Intercept2	1.11	.11	3.18	.21	9.26	.97
Intercept3	2.30	.14	8.98	.67	33.10	3.37
$\mathrm{FA}(r)_t$	-35.37	3.38	-637.81	49.47	-3368.42	353.75
$\mathrm{FA}(r)_{t-3}$	22.05	2.96	182.32	21.52	-1164.45	148.72

Table 3.13: Proportional Odds Model fit using $r = 0, 1, 3$ for $Y_t^1(4)$.

cases as well. Tables 3.15 and 3.16 summarize the diagnostics for each series of models respectively. The same situation is apparent again. Figures 3.3 and 3.4 exhibit plots of the observed versus the predicted rate and of the predicted probabilities for the proportional odds model respectively, applied to $Y_t^1(4)$ using as a covariate the fractional area exceeding level three. The plots are based on the test data set only. The same comments apply here as well.

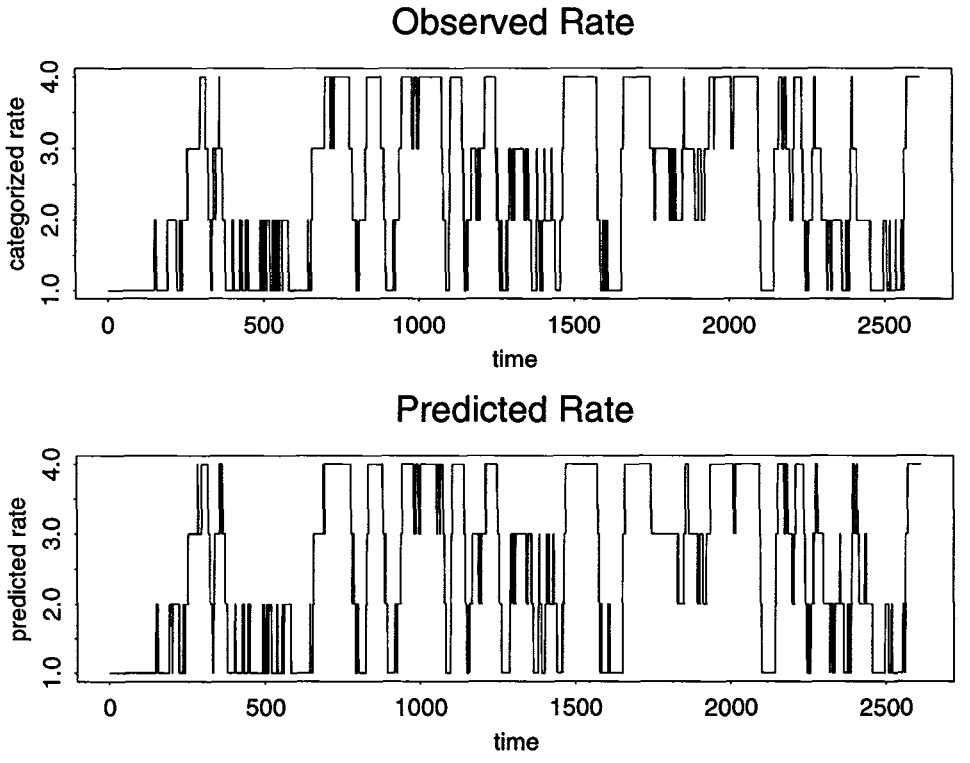

Figure 3.3: Plot of the predicted versus the observed rate using F3=FA(3) as covariate for $Y_t^1(4)$ for the proportional odds model. Graph based on the test data set only.

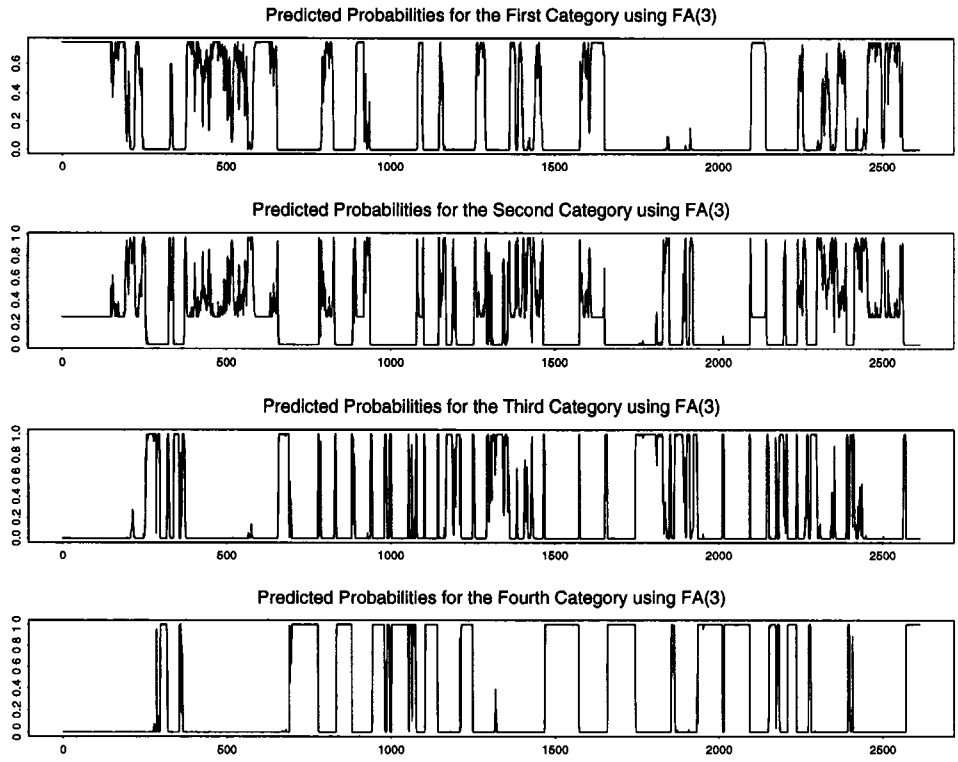

Figure 3.4: Time series plot of the predicted probabilities using F3=FA(3) as covariate for $Y_t^1(4)$ for the proportional odds model. Graph based on test data set only.

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	.54	.12	.35	.11	.12	.11
Intercept2	4.42	.34	3.71	.27	2.79	.19
Intercept3	15.50	1.29	12.08	.97	17.59	1.41
$\text{FA}(r)_t$	-2135.21	216.35	-2215.43	227.51	-2146.13	221.23
$\text{FA}(r)_{t-3}$	-1265.01	158.23	-1453.16	178.17	-1223.63	157.76

Table 3.14: Proportional Odds Model fit using $r = 5, 7, 9$ for $Y_t^1(4)$.

r	$-2\log pl$	χ^2_{10}	Probabilities of Misclassification				
			1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Total
0	1503.05	295.46	5.4%	78.5%	97.2%	29.6%	47.7%
1	793.14	201.27	2.6%	23.7%	27.8%	7.5%	13.7%
3	416.41	62.98	3.4%	27%	14.9%	1.3%	10.3%
5	608.43	77.46	2.1%	37.2%	19.2%	2.9%	13.6%
7	682.22	82.04	2.8%	38.4%	21.3%	3.9%	14.4%
9	787.45	119.89	2.7%	42.8%	16.9%	4.3%	16.3%

Table 3.15: Proportional Odds Model diagnostics for $Y_t^1(4)$.

r	$-2\log pl$	χ^2_{10}	Probabilities of Misclassification				
			1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Total
0	1492.85	258.36	1.5%	96.9%	53.9%	49.8%	43.7%
1	960.90	171.65	0.0%	82.8%	45.9%	26.5%	24.3%
3	415.60	25.35	1.1%	28.4%	6.4%	4.7%	8.1%
5	582.93	67.45	.6%	44.8%	10.4%	3.1%	11.7%
7	681.32	94.54	.3%	49.8%	13.1%	3.3%	13.4%
9	745.65	121.24	.1%	59.6%	15.9%	4.3%	16.1%

Table 3.16: Proportional Odds Model diagnostics for $Y_t^2(4)$.

Covariate	$r = 0$		$r = 1$		$r = 3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-.01	.06	.08	.06	-	-
Intercept2	1.85	.09	3.17	.22	-	-
$FA(r)_t$	-17.88	1.69	-97.08	9.11	-	-
$FA(r)_{t-3}$	11.06	1.55	33.73	6.23	-	-

Table 3.17: Cumulative Odds Model fit with probit link using $r = 0, 1, 3$ for $Y_t^1(3)$.

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-	-	.31	.06	-.06	.06
Intercept2	-	-	6.08	.59	4.63	.46
$FA(r)_t$	-	-	-578.23	74.88	-565.81	79.56
$FA(r)_{t-3}$	-	-	-312.92	54.26	-314.91	62.02

Table 3.18: Cumulative Odds Model with probit link using $r = 5, 7, 9$ for $Y_t^1(3)$.

r	$-2 \log pl$	χ_6^2	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1088.21	165.09	21.6%	16.5%	56.7%	25.7%
1	741.67	124.92	3.3%	8.2%	33.5%	11.3%
3	-	-	-	-	-	-
5	-	-	-	-	-	-
7	569.09	118.94	.2%	21.1%	3.3%	10.7%
9	613.93	135.63	95.5%	4.5%	4.3%	31.6%

Table 3.19: Cumulative Odds Model diagnostics with probit link for $Y_t^1(3)$.

Covariate	$r = 0$		$r = 1$		$r = 3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-.33	.07	-.06	.07	-	-
Intercept2	1.68	.09	4.23	.40	-	-
$FA(r)_t$	-18.29	1.89	-164.52	18.47	-	-
$FA(r)_{t-3}$	9.38	1.67	49.43	10.11	-	-

Table 3.20: Cumulative Odds Model with clog-log link using $r = 0, 1, 3$ for $Y_t^1(3)$.

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-	-	.74	.12	-.49	.07
Intercept2	-	-	16.64	1.89	3.41	.34
$FA(r)_t$	-	-	-1528.08	207.07	-361.22	68.28
$FA(r)_{t-3}$	-	-	867.51	137.62	-418.91	70.14

Table 3.21: Cumulative Odds Model with clog-log link using $r = 5, 7, 9$ for $Y_t^1(3)$.

r	$-2 \log pl$	χ_6^2	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1061.48	129.51	25.8%	13.3%	44.3%	23.2%
1	691.14	117.08	0%	20.7%	20.8%	13.8%
3	-	-	-	-	-	-
5	-	-	-	-	-	-
7	619.38	111.56	.5%	26.1%	2.2%	13.1%
9	664.92	125.61	96.5%	5.8%	2.4%	32.1%

Table 3.22: Cumulative Odds Model diagnostics with clog-log link for $Y_t^1(3)$.

Covariate	$r = 0$		$r = 1$		$r = 3$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-.54	.07	-	-	.14	.08
Intercept2	.52	.06	-	-	2.76	.26
Intercept3	1.25	.07	-	-	9.59	.88
FA$(r)_t$	-17.67	1.88	-	-	-1124.51	111.07
FA$(r)_{t-3}$	6.91	1.61	-	-	-395.07	57.95

Table 3.23: Cumulative Odds Model fit with clog-log link using $r = 0, 1, 3$ for $Y_t^1(4)$.

Covariate	$r = 5$		$r = 7$		$r = 9$	
	Estim.	S. E.	Estim.	S. E.	Estim.	S. E.
Intercept1	-.16	.07	-.24	.07	-.41	.07
Intercept2	1.46	.09	1.29	.08	.91	.06
Intercept3	4.12	.31	3.51	.25	4.39	.35
FA$(r)_t$	-682.23	74.86	-768.12	92.26	-489.31	72.70
FA$(r)_{t-3}$	-485.92	67.01	-628.43	86.23	-60251	76.46

Table 3.24: Cumulative Odds Model fit with clog-log link using $r = 5, 7, 9$ for $Y_t^1(4)$.

r	$-2\log pl$	χ^2_{10}	Probabilities of Misclassification				
			1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Total
0	1488.69	259.89	5.9%	81.0%	93.3%	24.7%	47.1
1	-	-	-	-	-	-	-
3	505.83	117.21	1.1%	34.9%	22.2%	.4%	12.7%
5	735.18	168.75	1.6%	40.8%	34.3%	1.2%	16.8%
7	797.57	180.61	2.4%	41.6%	38.9%	2.5%	18.5%
9	919.56	228.44	2.7%	44.7%	28.1%	1.6%	19.5%

Table 3.25: Cumulative Odds Model diagnostics using clog-log link for $Y_t^1(4)$.

We fitted also a series of models using probit and complementary log-log links. Unfortunately, in some cases, convergence was not attained due to singularities of the conditional information matrix. The rest of the tables (Tables 3.17-3.25) display the fit and the diagnostics. The probit model is inconclusive since we did not have convergence for $r = 3$. However a close look at the clog-log (complementary log-log) models for $Y_t^1(4)$ reveals the fact that we have observed before, namely, $r = 3$ is the optimal threshold.

3.5 A Random Coefficients Model

Up to this point, we considered models with time invariant parameters. This is not always necessarily the case. A possible generalization would be that some of the coefficients are realizations of a random variable. For a comprehensive account of this subject in the case of the usual autoregressive processes, see [75]. Let us be more specific. We use the same notation as in Chapter 2. Then,

according to what we discussed, set

$$p_{tj} = P[y_{tj} = 1 \mid \mathcal{F}_{t-1}, a_t] = h_j(\mathbf{Z}'_{t-1}\beta + \mathbf{U}'_{t-1}a_t), \;\; j = 1, \ldots, q, \tag{3.10}$$

with $\mathbf{U}_{t-1}$ is a $d \times q$ stochastic covariate process (possibly a subset of $\mathbf{Z}_{t-1}$) and a_t are independently identically distributed q-dimensional random variables with continuous density $g(a_t, \theta)$, θ being a parameter. Let's suppose that a_t is independent of $(\mathbf{U}_{t-1}\mathbf{Z}_{t-1})$ for all t. In addition assume that a_t is independent of $\mathcal{F}_{t-1}$ for all t.

We would like to compute the partial likelihood for this model. By definition of the partial likelihood we have

$$\begin{aligned} f(y_t \mid \mathcal{F}_{t-1}) &= \int f(y_t, a_t \mid \mathcal{F}_{t-1}) da_t \\ &= \int f(y_t \mid a_t, \mathcal{F}_{t-1}) g(a_t, \theta) da_t \\ &= \int \prod_{j=1}^{m} p_{tj}(\beta, a_t)^{y_{tj}} g(a_t, \theta) da_t \\ &= \prod_{j=1}^{m} (\int p_{tj}(\beta, a_t) g(a_t, \theta) da_t)^{y_{tj}} \end{aligned}$$

where the second equality follows by the independence assumption and the last equality upon noting that y_{tj} can take the value 0 or 1. In conclusion we see that the partial likelihood for the model (3.10) is given by

$$PL(\beta) = \prod_{t=1}^{N} \prod_{j=1}^{m} (\int p_{tj}(\beta, a_t) g(a_t, \theta) da_t)^{y_{tj}} \tag{3.11}$$

This shows that the random coefficients assumption results to a modification of the link function. In particular, suppose that

$$p_{tj} = h_j(\mathbf{Z}'_{t-1}\beta + a_t), \;\; j = 1, \ldots, q$$

the so called random intercept model. For the sake of argument, let h_j be a

cumulative distribution function. Then, the new link function, namely

$$\int h_j(\mathbf{Z}'_{t-1}\beta + a_t)g(a_t, \theta)da_t,$$

is just the convolution of two random variables; one that has distribution h_j and another which has density g.

Now, assume that θ is known and consider (3.11). Then, under the following two additional assumptions the results from Chapter 2 are applicable.

RC1

$$\frac{\partial}{\partial\beta}\int p_{tj}(\beta, a_t)g(a_t, \theta)da_t = \int \frac{\partial}{\partial\beta}p_{tj}(\beta, a_t)g(a_t, \theta)da_t.$$

RC2

$$\frac{\partial^2}{\partial\beta\partial\beta'}\int p_{tj}(\beta, a_t)g(a_t, \theta)da_t = \int \frac{\partial^2}{\partial\beta\partial\beta'}p_{tj}(\beta, a_t)g(a_t, \theta)da_t.$$

In particular, if the integrand of (3.11) is a convex function of both β and a, then the partial maximum likelihood estimator is unique and satisfies the conclusion of Theorem 2.6.1, according to [18].

If θ is not known the problem becomes more challenging, but we will not proceed any further. However, there is a growing literature on this subject (see for example [103]) in the independent data case which may be useful for time series data.

Chapter 4

Adaptive Control of Binary Time Series

4.1 Introduction

The usual ARMAX methodology provides a complete framework for the adaptive control of linear models. In this chapter we will extend this methodology to cover binary time series. A particular example is the adaptive control of Markov processes with two states. By employing a logistic model, we will analyze a recursive estimation procedure and an adaptive control law. This enables the observer to regulate the transition probabilities of the system. Some geometric properties of the algorithm are discussed. These indicate that the proposed control law is asymptotically optimal with respect to a certain criterion.

4.2 A Brief Tour of Adaptive Control of Linear Models

We are going to review here some basic results from the theory of adaptive control for linear models. Appropriate definitions will be given where it is necessary.

The most general linear model, the so-called ARMAX (Autoregressive Moving Average with External Input) model, is fully described by the following stochastic difference equation:

$$y_{t+1} = \sum_{i=0}^{p_1} a_i y_{t-i} + \sum_{i=0}^{p_2} b_i u_{t-i-d} + \sum_{i=0}^{p_3} c_i w_{t-i} + w_{t+1} \tag{4.1}$$

In the above, y_t is the output and u_t is the external input at time t, while w_t is a sequence of uncorrelated random variables with zero mean and common finite variance, i.e., $\{w_t\}$ is white noise. The assumed known nonnegative constant d is called the delay of the system. The above representation makes evident why such models are termed as ARMAX. The first term on the right hand side of (4.1) stands for the autoregressive part, the second term is the external input, and the last one represents the moving average part. A system of this kind can be regarded as an input-output model. The observer wants to control the input u_t so that the output is close to some target value by means of some feedback mechanism. The problem becomes even more complicated since the parameters $\{a_i, b_i, c_i\}$ are assumed to be unknown.

Our attention will be focused on ARX models (Autoregressive with External Input). In other words, we will assume that $c_i = 0$ for every $i = 0, \ldots p_3$. In addition we assume that there is no delay, i.e., $d = 0$, and $p_1 = p_2$. In case that $p_1 \neq p_2$ we can set $p = \max(p_1, p_2)$. For an excellent treatment of ARMAX models see [12], [17], [27], [39], and [56].

Under the above assumptions, the resulting model is reduced to

$$\begin{aligned} y_{t+1} &= \sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=0}^{p} b_i u_{t-i} + w_{t+1} \\ &= \beta' x_t + w_{t+1} \end{aligned} \tag{4.2}$$

with $\beta = (a_0, \ldots, a_p, b_0, \ldots, b_p) \in R^{2p+2}$ and $x_t = (y_t, \ldots, y_{t-p}, u_t, \ldots, u_{t-p})$. The task is to estimate the parameter vector β and then to control such a process. Our exposition is in line with [55].

4.2.1 Least Squares Estimation

Suppose first the we have N available observations from (4.2). Define $Y_N = (y_1, \ldots, y_N)$ and $\mathbf{X}_N = (x_0, \ldots, x_{N-1})'$ Then the well-known least squares estimator of β is given by

$$\hat{\beta}_N = (\mathbf{X}_N'\mathbf{X}_N)^{-1}\mathbf{X}_N'Y_N, \tag{4.3}$$

provided that the above inverse exists. Suppose now that a new piece of data becomes available, say (y_{N+1}, x_N). The new estimator, which is based on $N+1$ observations, can be calculated explicitly by updating (4.3). Define $\mathbf{R}_N = \sum_{t=0}^{N} x_t x_t'$. Then the following recursions can be used for updating (4.3):

$$\hat{\beta}_{N+1} = \hat{\beta}_N + \mathbf{R}_N^{-1} x_N (y_{N+1} - x_N'\hat{\beta}_N), \tag{4.4}$$

$$\mathbf{R}_N = \mathbf{R}_{N-1} + x_N x_N'. \tag{4.5}$$

The above recursions require the inversion of a $(2p+2) \times (2p+2)$ matrix. This can be avoided by using the following equivalent expressions:

$$\mathbf{P}_N = \mathbf{P}_{N-1} - \frac{\mathbf{P}_{N-1} x_N x_N' \mathbf{P}_{N-1}}{1 + x_N'\mathbf{P}_{N-1}x_N} \tag{4.6}$$

$$\hat{\beta}_{N+1} = \hat{\beta}_N + \frac{\mathbf{P}_{N-1}x_N}{1 + x_N'\mathbf{P}_{N-1}x_N}(y_{N+1} - x_N'\hat{\beta}_N) \tag{4.7}$$

with $\mathbf{P}_N = \mathbf{R}_N^{-1}$. These recursions can be proved by utilizing a matrix inversion lemma [81]. We state it here for the sake of completeness.

Lemma 4.2.1 Suppose that $\mathbf{S}$, $\mathbf{Q}$ are non-singular matrices of order n and m respectively, $\mathbf{G}$ is an $m \times n$ matrix, and $\mathbf{H}$ is an $n \times m$ matrix. Then

$$(\mathbf{S} + \mathbf{GQH})^{-1} = \mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{G}(\mathbf{HS}^{-1}\mathbf{G} + \mathbf{Q}^{-1})^{-1}\mathbf{HS}^{-1}.$$

Proof: By multiplying the right-hand side of the above relation by $(\mathbf{S} + \mathbf{GQH})$ we obtain the result. □

Assume now that $Var(w_t \mid\mid \mathcal{F}_{t-1}) = \sigma^2 v_{t-1}$, with v_t a set of known positive weights. Then the recursive formulas (4.4)-(4.5) and (4.6)-(4.7) are modified. By transforming the weighted least squares problem to an ordinary one by the usual transformation and by letting $\tilde{\mathbf{R}}_N = \sum_{t=0}^{N}(x_t x_t')/v_t$, we get that

$$\hat{\beta}_{N+1} = \hat{\beta}_N + \tilde{\mathbf{R}}_N^{-1}\frac{x_N}{v_N}(y_{N+1} - x_N'\hat{\beta}_N) \tag{4.8}$$

$$\tilde{\mathbf{R}}_N = \tilde{\mathbf{R}}_{N-1} + \frac{x_N x_N'}{v_N} \tag{4.9}$$

for (4.4)-(4.5). Furthermore, by letting $\tilde{\mathbf{P}}_N = \tilde{\mathbf{R}}_N^{-1}$, the expressions for (4.6)-(4.7) become

$$\tilde{\mathbf{P}}_N = \tilde{\mathbf{P}}_{N-1} - \frac{\tilde{\mathbf{P}}_{N-1}x_N x_N'\tilde{\mathbf{P}}_{N-1}}{v_N + x_N'\tilde{\mathbf{P}}_{N-1}x_N}, \tag{4.10}$$

$$\hat{\beta}_{N+1} = \hat{\beta}_N + \frac{\tilde{\mathbf{P}}_{N-1}x_N v_N}{v_N + x_N'\tilde{\mathbf{P}}_{N-1}x_N}(y_{N+1} - x_N'\hat{\beta}_N), \tag{4.11}$$

respectively. For the general asymptotic theory of least squares estimators we refer to [57]. We quote the results of that paper in the following theorem.

Theorem 4.2.1 Suppose that (4.2) holds, and $\{w_t\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{F}_t$ such that

$$\sup_t E[|w_t|^\alpha \mid\mid \mathcal{F}_{t-1}] < \infty \quad \text{for some} \quad \alpha > 2.$$

In addition assume that x_t is $\mathcal{F}_t$-measurable and let $\mathbf{R}_N$ be as before. Then if $\lambda_{\min}(\mathbf{R}_N) \to \infty$ a.s. and $\log \lambda_{\max}(\mathbf{R}_N) = O((\log \lambda_{\max}(\mathbf{R}_N)/\lambda_{\min}(\mathbf{R}_N))^{1/2})$ a.s. , we have that $\hat{\beta}_N$ converges a.s. to β.

Note that this theorem is true under the assumption of white noise or equivalently of the ARX model. We need to point out here that this is not the case if the noise entering the system is "colored", i. e. the ARMAX model.

4.2.2 Minimum Variance Control of ARX Models

Suppose once again that the data is generated by the model (4.2). Furthermore, assume that $b_0 \neq 0$, $E[w_t] = 0$ and $E[w_t^2] = \sigma^2$, and let's suppose for a moment that β is known. The task will be to choose a control law such that the average squared error, $\sum_{t=1}^{N} y_t^2/N$, becomes minimum. In other words, we would like to have the output as close to zero as possible. If we choose the control input at each time t by

$$u_t = -\frac{1}{b_0}[\sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=1}^{p} b_i u_{t-i}] = -\frac{A(z)}{B(z)} y_t \tag{4.12}$$

with $A(z) = \sum_{i=0}^{p} a_i z^i$ and $B(z) = \sum_{i=0}^{p} b_i z^i$, where z is the lag operator, then the variance of the output becomes $N\sigma^2$, and this is the minimum value we can attain. However a problem arises by using this control law. This becomes clear in the next example (see [55]).

Example 4.2.1 Consider the following system

$$y_{t+1} = -y_t + u_t - 5u_{t-1} + w_{t+1}.$$

By applying the procedure (4.12), the optimal control law is given by

$$u_t = y_t + 5u_{t-1}.$$

Under this control law $y_{t+1} = w_{t+1}$, so we have $u_t = 5u_{t-1} + w_t$ which is an unstable difference quation, since u_t explodes.

The problem is resolved by assuming the minimum phase condition (MPC), namely, that all roots of $B(z)$ lie strictly outside the unit circle.

Definition 4.2.1 A system which satisfies the minimum phase assumption is called a minimum phase system.

Under the assumption that we have a minimum phase system, the control law (4.12) is optimal in the sense that it minimizes the variance output. Note that we can use other criteria as cost functions. Then the control law would have been modified. For detailed examples see [56].

4.2.3 The Self-tuning Regulator

Consider again the ARX model. The problem is again to minimize the variance of the output. If the system satisfies the minimum phase assumption, then we saw that (4.12) is optimal provided that all the coefficients are known. But in practice the system is never completely known and an estimator must be used. From the least squares recursions (4.4)-(4.5) we have an estimator at each time t. Denote this by [1] $\hat{\beta}_t = (\hat{a}_0(t), \ldots, \hat{a}_p(t), \hat{b}_0(t), \ldots, \hat{b}_p(t))$. The principle of estimation and control (PEC) ([44, pp. 38]) chooses the optimal control law at time t by simply substituting estimates into the optimal control law. That is:

$$u_t = -\frac{1}{\hat{b}_0(t)}[\sum_{i=0}^{p} \hat{a}_i(t) y_{t-i} + \sum_{i=1}^{p} \hat{b}_i(t) u_{t-i}]. \tag{4.13}$$

In other words, we choose u_t implicitly from the equation $x_t'\hat{\beta}_t = 0$. Such a control law is called adaptive. This is an on-line procedure. Given the data up to time t, $(y_0, u_0, \ldots, y_t)$, we update the estimators and a control law is calculated by means of (4.13). This is applied to the system and an output y_{t+1} is obtained. The process goes on by calculating a control u_{t+1} and so on.

We now define two optimality properties that will be useful subsequently.

Definition 4.2.2 An adaptive control law is self-optimizing, with respect to

[1] In spite of the fact that this symbolism is usually reserved for continuous time stochastic processes, we use it here for convenience.

some criterion, if it yields (asymptotically) the same cost as the optimal control law that would have been used if the system were completely known.

Definition 4.2.3 An adaptive control law which asymptotically approaches the law that would have been used if the system were known is called self-tuning.

The least squares algorithm does not provide a convenient framework for the analysis of the system and the control law (see the discussion in [55]). So the following modification is used, which is called the stochastic gradient algorithm.

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \gamma \frac{x_t}{r_t}(y_{t+1} - x_t'\hat{\beta}_t); \tag{4.14}$$

$$r_t = r_{t-1} + x_t' x_t; \quad r(0) = 1 \tag{4.15}$$

where γ is a positive constant. Note that the difference between the above expressions and (4.4)-(4.5) is the replacement of the matrix $\mathbf{R}_N$ by its trace. The control law is chosen by (4.13).

The first work on self-tuning regulators is from [8]. For the analysis of this control law using the least squares algorithm by ordinary differential equations see [62], [63]. In [38] martingale methods were used to exhibit self-optimality of (4.14)-(4.15). The following theorem was proved there:

Theorem 4.2.2 Suppose that $E[w_{t+1} \mid\mid \mathcal{F}_t] = 0$, a.s. $\forall\ t$, $E[w_{t+1}^2 \mid\mid \mathcal{F}_t] = \sigma^2 > 0$, a.s. $\forall\ t$, $\sup_t E[w_{t+1}^4 \mid\mid \mathcal{F}_t] < \infty$ a.s. and the system (4.2) is of minimum phase. If the stochastic gradient based adaptive control (4.13)-(4.14)-(4.15) is used, the following statements are true.

1. The adaptive control law is self optimizing with respect to the minimum variance criterion. That is

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} y_t^2 = \sigma^2 \text{ a.s.}$$

2. Moreover

$$\sum_{t=1}^{\infty}\left(\frac{E[y_t \mid\mid \mathcal{F}_{t-1}]}{r_{t-1}}\right)^2 < \infty \text{ a.s.,}$$

$$\lim_{N\to\infty}\sum_{t=1}^{N}(E[y_t \mid\mid \mathcal{F}_{t-1}])^2 = 0 \quad \text{a.s.,}$$

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}u_t^2 < \infty,$$

$$||\hat{\beta}_t - \beta|| \quad \text{converges a.s.}$$

The main ingredient of the proof is the following result, which we state here since we are going to use it too. The proof can be found in [83].

Theorem 4.2.3 Let $\{M_t\}$, $\{s_t\}$, and $\{d_t\}$ be nonnegative stochastic processes adapted to an increasing sequence of σ-fields $\{\mathcal{F}_t\}$ such that

$$E[M_{t+1} \mid\mid \mathcal{F}_t] \leq M_t - s_t + d_t \quad \text{a.s. and } \sum_t d_t < \infty \quad \text{a.s.}$$

Then M_t converges a.s. and $\sum_t s_t < \infty$ a.s.

Until now, we have seen that the recursions (4.13)-(4.14)-(4.15) give a self-optimizing law. The problem of self-tuning has been considered in [10]. In that work, it was proved that the aforementioned algorithm has crucial geometric properties that can be heavily utilized for proving convergence of the adaptive control law. Before we state the main theorems of [10], we will need a couple of definitions.

Suppose that $\beta = (\beta_1, \ldots, \beta_{2p+2})'$ is the estimated parameter at time t. Then the control law

$$u_t = -\frac{\beta_1 + \beta_2 z + \ldots + \beta_{p+1}z^p}{\beta_{p+2} + \beta_{p+3}z + \ldots + \beta_{2p+2}z^p}y_t$$

is applied to the system at each time t, with z the lag operator. Recall now that the minimum variance control law for (4.2) is given by (4.12).

Definition 4.2.4 A parameter β produces a minimum variance control law if

$$\frac{\beta_1 + \beta_2 z + \ldots + \beta_{p+1} z^p}{\beta_{p+2} + \beta_{p+3} z + \ldots + \beta_{2p+2} z^p} = \frac{a_0 + a_1 z + \ldots + a_p z^p}{b_0 + b_1 z + \ldots + b_p z^p}$$

Let $\mathcal{B}$ denote the set of all parameters that produce a minimum variance control law.

Definition 4.2.5 The adaptive control-law is self-tuning in the Cesaro sense if

$$\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} I_{[\hat{\beta}_t \in O]} = 1 \quad \text{a.s.}$$

for every open set O contained in $\mathcal{B}$.

Definition 4.2.6 We say that the ARX model (4.2) does not have reduced order minimum variance control law if there do not exist polynomials $F(z)$ and $G(z)$, both of degree less or equal to $p-1$, such that

$$\frac{F(z)}{G(z)} = \frac{a_0 + a_1 z + \ldots + a_p z^p}{b_0 + b_1 z + \ldots + b_p z^p}$$

We can now state the main results of [10].

Theorem 4.2.4 The stochastic gradient adaptive control law (4.13),(4.14), (4.15) is self-tuning in the Cesaro sense.

Theorem 4.2.5 Assume the same hypotheses of Theorem 4.2.2. In addition, suppose that the system has no reduced order minimum variance control law. Then, if the stochastic gradient adaptive control law is used, we have the following.

1. The parameter estimates $\{\hat{\beta}_t\}$ converge to some random multiple of β. That is

$$\lim_{t\to\infty} \hat{\beta}_t = k\beta \quad \text{a.s. for some scalar random variable } k.$$

2. The adaptive control law is self-tuning. That is,

$$\lim_{t\to\infty} (u_t - \frac{A(z)}{B(z)} y_t) = 0 \quad \text{a.s.}$$

4.3 An Extension to the Logistic Model

In this section, we establish the framework for generalizing the previous results to the case of binary data. Let us be more specific. Assume that we observe a system. The output of the system, denoted again by $\{y_t\}$, is a binary variable. We would like to build a class of models, such as the ARX class, so that the observer can control the transition probabilities of the system. By our exposition from the previous chapters, a natural candidate for modeling such a process would be

$$\lambda_{t+1} = \log \frac{p_{t+1}}{1 - p_{t+1}} = \sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=0}^{p} b_i u_{t-i}. \tag{4.16}$$

In the above, $p_{t+1} = P[y_{t+1} = 1 \mid\mid \mathcal{F}_t]$, where $\mathcal{F}_t$ stands for the σ-field generated by past observations, $\{a_i, b_i\}$ are unknown coefficients and $\{u_t\}$ is the control sequence. Let $\beta = (a_0, \ldots, a_p, b_0, \ldots, b_p)' \in R^{2p+2}$, $z_t = (y_t, \ldots, y_{t-p}, u_t, \ldots, u_{t-p})'$ and $b_0 \neq 0$. Therefore (4.16) can be written as

$$\lambda_{t+1} = \beta' z_t, \tag{4.17}$$

which closely resembles equation (4.2). We will first consider the problem of regulation. Equivalently we will try to keep the transition probabilities close to

1/2, so that in the long run we can expect equal numbers of occurrences of both states. To this end, we will need a recursive formula for the estimator and a control law as in the case of ARX models. Our exposition is as follows. We first prove a recursive formula for updating the estimators and then we propose a control law. We will show that this law is self-optimizing with respect to some criteria. We will also see that this law is self-tuning by using a similar line of argument as in [10]. As a matter of fact, our recursive estimation algorithm possesses the same geometric properties as theirs. This does not come as a surprise when one recalls the definition of generalized linear models. The linear predictor is the one that influences the expected value of the response.

Note that (4.17) can be regarded as a regression model for the adaptive control of Markov processes. It is clear that $\{y_t\}$ is a Markov chain of order p. For a comprehensive account of the adaptive control of Markov processes see [44] and the references cited there. In spite of this fact, we would like to view (4.17) as an extension of the ARX model. It will become clear in what follows, that the cost function we use points to a close connection between (4.17) and the ARX models.

4.4 A Recursive Estimation Procedure

The assumed model is given by eq. (4.17). Denote the estimator of β, up to time t, by $\hat{\beta}_t$. Suppose that a new observation, say (y_{t+1}, z_t) becomes available. How can we update $\hat{\beta}_t$ to $\hat{\beta}_{t+1}$? Note that $\hat{\beta}_t$ does not admit a closed expression as in the case of least squares. However, a recursive formula can be proved, and we give two different approaches to this fact.

4.4.1 Expanding the Partial Score

We know from Chapter 2 that the maximum partial likelihood estimator of β is given as a solution to the following nonlinear equations system (see (2.12)):

$$ps_t(\beta) = \sum_{s=1}^{t} z_{s-1}(y_s - p_s(\beta)).$$

Furthermore, the conditional information matrix is

$$\mathbf{G}_t(\beta) = \sum_{s=1}^{t} z_{s-1} z'_{s-1} p_s(\beta)(1 - p_s(\beta)).$$

Now, following [30] and [85] (for a comprehensive account of recursive estimation see [74]) we have, by noting that $ps_{t+1}(\hat{\beta}_{t+1}) = 0$ (since $\hat{\beta}_{t+1}$ is a root of this equation),

$$ps_{t+1}(\hat{\beta}_{t+1}) = ps_{t+1}(\hat{\beta}_t) - (\hat{\beta}_{t+1} - \hat{\beta}_t)\mathbf{G}_{t+1}(\tilde{\beta}) + o_p(||\hat{\beta}_{t+1} - \hat{\beta}_t||),$$

where $\tilde{\beta}$ lies in the line segment connecting $\hat{\beta}_t$ and $\hat{\beta}_{t+1}$. The above relations lead to

$$0 = z_t(y_{t+1} - p_{t+1}(\hat{\beta}_t)) - (\hat{\beta}_{t+1} - \hat{\beta}_t)\mathbf{G}_{t+1}(\tilde{\beta}) + o_p(||\hat{\beta}_{t+1} - \hat{\beta}_t||),$$

since $ps_{t+1}(\hat{\beta}_t) = ps_t(\hat{\beta}_t) + z_t(y_{t+1} - p_{t+1}(\hat{\beta}_t))$ and $ps_t(\hat{\beta}_t) = 0$. A summary of the above calculations shows that:

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \mathbf{G}_{t+1}^{-1}(\tilde{\beta}) z_t(y_{t+1} - p_{t+1}(\hat{\beta}_t)) + o_p(||\mathbf{G}_{t+1}^{-1}(\tilde{\beta})\hat{\beta}_{t+1} - \hat{\beta}_t||),$$

provided that the above inverse exists. Now, $\tilde{\beta}$ lies between $\hat{\beta}_t$ and $\hat{\beta}_{t+1}$, and we know that the conditional information matrix is a continuous function of β under some assumptions. The above indicates the possibility of replacing $\tilde{\beta}$ by $\hat{\beta}_t$. Note however that this sequence does not necessarily produce the maximum

partial likelihood estimator. Therefore, the recursive algorithm we propose is the following

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \mathbf{G}_{t+1}^{-1}(\hat{\beta}_t) z_t (y_{t+1} - p_{t+1}(\hat{\beta}_t)) \tag{4.18}$$

The above discussion makes evident that this recursive estimation procedure can be generalized to any time series which is modeled by means of generalized linear models with any link function, such that the maximum partial likelihood estimator is unique and the inverse of the conditional information matrix exists. We need to note however that in the Taylor expansion, we used the fact that the link is canonical. In other words the conditional information matrix coincides with the negative matrix of second derivatives of the log-partial likelihood.

4.4.2 Utilizing the Fitting Procedure

The second derivation of (4.18) utilizes the fitting procedure, i.e., the iterative reweighted least squares method as explained in Section 2.4. Suppose that we have a fixed number of observations, say t. In the limit of the Fisher-scoring method, the solution of the partial score equations, $\hat{\beta}_t$, is given approximately by the following least squares problem

$$(\tilde{y}_1, \ldots, \tilde{y}_t)' = (z_0, \ldots, z_{t-1})'\beta + \mathbf{W}_t^{-1} r_t \tag{4.19}$$

in the notation of Section 2.4, with

$$\mathbf{W}_t = \mathrm{diag}(p_s(\beta)(1 - p_s(\beta)))_{s=1}^t$$

and

$$r_t = (y_1 - p_1, \ldots, y_t - p_t)'.$$

Observe that $E[\mathbf{W}_t^{-1} r_t \mid\mid \mathcal{F}_{t-1}] = 0$ and $Var[\mathbf{W}_t^{-1} r_t \mid\mid \mathcal{F}_t] = \mathbf{W}_t^{-1}$ under the true model. Therefore, since this is a weighted least squares problem, the recursive

formulas (4.8) and (4.9) can be applied by using $v_t = 1/(p_{t+1}(\hat{\beta}_t)(1 - p_{t+1}(\hat{\beta}_t)))$. The last quantity really depends only on t, since $p_{t+1}(\hat{\beta}_t) = 1/(1 + \exp(-\hat{\beta}_t z_t))$. Then we have

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \mathbf{G}_{t+1}^{-1}(\hat{\beta}_t) z_t (y_{t+1} - p_{t+1}(\hat{\beta}_t))$$

upon noting that $v_t z_t(\tilde{y}_{t+1} - z_t\hat{\beta}_t) = z_t(y_{t+1} - p_{t+1}(\hat{\beta}_t))$ by employing (4.19). This result is in accordance with (4.18). Furthermore

$$\mathbf{G}_{t+1}(\hat{\beta}_t) = \mathbf{G}_t(\hat{\beta}_t) + z_t z_t' v_t^{-1}. \tag{4.20}$$

Now, if we use (4.10) and (4.11), we get that

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \frac{\mathbf{D}_t(\hat{\beta}_t) z_t}{1 + z_t' \mathbf{D}_t(\hat{\beta}_t) z_t v_{t+1}^{-1}(\hat{\beta}_t)} (y_{t+1} - p_{t+1}(\hat{\beta}_t)), \tag{4.21}$$

$$\mathbf{D}_{t+1}(\hat{\beta}_t) = \mathbf{D}_t(\hat{\beta}_t) - \frac{\mathbf{D}_t(\hat{\beta}_t) z_t z_t' \mathbf{D}_t(\hat{\beta}_t)}{v_{t+1}^{-1}(\hat{\beta}_t) + z_t \mathbf{D}_t(\hat{\beta}_t) z_t'}, \tag{4.22}$$

with $\mathbf{D}_t(\hat{\beta}_t) = \mathbf{G}_t^{-1}(\hat{\beta}_t)$. The result is in line with [98], the authors of which used a very similar argument for the derivation of (4.21)-(4.22). The last argument just takes advantage of the fitting procedure. Namely, at each time t, the corresponding estimator at the limit of the iterations is a solution to a weighted least squares problem. Notice, however, that the weights are determined by the fitting procedure.

Now, we have all the necessary ingredients to propose and study a control law with respect to some criteria.

4.5 Self-Optimality

Having developed the recursive estimation scheme , one would like to know how to control a system such as (4.16). Although there is no available corresponding

minimum variance control theory for such systems, it would be sensible to choose the control law such that the right hand side of (4.16) becomes zero. In other words we would be able to regulate such a system if we choose

$$u_t = -\frac{1}{b_0}[\sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=1}^{p} b_i u_{t-i}] \tag{4.23}$$

in the case that the coefficients are known. Equivalently, this can be written as

$$u_t = -\frac{A(z)}{B(z)} y_t \tag{4.24}$$

with $A(z) = \sum_{i=0}^{p} a_i z^i$, $B(z) = \sum_{i=0}^{p} b_i z_i$ and z denotes the lag operator. Even though that y_t is binary, we define (4.24) by means of (4.23). It is clear again, from the discussion of the linear systems, that we need to assume that $B(z)$ has all its roots outside the unit circle. That is, our system satisfies the minimum phase condition.

Now, by acting in the same way as in the linear case, we propose the following stochastic approximation type algorithm ([82]):

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \gamma \frac{z_t}{r_t}(y_{t+1} - p_{t+1}(\hat{\beta}_t)); \tag{4.25}$$

$$r_t = r_{t-1} + z_t' z_t p_{t+1}(\hat{\beta}_t)(1 - p_{t+1}(\hat{\beta}_t)) \ ; r(0) = 1 \tag{4.26}$$

where γ is a constant greater than or equal to 1. Based on the parameter estimate $\hat{\beta}_t$ available at time t, the control input u_t is chosen to satisfy $z_t'\hat{\beta}_t = 0$. In other words,

$$u_t = -\frac{1}{\hat{b}_0(t)}[\sum_{i=0}^{p} \hat{a}_i(t) y_{t-i} + \sum_{i=1}^{p} \hat{b}_i(t) u_{t-i}]. \tag{4.27}$$

A natural criterion for the performance of this system would be the following cost function

$$C_N = \frac{\sum_{t=1}^{N} y_t}{N}.$$

If C_N is close to $1/2$, then the self-optimality would hold. Indeed, if the system was completely known and (4.23) had been used, then $p_t = 1/2$ for all t. This means that we have a homogeneous stationary Markov chain, so

$$C_N \to 1/2 \text{ a.s.}$$

However, this is not a usual cost function in the sense that it doesn't depend on previous observations, control inputs or parameters. An appealing measure of performance would be

$$D_N = \frac{\sum_{t=1}^{N} \lambda_t^2}{N}$$

which depends on the unknown parameters, past observations and control actions. Convergence of D_N to 0 indicates again self-optimality. The last criterion can be interpreted as "variance output" in line with the linear models. We will examine both measures of performance in the following theorem. First, we need to state some assumptions.

C.1 The polynomial $B(z)$ has all its roots strictly outside the unit circle.

C.2 The true probability measure which governs $\{y_t, z_t\}$ obeys (4.16) with the true parameter being β.

C.3 The random variables z_t belong to a non-random compact subset of R^{2p+2}.

C.4 There exists a probability measure μ such that under the true parameter and for all Borel subsets A of R^{2p+2},

$$\frac{\sum_{t=1}^{N} I_{[z_t \in A]}}{N} \to \mu(A) \text{ a.s.}$$

Theorem 4.5.1 Consider the system (4.16). Suppose that assumptions C.1–C.4 hold. If the stochastic gradient based adaptive control law (4.25)-(4.26)-(4.27) is used then we have:

1. the adaptive control law is self-optimizing with respect to both criteria C_N and D_N;

2. $||\hat{\beta}_t - \beta||$ converges a.s.

Proof: The proof will be based on the following quadratic form

$$V_t = ||\hat{\beta}_t - \beta||^2$$

which is also known as stochastic Lyapounov function. We will first calculate $E[V_{t+1} \mid\mid \mathcal{F}_t]$. Notice that

$$\hat{\beta}_{t+1} - \beta = (\hat{\beta}_t - \beta) + \gamma \frac{z_t}{r_t}(y_{t+1} - p_{t+1}(\hat{\beta}_t))$$

Since $z_t'\hat{\beta}_t = 0$, we have that

$$\hat{\beta}_{t+1} - \beta = (\hat{\beta}_t - \beta) + \gamma \frac{z_t}{r_t}(y_{t+1} - \frac{1}{2})$$

We now square both sides and take conditional expectations given $\mathcal{F}_t$. Therefore, we get

$$\begin{aligned} E[V_{t+1} \mid\mid \mathcal{F}_t] &= V_t + 2\gamma \tfrac{z_t'(\hat{\beta}_t - \beta)}{r_t} E[(y_{t+1} - \tfrac{1}{2}) \mid\mid \mathcal{F}_t] \\ &\quad + \gamma^2 \tfrac{||z_t||^2}{r_t^2} E[(y_{t+1} - \tfrac{1}{2})^2 \mid\mid \mathcal{F}_t]. \end{aligned}$$

But $E[(y_{t+1} - 1/2) \mid\mid \mathcal{F}_t] = p_{t+1} - 1/2$ and $E[(y_{t+1} - 1/2)^2 \mid\mid \mathcal{F}_t] = 1/4$. It is also helpful to note that $z_t'(\hat{\beta}_t - \beta) = -z_t'\beta = -\lambda_{t+1}$. Thus

$$E[V_{t+1} \mid\mid \mathcal{F}_t] = V_t - 2\gamma \frac{\lambda_{t+1}}{r_t}(p_{t+1} - \frac{1}{2}) + \gamma^2 \frac{||z_t||^2}{4r_t^2}$$

Define now the function,

$$f(x) = x(\frac{1}{1 + \exp(-x)} - \frac{1}{2}).$$

This is a continuous function which is positive except at the point $x = 0$. At this point, $f(0) = 0$. So, we can write now, in a somewhat more compact form

$$E[V_{t+1} \mid\mid \mathcal{F}_t] \;\leq\; V_t - \frac{2\gamma}{r_t} f(z_t'\beta) + \gamma^2 \frac{||z_t||^2}{4r_t^2}$$

Now, observe that $f(z_t'\beta)/r_t \geq 0$, and $||z_t||^2/4r_t^2 \geq 0$ as well. Furthermore, using assumption C.3,

$$\sum_t \frac{||z_t||^2}{4r_t^2} \leq M \sum_t \frac{1}{(t+1)^2} < \infty \text{ a.s.}$$

It follows from Theorem 4.2.3 that

I.

$$\sum_t \frac{f(z_t'\beta)}{r_t} < \infty \text{ a.s.}$$

II.

$$||\hat{\beta}_t - \beta||^2 < \infty \text{ converges a.s.}$$

Statement (II) above proves the second assertion of the theorem. Now, using (I) and Kronecker's lemma, we have that

$$\frac{1}{r_N} \sum_{t=1}^{N} f(z_t'\beta) \to 0 \text{ a.s.}$$

since $r_t > 0$ and $\lim_{t\to\infty} r_t = \infty$. However, if we observe the fact that

$$\frac{r_N}{N} \leq \sum_{t=1}^{N} \frac{||z_t||^2}{N} < M \text{ a.s.,}$$

we get that

$$\frac{1}{N} \sum_{t=1}^{N} f(z_t'\beta) \to 0 \text{ a.s.}$$

On the other hand, by assumption C.4

$$\frac{1}{N} \sum_{t=1}^{N} f(z_t'\beta) \to \int f(z'\beta)\mu(dz)$$

since the function f is continuous on a bounded set. Now, from the uniqueness of the limit we conclude that $\int f(z'\beta)\mu(dz) = 0$. Thus, by the properties of f it follows that

$$z'\beta = 0 \text{ a.s. } \mu. \tag{4.28}$$

Therefore we get that

$$D_N = \frac{1}{N}\sum_{t=1}^{N}\lambda_t^2 \to \int (z'\beta)^2\mu(dz) = 0 \text{ a.s.}$$

Therefore, asymptotic optimality of the proposed control law with respect to D_N is established. Now, we examine the behavior of C_N. Upon noting that $y_t - p_t$ is a uniformly bounded martingale difference sequence with finite second moment, we have that

$$\sum_t \frac{E[(y_t - p_t)^2 \mid\mid \mathcal{F}_{t-1}]}{t^2} < \infty \text{ a.s.}$$

By the martingale stability theorem ([95, theorem 3.3.1]), we can conclude therefore that

$$\frac{\sum_{t=1}^{N}(y_t - p_t)}{N} \to 0 \text{ a.s.}$$

It follows that

$$\begin{aligned} C_N &= \frac{\sum_{t=1}^{N}(y_t - p_t)}{N} + \frac{\sum_{t=1}^{N} p_t}{N} \\ &\to 0 + \int \frac{1}{1+\exp(-z'\beta)}\mu(dz) \\ &= 1/2 \end{aligned}$$

as was supposed to be proved. The theorem therefore follows. □

Remark 4.5.1 Another possible measure of performance is the following

$$\tilde{D}_N = \frac{1}{N}\sum_{t=1}^{N}(p_t - \frac{1}{2})^2.$$

Then if the above quantity converges at 0 as $N \to \infty$, the control law would be self-optimal with respect to this criterion too. But this is the case since

$$\tilde{D}_N \to \int (\frac{1}{1+\exp(-z'\beta)} - \frac{1}{2})^2 \mu(dz) = 0.$$

Remark 4.5.2 The relation (4.28) is a strong conclusion which follows from the hypotheses C.3 and C.4. These are very strong assumptions but we use them here as a first approximation to the problem at hand. It is an interesting problem to consider relaxation of these conditions.

4.6 Geometric Properties of the Proposed Algorithm

We now investigate the geometric properties of the proposed algorithm (4.25)-(4.26)-(4.27). These properties are crucial for the proof of self-tuning. We first give the following proposition.

Proposition 4.6.1 Consider the recursions (4.25)-(4.26)-(4.27). Then $\hat{\beta}_{t+1} - \hat{\beta}_t$ is orthogonal to $\hat{\beta}_t$.

Proof: Since

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \gamma \frac{z_t}{r_t}(y_{t+1} - .5)$$

due to the relation $z_t'\hat{\beta}_t = 0$, we have that $\hat{\beta}_{t+1} - \hat{\beta}_t$ is parallel to z_t. However, $z_t'\hat{\beta}_t = 0$ implies that $\hat{\beta}_t$ is orthogonal to z_t. Therefore the conclusion of the proposition holds true. □

We must point out here that the last proposition is a property of the algorithm and has no relation with the system being controlled. We state now another proposition which is a direct consequence of Proposition 4.6.1.

Proposition 4.6.2 Consider again the recursions (4.25)-(4.26)-(4.27). Then the following are true:

1. $||\hat{\beta}_t||^2 = ||\beta_0||^2 + \sum_{s=1}^{t} ||\hat{\beta}_s - \hat{\beta}_{s-1}||^2$;

2. $||\hat{\beta}_{t+1}||^2 \geq ||\hat{\beta}_t||^2$ for $t \geq 0$;

3. If $\sup_t ||\hat{\beta}_t|| < \infty$, then $\{||\hat{\beta}_t||\}$ converges a.s.;

4. If $\sup_t ||\hat{\beta}_t|| < \infty$, then $||\hat{\beta}_{t+1} - \hat{\beta}_t|| \to 0$ a.s. as $t \to \infty$;

5. If $||\hat{\beta}_t - \beta||$ converges, so does $\{||\hat{\beta}_t||\}$.

Proof:

1. By Pythagoras's theorem we get that

$$||\hat{\beta}_t||^2 = ||\hat{\beta}_{t-1}||^2 + ||\hat{\beta}_t - \hat{\beta}_{t-1}||^2 \quad \text{for } t \geq 1$$

   The assertion follows by summation.

2. This is a consequence of (1). In other words the sequence of estimators is nondecreasing.

3. Since $\sup_t ||\hat{\beta}_t|| < \infty$ and $\{||\hat{\beta}_t||\}$ is a nondecreasing sequence, it follows that $\hat{\beta}_t$ converges.

4. If the sequence of estimators is bounded, then necessarily the summation in (1) converges and therefore the desired result follows.

5. If $||\hat{\beta}_t - \beta||$ converges, then $\sup_t ||\hat{\beta}_t|| < \infty$ and the result follows from (2). □

The last item of proposition 4.6.2 and the second assertion of proposition 4.6.1 give the following lemma:

Lemma 4.6.1 $||\hat{\beta}_t||$ and $||\hat{\beta}_t - \beta||$ both converge.

The above lemma shows that $\hat{\beta}_t$ converges to the intersection of two random spheres. The one sphere is centered at the origin and the other at β. The intersection of two spheres is a hypersphere, say H, of strictly smaller dimension. If we want to show that H consists of only one point, the we need to show that the two spheres are tangential. But if this is the case, then the point of intersection belongs to the straight line L which passes through 0 and β. Now H consists of only one point if and only if $H \cap L \neq \emptyset$. To show that H and L have only one point in common , it suffices to show that there exists a subsequence t_l such that

$$\lim_{l \to \infty} \hat{\beta}_{t_l} = k\beta$$

since we know that all the limit points belong to H. For the sake of completeness, we give the following proposition (see [56]), which makes the argument precise.

Proposition 4.6.3 Let $\beta(i)$ and $\hat{\beta}_t(i)$ denote the ith component of the vectors β and $\hat{\beta}_t$, respectively, for $i = 1, \ldots, 2p+2$. Then the following statements are equivalent:

1. There exists a subsequence t_l such that

$$\lim_{l \to \infty} \hat{\beta}_{t_l} = k\beta;$$

2. There exists a subsequence t_l such that

$$\lim_{l \to \infty} (\beta(i)\hat{\beta}_{t_l}(p+2) - \beta(p+2)\hat{\beta}_{t_l}(i)) = 0 \text{ for } i = 1, \ldots, 2p+2; \quad (4.29)$$

3.

$$\lim_{t\to\infty} \hat{\beta}_t = k\beta.$$

Proof: It is clear that (1) implies (2) and (3) implies (1). We only need to verify that (2) implies (3). To this end, noticing that $\beta(p+2) = b_0 \neq 0$, (2) implies

$$\lim_{l\to\infty} \hat{\beta}_{t_l} = \frac{\bar{\beta}(p+2)}{\beta(p+2)}\beta(i) \text{ for } i = 1, \ldots, 2p+2,$$

where $\bar{\beta}(p+2)$ is the $(p+2)^{\text{th}}$ component of $\bar{\beta} = \lim_{l\to\infty} \hat{\beta}_{t_l}$. This limit exists, from Lemma 4.6.1. The above relation can be written as

$$\lim_{l\to\infty} \hat{\beta}_{t_l} = k\beta \tag{4.30}$$

with $k = \bar{\beta}(p+2)/\beta(p+2)$. Consider now the quadratic form

$$||\hat{\beta}_t - k\beta||.$$

We have, by properties of the inner product, that

$$||\hat{\beta}_t - k\beta|| = ||\hat{\beta}_t||^2 - 2k < \beta, \hat{\beta}_t > +k^2||\beta||^2.$$

From the convergence of $||\hat{\beta}_t - \beta||$, we conclude that $< \hat{\beta}_t, \beta >$ converges. It follows from (4.30) that

$$\lim_{t\to\infty} ||\hat{\beta}_t - k\beta||^2 = 0 \text{ a.s.}$$

proving the proposition. □

4.7 Self-tuning of the Control Law

We now prove that the proposed control law is self-tuning. The proof will be based on the following three lemmas. The first and the third can be found in [10]. We quote their proof for the sake of completeness.

Lemma 4.7.1 Suppose that s_t is a real-valued sequence that satisfies

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} s_t^2 = 0.$$

Then for any real valued sequence q_t,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} q_t^2 = 0 \Longleftrightarrow \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N}(s_t + q_t)^2 = 0.$$

Proof: Assume first that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} s_t^2 = 0.$$

Then, we have that

$$\frac{1}{N}\sum_{t=1}^{N}(s_t + q_t)^2 \leq \frac{2}{N}\sum_{t=1}^{N} s_t^2 + \frac{2}{N}\sum_{t=1}^{N} q_t^2,$$

and this proves the if part of the lemma. Now, assume that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N}(s_t + q_t)^2 = 0,$$

Using the inequality

$$q_t^2 \leq 2s_t^2 + 2(q_t + s_t)^2,$$

the conclusion of the lemma follows. □

Lemma 4.7.2 Assume that $\{\eta_t\}$ and $\{s_t\}$ are stochastic processes adapted to $\mathcal{F}_t$, such that $\{\eta_t\}$ is bounded. Let $\{y_t\}$ be a binary stochastic process such that $E[y_t \mid\mid \mathcal{F}_{t-1}] = p_t$. Furthermore, assume that

$$\lim_{N\to\infty} \sum_{t=1}^{N} \frac{1}{N}(\eta_{t-1}y_t + s_{t-1})^2 = 0 \text{ a.s.} \tag{4.31}$$

Then

$$\lim_{N\to\infty} \sum_{t=1}^{N} \frac{1}{N}\eta_t^2 = 0 \text{ a.s.}$$

and

$$\lim_{N\to\infty}\sum_{t=1}^{N}\frac{1}{N}s_t^2 = 0 \text{ a.s.}$$

Proof: Since $E[y_t \mid\mid \mathcal{F}_{t-1}] = p_t$, it follows that $E[y_t - p_t \mid\mid \mathcal{F}_{t-1}] = 0$ a.s. and therefore the sequence $\{y_t - p_t\}$ is a martingale difference. Relation (4.31) can be written as

$$\frac{1}{N}\sum_{t=1}^{N}(\eta_{t-1}(y_t - p_t) + \eta_{t-1}p_t + s_{t-1})^2 = \frac{1}{N}\sum_{t=1}^{N}(\eta_{t-1}w_t + v_{t-1})^2$$

with $w_t = y_t - p_t$ and $v_{t-1} = \eta_{t-1}p_t + s_{t-1}$. The last equation implies that

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^{N}(\eta_{t-1}w_t + v_{t-1})^2 &= \frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}^2 w_t^2 + \frac{1}{N}\sum_{t=1}^{N}v_{t-1}^2 \\ &+ \frac{2}{N}\sum_{t=1}^{N}\eta_{t-1}v_{t-1}w_t. \end{aligned} \tag{4.32}$$

Using [57, lemma 2.iii], we have that $\sum_{t=1}^{N}\eta_{t-1}v_{t-1}w_t$ converges a.s. on $\Omega = \{\omega : \sum_{t=1}^{\infty}\eta_{t-1}^2 v_{t-1}^2 < \infty\}$. Therefore, we have that

$$\frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}v_{t-1}w_t \to 0 \text{ a.s.}$$

on Ω. Thus, we conclude from (4.32) and (4.31) that

$$\frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}^2 w_t^2 \to 0 \text{ a.s.}$$

and

$$\frac{1}{N}\sum_{t=1}^{N}v_{t-1}^2 \to 0 \text{ a.s.}$$

on Ω. Now, on $\Omega^c = \{\omega : \sum_{t=1}^{\infty}\eta_{t-1}^2 v_{t-1}^2 = \infty\}$ we have again from the same lemma that

$$\sum_{t=1}^{N}\eta_{t-1}v_{t-1}w_t = o(\sum_{t=1}^{N}\eta_{t-1}^2 v_{t-1}^2) = o(\sum_{t=1}^{N}v_{t-1}^2).$$

The last inequality follows from the fact that $\{\eta_t\}$ is bounded. Thus, from (4.32) we have

$$\frac{1}{N}\sum_{t=1}^{N}(\eta_{t-1}w_t + v_{t-1})^2 = \frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}^2 w_t^2 + \frac{1}{N}\sum_{t=1}^{N} v_{t-1}^2 \left[1 + \frac{o(\sum_{t=1}^{N} v_{t-1}^2)}{\sum_{t=1}^{N} v_{t-1}^2}\right]$$

a.s. on Ω^c. Because of (4.31) we once again have that

$$\frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}^2 w_t^2 \to 0 \text{ a.s.}$$

and

$$\frac{1}{N}\sum_{t=1}^{N} v_{t-1}^2 \to 0 \text{ a.s.}$$

on Ω^c. Hence, we obtain that

$$\frac{1}{N}\sum_{t=1}^{N}(\eta_{t-1}p_t + s_t)^2 \to 0 \text{ a.s.}$$

and

$$\frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}^2(y_t - p_t)^2 \to 0 \text{ a.s.}$$

Now, note that

$$\sup_t E[|\eta_{t-1}^2(y_t - p_t)^2 - E[\eta_{t-1}^2(y_t - p_t)^2 \mid\mid \mathcal{F}_{t-1}]|^{1+\delta/2} \mid\mid \mathcal{F}_{t-1}] < \infty \text{ a.s.}$$

since $\{\eta_t\}$ is bounded. Hence from [95, theorem 3.3.1], it follows that

$$\frac{1}{N}\sum_{t=1}^{N}\eta_t^2 \to 0 \text{ a.s.}$$

proving the first part of the lemma. Upon noting again that

$$\begin{aligned} s_t^2 &\leq 2\eta_{t-1}^2 p_t^2 + 2(\eta_{t-1}p_t + s_t)^2 \\ &\leq 2\eta_{t-1}^2 + 2(\eta_{t-1}p_t + s_t)^2, \end{aligned}$$

we conclude also that the second assertion of the lemma holds true. □.

Lemma 4.7.3 Assume that $\{x_t\}$, $\{u_t\}$ and $\{s_t\}$ are real valued sequences satisfying

I. $\limsup_{N\to\infty}(1/N)\sum_{t=1}^{N} u_t^2 < \infty$,

II. $\lim_{t\to\infty}(x_t - x_{t-1}) = 0$,

III. $\lim_{N\to\infty}(1/N)\sum_{t=1}^{N}(x_t u_t + s_t)^2 = 0$.

Then, we have that

$$\frac{1}{N}\sum_{t=1}^{N}(x_{t-l}u_t + s_t)^2 = 0 \quad \text{for every } l \geq 1.$$

Proof: From (II), we get that for all $\epsilon > 0$, $|x_t - x_{t-l}| < \epsilon$ for all $t > t_0$. Now, for $N > t_0$,

$$\frac{1}{N}\sum_{t=1}^{N}[(x_{t-l} - x_t)u_t]^2 \leq \frac{1}{N}\sum_{t=1}^{t_0}(x_{t-l} - x_t)u_t]^2 + \frac{\epsilon}{N}\sum_{t=t_0+1}^{N} u_t^2.$$

Letting $N \to \infty$, since $\epsilon > 0$

$$\frac{1}{N}\sum_{t=1}^{N}[(x_{t-l} - x_t)u_t]^2 \to 0.$$

From lemma 4.7.1 the desired result follows. □

We summarize our results in the following theorems. The proofs follow the lines of [10], [56]. We will not repeat the arguments here. However their results are quite applicable ([10, Theorem 22]).

Theorem 4.7.1 Asuume that C1-C4 hold. Then the stochastic adaptive control law (4.25)-(4.26)-(4.27) is self-tunning in the Cesaro sense for the system (4.17). In other words,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N} I_{[\hat{\beta}_t \in O]} = 1 \text{ a.s.}$$

for every open set O contained in $\mathcal{B}$ (see Definition 4.2.4 and 4.2.5).

To prove self-tuning of the proposed control law, Lemmas 4.7.1, 4.7.2, 4.7.3 and the fact that $D_N \to 0$ a.s. play a crucial role. By combining these facts, the main idea is to show that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N}(\beta_i\hat{\beta}_t(p+2) - \beta(p+2)\hat{\beta}_t(i)) = 0 \quad \text{a.s.}$$

for $i = 1, \ldots, 2p+2$, verifying (2) of Proposition 4.6.3. In summary, we get the following.

Theorem 4.7.2 Consider the system (4.17). Suppose that the system has no reduced minimum variance controllers [(see (4.2.6)] and assume C1-C4. If the control law (4.25)-(4.26)-(4.27) is used then

1. the parameter estimates $\{\hat{\beta}_t\}$ converge to a random multiple of β, vis.

$$\lim_{t\to\infty} \hat{\beta}_t = k\beta$$

   where

$$k^2 = \frac{||\hat{\beta}(0)||^2}{||\beta||^2} + \frac{\gamma^2}{4||\beta||^2}\sum_{t=1}^{\infty}\frac{||z_{t-1}||^2}{r_{t-1}^2}$$

2. the adaptive control law is self-tuning:

$$\lim_{t\to\infty}\left(u_t - \frac{A(z)}{B(z)}y_t\right) = 0.$$

For the calculation of the random constant, note from (1) of Proposition (4.6.2) that as $t \to \infty$ we have

$$k^2||\beta||^2 = ||\hat{\beta}(0)||^2 + \frac{\gamma^2}{4}\sum_{t=1}^{\infty}\frac{||z_{t-1}||^2}{r_{t-1}^2}$$

since $\{y_t\}$ is taking only the values 0 or 1.

4.8 Simulations

We now illustrate some important points by simulations. We first generated a time series of length equal to 850 according to the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = \sin(t) + \cos(t)$. We ran 15 simulations and then we averaged the results. The first two hundred observations gave a preliminary estimator using the method of partial likelihood. We found $k = 1.1209$, $D_N = 0.0590$, $C_N = 0.4958$ and $\tilde{D}_N = 0.0033$.

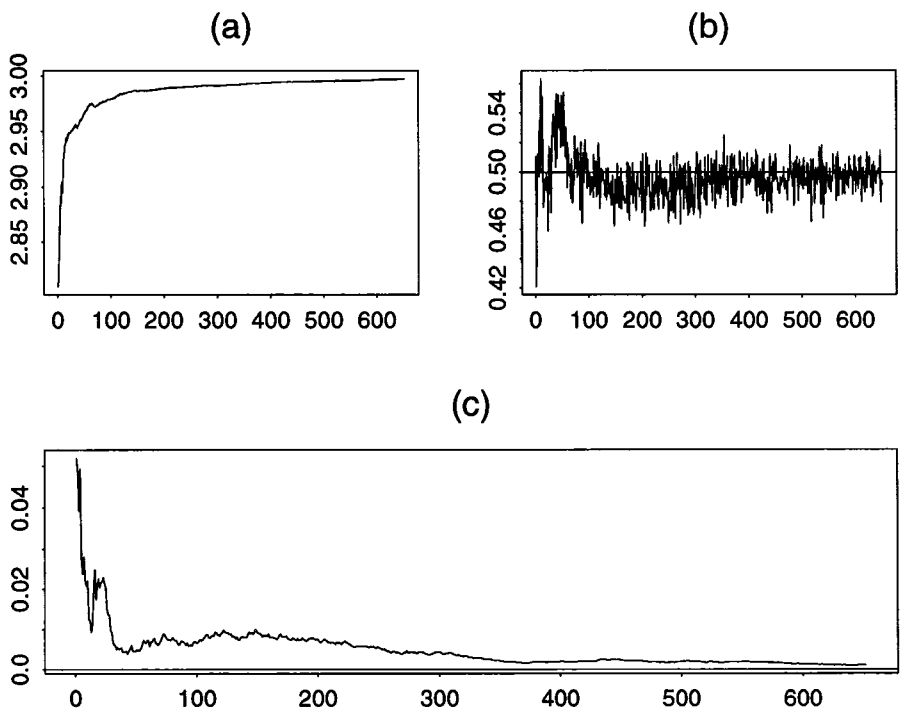

Figure 4.1: (a) Norm of the estimators (b) Controlled probabilities around 1/2 (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = \sin(t) + \cos(t)$.

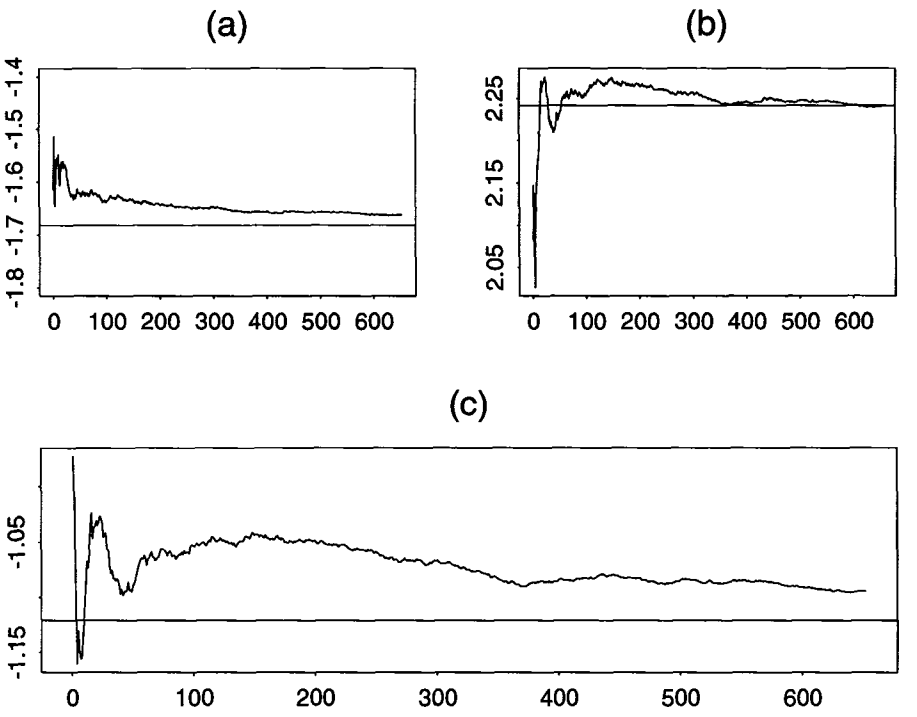

Figure 4.2: (a) Iterations for β_1 (b) Iterations for β_2 (c) Iterations for β_3 for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = \sin(t) + \cos(t)$.

Figure 4.1 illustrates (a) the norm of the estimators, (b) the transition probabilities of the system and (c) the norm of the difference i.e. $||\hat{\beta}_t - k\beta||^2$. We see that the norm of estimators seems to be an increasing and bounded sequence. The transition probabilities fluctuate around 0.5 and the norm of the difference converges to 0. An unfortunate characteristic is the slow rate of convergence being evident in Figure 4.1(c). Although the initial partial likelihood estimator used to start the recursions is satisfactory, we see that the rate of convergence is very slow.

Figure 4.2 demonstrates the recursions for the estimators. The iterations for the first two estimators are quite satisfactory. However, the iterations for

the third estimator do not exhibit the expected result. Note however that the difference between recursion and the horizontal line is of the order of 10^{-2}. We did not draw all the graphs in the same horizontal scale. However, we can see that there is some early oscillation and the recursions seem to converge after the first two to three hundred observations. We fitted the same model using the same number of data points. We now generate u_t according to a first-order autoregressive process with parameter 0.3. That is, $u_t = .3u_{t-1} + e_t$. Now $k = 1.0897$, $D_N = 0.0521$, $C_N = 0.4973$ and $\tilde{D}_N = 0.0029$.

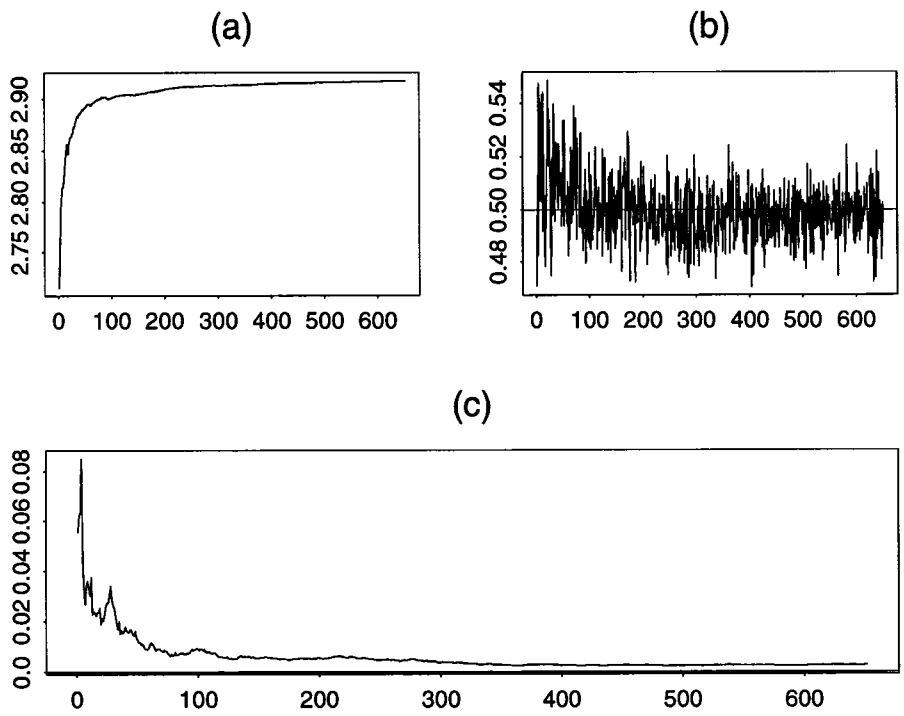

Figure 4.3: (a) Norm of the estimators (b) Controlled probabilities around 1/2. (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = .3u_t + e_t$.

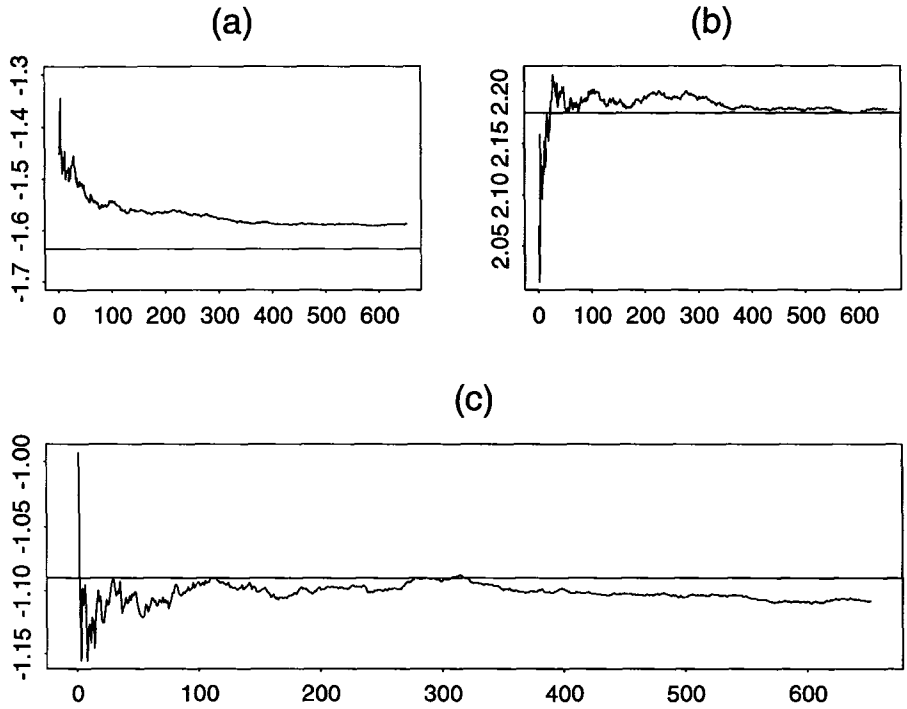

Figure 4.4: (a) Iterations for β_1 (b) Iterations for β_2 (c) Iterations for β_3 for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = .3u_t + e_t$.

We again use the partial likelihood estimator as a starting value for the recursion based on the first two hundred observations. Figures 4.3 and 4.4 correspond to Figures 4.1 and 4.2. The same conclusions as in the previous case can be drawn.

By changing the starting values for the estimators to $(-2, 3.1, 2)$ and still using the same model we got $k = 1.4360$, $D_N = 0.0981$, $C_N = 0.4566$ and $\tilde{D}_N = 0.0070$.

Figure 4.5 demonstrates the same results as Figure 4.1. Note that Figure 4.5(c) shows that the norm of the difference converges toward zero but again in a very slow rate (the horizontal line indicates 0). A more striking point emerges from Figure 4.6 which illustrates the recursions for the estimators. We see that

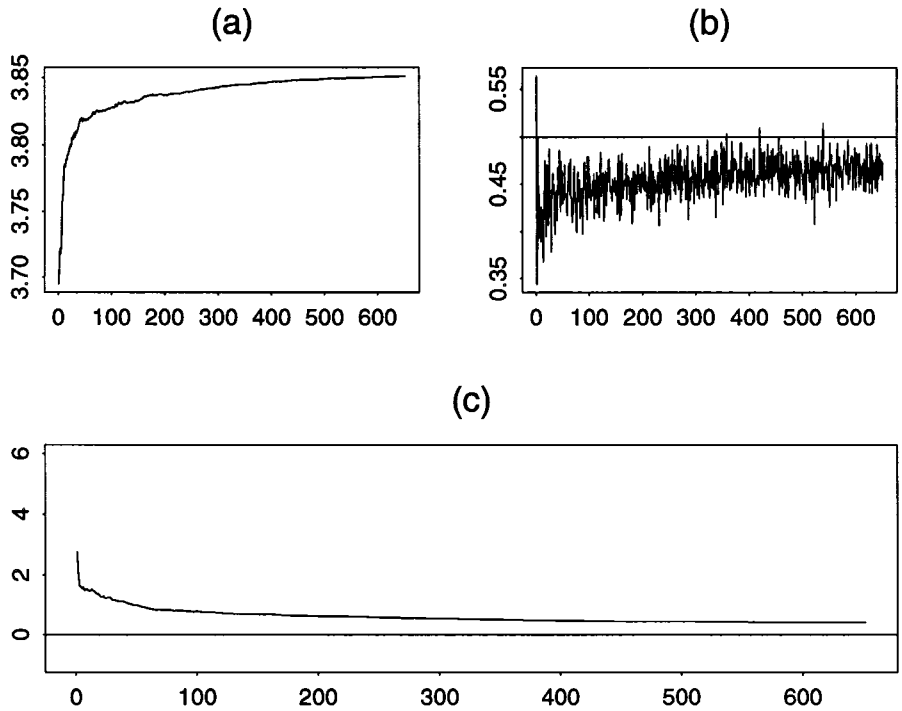

Figure 4.5: (a) Norm of the estimators (b) Controlled probabilities around 1/2 (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with starting values (-2,3.1,2).

the recursions for the first and the third estimator are far from convergence. Thus, if we want to control such a system it would be better to obtain a preliminary estimator, like the maximum partial likelihood estimator, and then use this as a starting value.

We next fitted now the model $\lambda_{t+1} = y_t - .5y_{t-1} - 2u_t + u_{t-1}$, with u_t being generated by a first-order autoregressive process with parameter 0.3. There we used a time series of length 1850. The first two hundred observations were used to compute a preliminary estimate by the method of partial likelihood. We found $k = 1.1718$, $D_N = 0.0846$, $C_N = 0.4810$ and $\tilde{D}_N = 0.0041$.

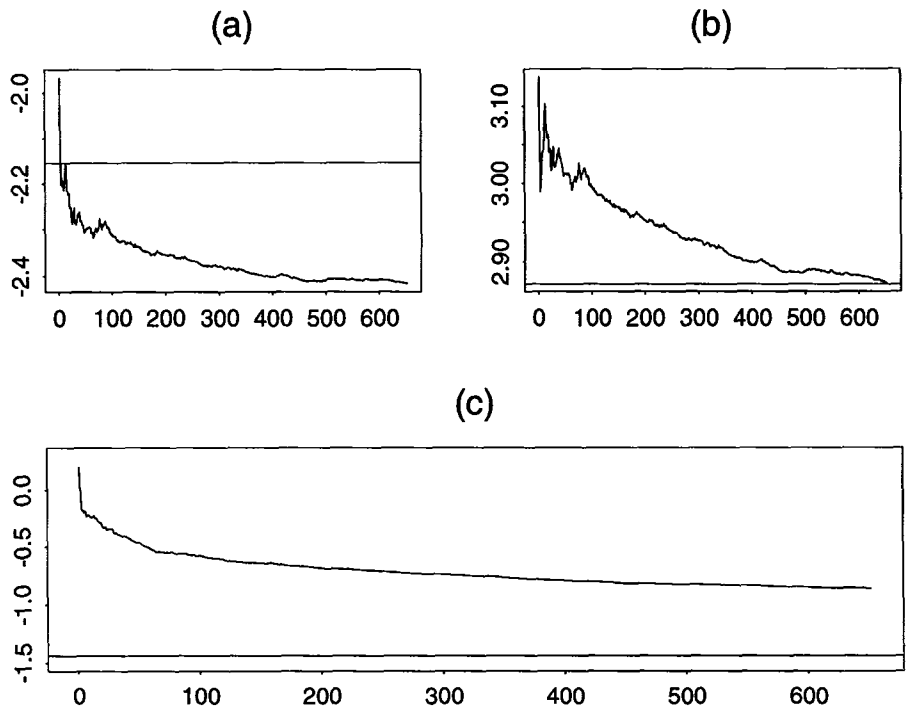

Figure 4.6: (a) Iterations for β_1 (b) Iterations for β_2 (c) Iterations for β_3 for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with starting values (-2,3.1,2).

Figure 4.7 exhibits the same phenomena as before. Figure 4.8 shows that the parameter estimates do not quite converge, because the algorithm, as we have already mentioned, does not posses a satisfactory rate of convergence. In addition the number of parameters have been increased.

The next model we fitted was $\lambda_{t+1} = -1.2y_t - 1.32u_t + .1u_{t-1} + u_{t-2}$. We used 1050 data of points and the control was generated again by a first-order autoregressive process with parameter 0.3. Our results are $k = 1.2627$, $D_N = 0.0647$, $C_N = 0.4881$ and $\tilde{D}_N = 0.0084$. Figures 4.9 anf 4.10 illustrate the results.

Figures 4.11 and 4.12 display data from the model first considered, namely, $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$. But now we would like to control the probabilities

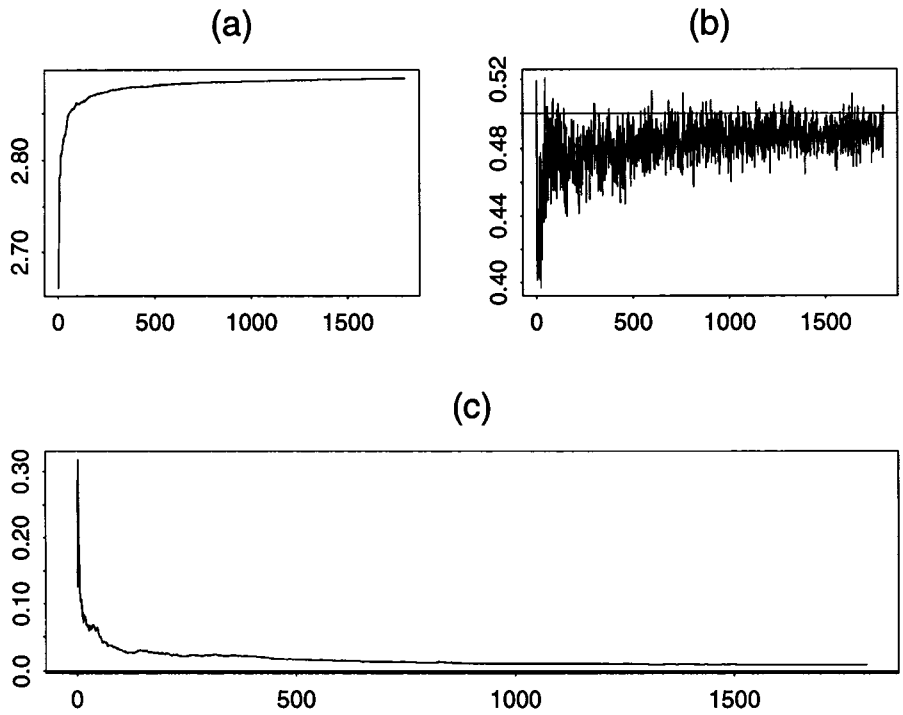

Figure 4.7: (a) Norm of the estimators (b) Controlled probabilities around 1/2 (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = y_t - .5y_{t-1} - 2u_t + u_{t-1}$.

around .2 (see Figure 4.11(b)). Suppose that we want to control the probabilities around a known number r, $0 < r < 1$. Then the control law should be modified to

$$u_t = \frac{1}{\hat{b}_0(t)}[c - \sum_{i=0}^{p} \hat{a}_i(t)y_{t-i} - \sum_{i=1}^{p} \hat{b}_i(t)u_{t-i}]$$

with

$$c = \log(\frac{r}{1-r})$$

Note, however, that this law does not preserve the orthogonality property of the proposed algorithm. Hence some adjustments must be made. Our idea, in this particular example was not to model the response probabilities but rather the

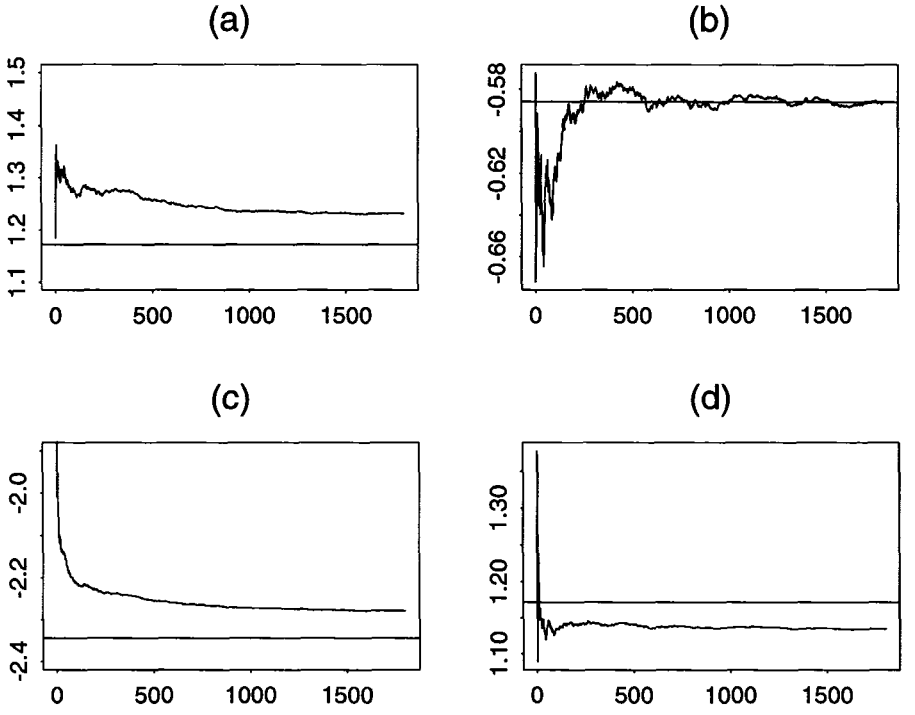

Figure 4.8: (a) Iterations for β_1 (b) Iterations for β_2 (c) Iterations for β_3 (d) Iterations for β_4 for the model $\lambda_{t+1} = y_t - .5y_{t-1} - 2u_t + u_{t-1}$.

log-odds ratio. In other words we can consider the following model

$$\log\left[\frac{p_{t+1}(1-r)}{(1-p_{t+1})r}\right] = \sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=0}^{p} b_i u_{t-i}$$

and then use the proposed control law. The cost functions must modified to

$$D_N = \frac{1}{N}\sum_{t=1}^{N}(\lambda_t - \log\frac{r}{1-r})^2$$

and

$$\tilde{D}_N = \frac{1}{N}\sum_{t=1}^{N}(p_t - r)^2$$

With this discussion in mind, we have then that $k = 1.0239$, $D_N = 0.0264$, $C_N = 0.1971$ and $\tilde{D}_N = 0.0007$. Figure (4.12) gives the iterations for the estimators.

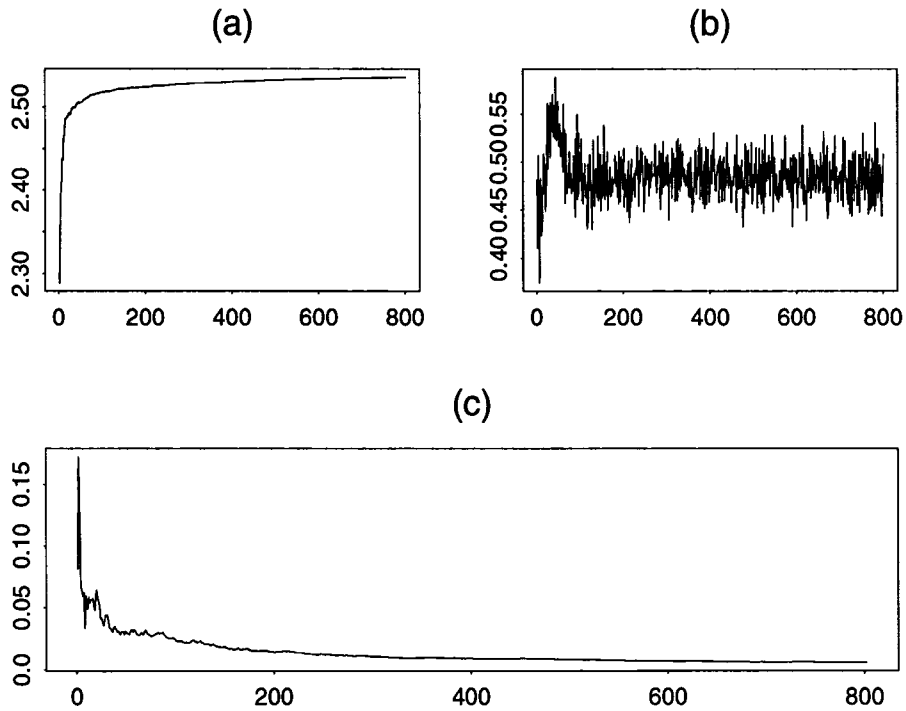

Figure 4.9: (a) Norm of the estimators (b) Controlled probabilities aroun 1/2 (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = -1.2y_t - 1.32u_t + .1u_{t-1} + u_{t-2}$.

We would like to point out however that proving self-tuning and self optimality are more difficult problems which are not going to be discussed here.

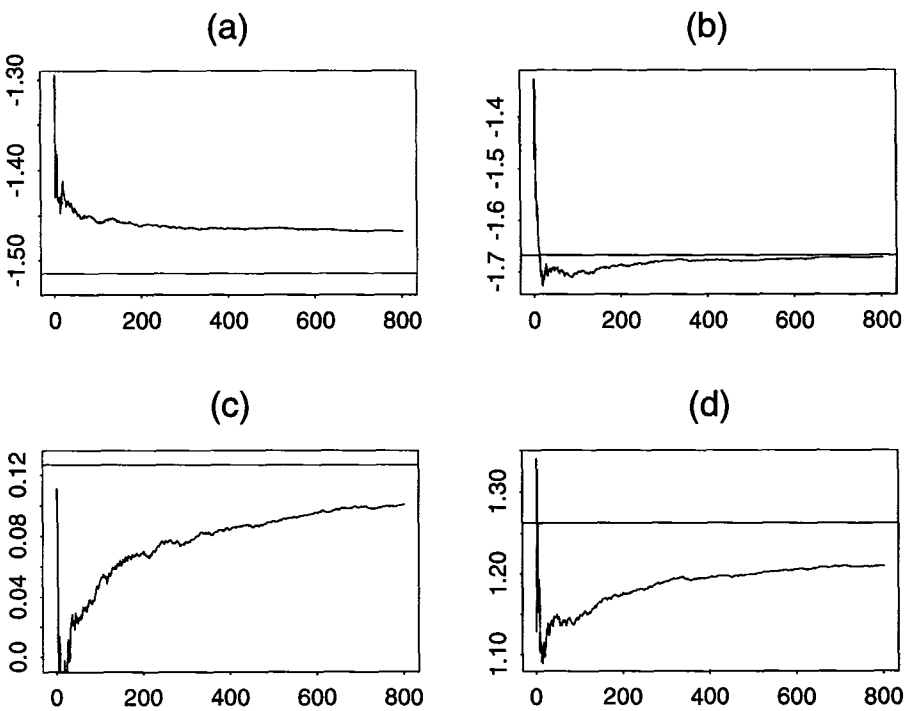

Figure 4.10: (a) Iterations for β_1 (b) Iterations for β_2 (c) Iterations for β_3 (d) Iterations for β_4 for the model $\lambda_{t+1} = -1.2y_t - 1.32u_t + .1u_{t-1} + u_{t-2}$.

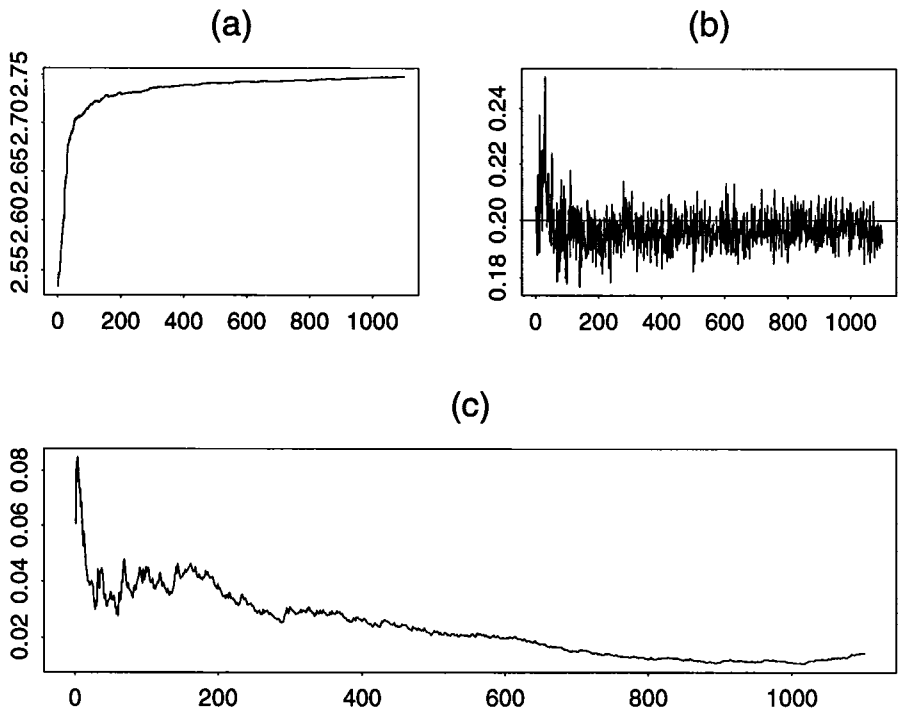

Figure 4.11: (a) Norm of the estimators (b) Controlled probabilities around .2 (c) Norm of the difference $\hat{\beta}_t - k\beta$.

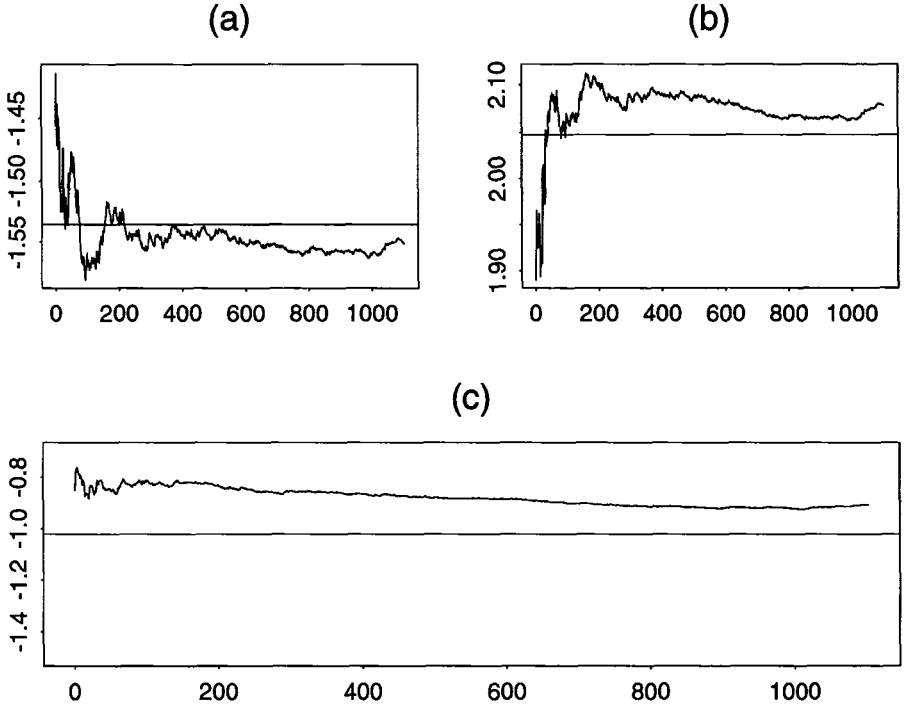

Figure 4.12: (a) Iterations for β_1 (b) Iterations for β_2 (c) Iterations for β_3 for the control of probabilities around .2.

Chapter 5

Main Results and Further Research

5.1 Main Results

We now summarize the chief results of this dissertation.

In Chapters 2 and 3 we studied regression models for nonstationary categorical time series. Specifically, we investigated the asymptotic properties of the maximum partial likelihood estimator. We proved that this estimator exists, it is unique, consistent and asymptotically normally distributed (Theorem 2.6.1). The results were applied to rainfall data. We judged the quality of the various models by a goodness of fit test (Section 2.7). We saw the importance of a certain covariate in the categorical prediction of rain rate.

In Chapter 4 we studied the adaptive control of binary time series. We derived a method of updating estimators in generalized linear models as data keep coming sequentially in time. The recursions are given by (4.21)-(4.22). These recursions suitably modified lead to a stochastic approximation type scheme. The last one was used to connect the areas of generalized linear models and control theory. Some optimality criteria (self-optimality and self-tuning) were proved for the proposed control law.

5.2 Further Research

In this chapter we discuss some further problems for study that follow from this dissertation.

An essential feature of every model building procedure is the assessment of its validity. Our goodness of fit tests (Section 2.7) are not the only ones that can be used. The area of regression diagnostics for categorical time series models has not been developed fully yet.

Estimation of parameters in the random coefficients model (Section 3.5) is another area that deserves special attention. We conjecture that the Gibbs sampler (see [29]) is a promising way to estimate parameters and predict the random effects.

The recursive algorithm (4.22) for updating the estimators is a promising way to obtain strongly consistent estimators for time series models based on generalized linear models. It can be easily generalized in the case of the canonical link ([36]). Furthermore this thesis did not address the statistical properties of the algorithm. This can be a very promising area of research. In addition, we believe that these recursions can be extended to the case of arbitrary link as well, under some assumptions. The modification of this algorithm to a stochastic approximation with the help of a logistic regression model gave us an opportunity to link the area of control and the area of generalized linear models. Further research needs to be done on this topic. Especially proof, of self-optimality and self-tuning for an adaptive control law that tracks the probabilities along a specified trajectory is of interest (see the discussion at the end of Section 4.8). Applications of this theory is certainly of importance. If we generalize it, say, to the case of a categorical time series with three categories, then a possible

application will be in the area of quality control. In other words, one would like to control the observed process within some limits (see [17]).

Appendix A

The TOGA/COARE Data

Radar rainfall measurements over the western Pacific warm pool were collected by two shipboard Doppler radars as part of the Tropical Oceans Global Atmosphere (TOGA) Coupled Ocean Atmosphere Response Experiment (COARE) during the Intensive Observing Period (IOP), November 1992— February 1993. The MIT and TOGA radars were carried by the Research Vessel (R/V) John V. Vickers (U.S.A.) and R/V Xiang Yang-Hong #5 (People's Republic of China) in a series of three cruises during which the ships maintained station with the Intensive Flux Array (IFA) near 2°S, 156°E. Merged and single radar rainfields having a spatial resolution of 2 km $\times$ 2 km and a temporal resolution of 10 minutes were produced from these observations. In order to obtain a constant sized averaging domain under a wide variety of meteorological conditions, the gridded rainfall data used in this study are from the MIT radar, covering a circle of diameter 290 km, during Cruise 3, January 29 to February 23, 1993. For more details see [89].

Bibliography

[1] A. Agresti. *Categorical Data Analysis.* Wiley, New York, 1990.

[2] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

[3] K. P. Andersen, O. Borgan, D. R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer-Verlag, New York, 1993.

[4] P. K. Andersen and R. D Gill. Cox's regression models for counting process: a large sample approach. *Annals of Statistics*, 10:1100–1120, 1982.

[5] J. A. Anderson and R. P. Phillips. Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics*, 30:22–31, 1981.

[6] F. J. Aranda-Ordaz. On two families of transformations to additivity for binary response data. *Biometrika*, 68:357–363, 1981.

[7] E. Arjas and P. Haara. A logistic regression model for hazard: Asymptotic results. *Scandinavian Journal of Statistics*, 14:1–18, 1987.

[8] K. J. Astrom and B. Wittenmark. On self-tuning regulators. *Automatica*, 9:185–199, 1973.

[9] I. V. Basawa and R. L. S. Prakasa Rao. *Statistical Inference for Stochastic Processes.* Academic Press, London, 1980.

[10] A. Becker, P. R. Kumar, and C. Z. Wei. Adaptive control with the stochastic approximation algorithm: geometry and convergence. *IEEE Trans. Autom. Control*, AC-30:330–338, 1985.

[11] J. Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357–365, 1944.

[12] D. Bertsekas and S.E. Shreve. *Stochastic Optimal Control: The Discrete Time Case.* Academic Press, New York, 1978.

[13] P. Billingsley. *Statistical Inference for Markov Processes.* Univ. Chicago Press, Chicago, 1961.

[14] P. Billingsley. *Probability and Measure.* Wiley, New York, 2nd edition, 1986.

[15] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis. Theory and Practice.* MIT Press, Cambridge:Mass, 1975.

[16] E. G. Bonney. Logistic regression for dependent binary observations. *Biometrics*, 43:951–973, 1987.

[17] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, 2nd edition, 1976.

[18] H. J. Brascamp and E. H. Lieb. Some inequalities for Gaussian measures and the long-range order of the one-dimensional plasma. In Arthurs A. M.,

editor, *Functional Integration and its Applications*, pages 1–14. Clarendon Press, Oxford, 1975.

[19] P. J. Brockwell and R. A. Davis. *Time Series: Data Analysis and Theory.* Springer-Verlag, New York, 2nd edition, 1991.

[20] J. Burridge. Some unimodality properties of likelihood derived from grouped data. *Biometrika*, 69:145–151, 1982.

[21] Y. S. Chow. Local convergence of martingales and the law of large numbers. *Annals of Mathematical Statistics*, 36:552–558, 1965.

[22] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression.* Chapman and Hall, London, 1982.

[23] D. R. Cox. Partial likelihood. *Biometrika*, 62:69–76, 1975.

[24] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[25] D. R. Cox and E. J. Snell. *The Analysis of Binary Data.* Chapman and Hall, London, 2nd edition, 1989.

[26] C. Czado. On link selection in generalized linear models. In L. Fahrmeir, editor, *Advances in GLIM and Statistical Modelling*, volume 78, pages 60–65. Springer Lecture Notes in Statistics, 1992.

[27] M.H.A. Davis and R.B. Vinter. *Stochastic Modelling and Control.* Chapman and Hall, London, 1985.

[28] J. P. Diggle, K-Y. Liang, and L. S. Zeger. *Analysis of Longitudinal Data.* Oxford University Press, New York, 1994.

[29] Gelfand A. E. and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85:398–409, 1990.

[30] Y. M. El-Fattah. Gradient approach for recursive estimation and control in finite markov chains. *Advances in Applied Probability*, 13:778–803, 1981.

[31] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimates in generalized linear models. *Annals of Statistics*, 13:342–368, 1985.

[32] L. Fahrmeir and H. Kaufmann. Regression models for nonstationary categorical time series. *Journal of Time Series Analysis*, 8:147–160, 1987.

[33] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag, New York, 1994.

[34] D. Finney. *Probit Analysis.* Cambridge University Press, Cambridge, 3rd edition, 1971.

[35] R. A. Fisher and F. Yates. *Statistical Tables.* Oliver and Boyd, Edinburg, 1938.

[36] K. Fokianos and B. Kedem. Recursive estimation for time series models following generalized linear models. Technical Report 96-48, Institute for Systems Research, University of Maryland at College Park, 1996.

[37] F. C. Genter and V. T. Farewell. Goodness-of-link testing in ordinal regression models. *The Canadian Journal of Statistics*, 13:37–44, 1985.

[38] G. Goodwin, P. Ramadge, and Caines P. Discrete time stochastic adaptive control. *SIAM Journal of Control and Optimization*, 19:829–853, 1981.

[39] G. C. Goodwin and K. S. Sin. *Adaptive Filtering, Prediction and Control.* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[40] A. F. Graybill. *Matrices with Applications in Statistics.* Wadsworth, Belomnt, California, 2nd edition, 1983.

[41] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of Royal Statistical Society*, B46:149–192, 1984.

[42] S. J. Haberman. *The Analysis of Frequency Data.* University of Chicago Press, Chicago, 1974.

[43] P. Hall and C. C. Heyde. *Martingale Limit Theorems and its Applications.* Academic Press, New York, 1980.

[44] O. Hernandez-Lerma. *Adaptive Markov Control Processes.* Springer-Verlag, New York, 1989.

[45] P. A. Jacobs and P. A. W. Lewis. Discrete time series generated by mixtures. I. correlation and runs properties. *Journal of Royal Statistical Society*, B40:94–105, 1978.

[46] P. A. Jacobs and P. A. W. Lewis. Discrete time series generated by mixtures. II. asymptotic properties. *Journal of Royal Statistical Society*, B40:222–228, 1978.

[47] H. Kaufmann. Regression models for nonstationary time series: Asymptotic estimation theory. *Annals of Statistics*, 15:79–98, 1987.

[48] H. Kaufmann. On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models. *Metrika*, 13:291–313, 1989.

[49] B. Kedem. *Binary Time Series.* Marcel Dekker, New York, 1980.

[50] B. Kedem. *Time Series Analysis by Higher Order Crossings.* IEEE Press, New York, 1994.

[51] B. Kedem, L. S. Chiu, and Z. Karni. An analysis of the threshold method for measuring area average rainfall. *Journal of Applied Meteorology*, 29:3–20, 1990.

[52] B. Kedem and H. Pavlopoulos. On the threshold method for rainfal estimation: Choosing the optimal threshold level. *Journal of American Statistical Association*, 86:626–633, 1991.

[53] D. M. Keenan. A time series analysis of binary data. *Journal of American Statistical Association*, 77:816–821, 1982.

[54] E. L. Korn and A. S. Whittemore. Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, 35:795–802, 1979.

[55] P.R. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal of Control and Optimization*, 23:329–380, 1985.

[56] P.R Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification and Control.* Prentice-Hall, Englewood Cliffs, NJ, 1986.

[57] T. Z. Lai and C. Z. Wei. Least squares estimation in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10:154–166, 1982.

[58] E. L. Lehmann. *Theory of Point Estimation.* Wiley, New York, 1983.

[59] W. K. Li. Time series models based on generalized linear models: some further results. *Biometrics*, 50:506–511, 1994.

[60] K.-Y. Liang and S. L. Zeger. A class of logistic regression models for multivariate binary time series. *Journal of American Statistical Association*, 84:447–451, 1989.

[61] J. K. Lindsey. *Models for Repeated Measurements.* Oxford University Press, New York, 1993.

[62] K. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Autom. Control*, AC-22:551–575, 1977.

[63] K. Ljung. On positive real transfer functions and the convergence of some recursive schemes. *IEEE Trans. Autom. Control*, AC-22:539–551, 1977.

[64] Z. A. Lomnicki and S. K. Zaremba. Some applications of zero-one processes. *Journal of Royal Statistical Society*, B17:243–255, 1955.

[65] C. F. Manski and D. McFadden. Alternative estimators and sample designs for discrete choice analysis with econometric applications. In C. F. Manski and D. McFadden, editors, *Structural Analysis of Discrete Data*, pages 2–50. MIT Press, Cambridge, MA, 1981.

[66] P. McCullagh. Regression models for ordinal data (with discussion). *Journal of Royal Statistical Society*, B42:109–142, 1980.

[67] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, London, 2nd edition, 1989.

[68] D. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

[69] R. Meneghini and A. J. Jones. An approach to estimate the areal rain-rate distribution from spaceborne radar by the use of multiple thresholds. *Journal of Applied Meteorology*, 32:386–398, 1993.

[70] M. L. Morrisey. The effects of data resolution on the area threshold method. *Journal of Applied Meteorology*, 33:1263–1270, 1994.

[71] L. R. Muenz and L.V. Rubinstein. Markov models for covariate dependence of binary sequences. *Biometrics*, 41:91–101, 1985.

[72] S. Murphy and B. Li. Projected partial likelihood and its application to longidutinal data. *Biometrika*, 82:399–406, 1995.

[73] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, A135:370–384, 1972.

[74] M. B. Nevel'son and R. Z. Has'minski. *Stochastic Approximation and Recursive Estimation.* American Mathematical Society Translations, Providence, Rhode Island, 1973.

[75] B. G. Nicholls, D. F. anf Quinn. *Random Coefficients Autoregressive Models: An Introduction*, volume 11 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1982.

[76] W. J. Pratt. Concavity of the log-likelihood. *Journal of the American Statistical Association*, 76:103–106, 1981.

[77] D. Pregibon. Goodness of link tests for generalized linear models. *Applied Statistics*, 29:15–24, 1980.

[78] R. L. Prentice. A generalization of the probit and logit methods for dose response curves. *Biometrics*, 32:761–768, 1976.

[79] M. B. Priestley. *Spectral Analysis and Time Series.* Academic Press, London, 1981.

[80] M. B. Priestley. *Nonlinear and Nonstationary Time Series.* Academic Press, New York, 1988.

[81] C. R. Rao. *Linear Statistical Inference and Its Applications.* Wiley, New York, 2nd edition, 1973.

[82] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[83] H. Robbins and D. Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimization Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.

[84] D. Rosenfeld, D. Atlas, and D. Short. The estimation of convective rainfall by area integrals 2. the height-area threshold (hart) method. *Journal of Geophysical Research*, 95(D3):2161–2176, 1990.

[85] D. J. Sakrison. Efficient recursive estimation: application to estimating the parameters of a covariance function. *International Journal of Engineering Science*, 3:461–483, 1965.

[86] T. J. Santner and Duffy E. D. *Statistical Analysis of Discrete Data.* Springer-Verlag, New York, 1989.

[87] D. Schoenfeld. Chi-square goodness-of-fit test for the proportional hazards regression model. *Biometrika*, 67:145–153, 1980.

[88] D. Short, B. D. Wolff, D. Rosenfeld, and D. Atlas. A study of the threshold method utilizing rain-gauge data. *Journal of Applied Meteorology*, 32:1379–1387, 1993.

[89] D. A. Short, P. A. Kucera, B. S. Ferrier, and O. W. Thiele. Coare iop rainfall from shipborne radars: 1. rain mapping algorithms. In *27'th Conference on Radar Meteorology*, Vail, Colorado, 1995. to appear.

[90] M.J Silvapulle. On the existence of maximum likelihood estimates for the binomial response models. *Journal of Royal Statistical Society*, B43:310–313, 1981.

[91] E. Slud. Partial likelihood for continuous time stochastic processes. *Scandinavian Journal of Statistics*, 19:97–109, 1992.

[92] E. Slud and B. Kedem. Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica*, 4:89–106, 1994.

[93] E. J. Snell. A scaling procedure for ordered categorical data. *Biometrics*, 20:592–607, 1964.

[94] R. D. Stern and R. Coe. A model fitting analysis of daily rainfall data. *Journal of Royal Statistical Society*, A47:1–34, 1984.

[95] W. F. Stout. *Almost Sure Convergence.* Academic Press, New York, 1974.

[96] O. D. Stram, J. L. Wei, and H. J. Ware. Analysis of repeated ordered categorical outcomes with possibly missing observations and time dependent covariates. *Journal of American Statistical Association*, 83:631–637, 1988.

[97] T. A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83:426–431, 1988.

[98] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–179, 1967.

[99] R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates. *Biometrika*, 63:27–32, 1976.

[100] W. H. Wong. Theory of partial likelihood. *Annals of Statistics*, 14:88–123, 1986.

[101] K. Yin. Adaptive control of chemical processes: Survey and a case study. Master's thesis, University of Maryland at College Park, 1990.

[102] S. L. Zeger. A regression model for time series of counts. *Biometrika*, 75:621–629, 1988.

[103] S. L. Zeger and R. Karim. Generalized linear models with random effects. *Journal of the American Statistical Association*, 86:79–86, 1991.

[104] S. L. Zeger and K.-Y. Liang. Feedback models for discrete and continuous time series. *Statistica Sinica*, pages 51–64, 1991.

[105] S. L. Zeger, K.-Y. Liang, and S. G. Self. The analysis of binary longitudinal data with time independent covariates. *Biometrika*, 72:8–31, 1985.

[106] S. L. Zeger and B. Qaqish. Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, 44:1019–1031, 1988.