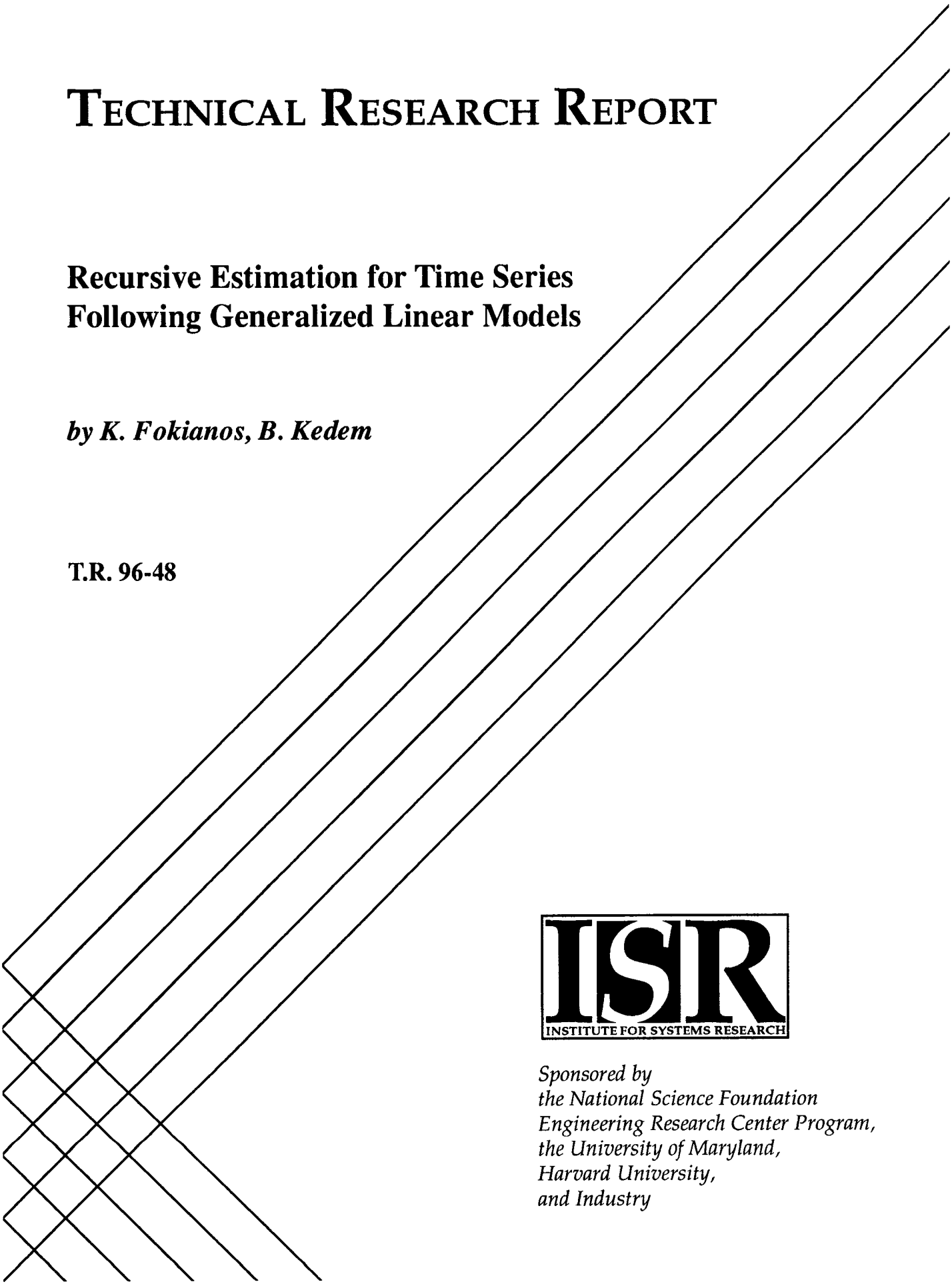
TECHNICAL RESEARCH REPORT

Recursive Estimation for Time Series Following Generalized Linear Models

by K. Fokianos, B. Kedem

T.R. 96-48



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

Recursive Estimation for Time Series Following Generalized Linear Models

Konstantinos Fokianos & Benjamin Kedem
Department of Mathematics and Institute for Systems Research
University of Maryland, College Park, MD 20742, U.S.A.

May 1996

Abstract

A recursive estimation method for time series models following generalized linear models is studied in two ways. The estimation procedure, suitably modified, gives rise to a stochastic approximation scheme. We use the modified estimation procedure to illustrate a connection between control theory and generalized linear models by employing a logistic regression model.

Keywords: longitudinal data, stochastic approximation, partial likelihood, tracking, control, iterative reweighted least squares.

1 Introduction

Generalized Linear Models (GLM)–as described by McCullagh & Nelder (1989)–provide a useful framework for modeling time series, and there is a growing interest among statisticians regarding this research topic. The papers by Fahrmeir & Kaufmann (1987), Zeger (1988), Zeger & Liang (1991), Slud & Kedem (1994) and Li (1994), to name only a few, make this claim evident. Primarily, the scientific interest is focused on the regression parameters. The main approaches that have been used for the estimation of the parameters are either likelihood or estimating functions based. This implies in particular that we fit a model after the termination of an expirement. But there may be occasions where the analyst observes data sequentially in time and wishes to update the estimators as soon as a new datum becomes available, without waiting for the expirement to end.

The present work addresses this question in the context of GLM in proposing a recursive estimation method for time series models. Our main results is the recursive relation (3.6). It is derived by considering only canonical link models. We give two different proofs for the recursive relation. One is based on the score and the other on the fitting procedure.

The next main result of this paper is a connection between GLM and adaptive control theory. We show how to control a probability by using a simple logistic model. The logistic paradigm throws light on the behavior of more general models. This will be reported in future work.

Finally, motivated by the adaptive control of linear systems, we modify the recursion (3.6). This yields a new stochastic approximation scheme (4.12). We study briefly the adaptive control law and derive several optimality results. The last problem is attacked along the lines of Becker, Kumar & Wei (1985).

The plan of the paper is as follows. The next section describes the connection between GLM and time series. Then we proceed by proving the recursive estimation method. In section 4 we show how one can cross-fertilize ideas from GLM and control theory. This leads us to a stochastic approximation type algorithm. We conclude the paper by some simulations and some comments.

2 Generalized Linear Models and Time Series

Suppose that $\{y_t\}$, $t = 1, 2, \ldots$, is a univariate time series. Assume that there exists a vector of covariates, say $\{\mathbf{x}_t\}$, that influences the evolution of $\{y_t\}$. Let $\mathcal{F}_t$ denote the history of the joint process $\{y_t; \mathbf{x}_t\}$ up to and including time time t. Generalized Linear Models for time series are defined by the following assumptions:

1. We assume that $f_\theta(y_{t+1} \mid \mathcal{F}_t)$, the conditional distribution of y_{t+1} given the past, follows the canonical exponential family of distributions

$$f_\theta(y_{t+1} \mid \mathcal{F}_t) = c(y_{t+1}) \exp(\theta_{t+1} y_{t+1} - b(\theta_{t+1}))$$

with $c \geq 0$ and measurable. It is not difficult to check that $E(y_{t+1} \mid \mathcal{F}_t) = \partial b(\theta_{t+1})/\partial \theta_{t+1} = \mu(\theta_{t+1}) = \mu_{t+1}$ and $Var(y_{t+1} \mid \mathcal{F}_t) = \partial^2 b(\theta_{t+1})/\partial^2 \theta_{t+1} = v(\theta_{t+1}) = v_{t+1}$ Note, that we do not take into account any dispersion parameters. This can be done without loss of generality as we will see subsequently.

2. The covariate vector $\mathbf{x_t}$ influences the response process by means of a linear predictor

$$\gamma_{t+1} = \boldsymbol{\beta}' \mathbf{z}_t$$

Here $\boldsymbol{\beta}$ denotes a p-dimensional vector of parameters and $\mathbf{z}_t$ is a vector of random time dependent covariates. It may include past values of the response, past values of $\mathbf{x}_t$, or any interactions between them.

3. The linear predictor is related to the mean μ_{t+1} by the *link* function g

$$\gamma_{t+1} = g(\mu_{t+1})$$

where g is monotone and differentiable function. In particular if $g = \mu^{-1}$ then we obtain the so called canonical link.

Throughout the paper we use canonical link. It simplifies all subsequent calculations, and guarantees existence and uniqueness of the estimators (Weddeburn (1976)). The case of non-canonical link can be treated, in principle in the same manner. However existence and uniqueness of the estimators can be problematic issues.

A well known example of the canonical link is the logistic model (Cox & Snell (1989)). Accordingly, y_t is a binary time series (for example yes/no), and the model is described by the equation

$$\lambda_{t+1} = \log \frac{p_{t+1}}{1 - p_{t+1}} = \boldsymbol{\beta}' \mathbf{z}_t \tag{2.1}$$

with $p_{t+1} = E(y_{t+1} \mid \mathcal{F}_t) = P(y_{t+1} = 1 \mid \mathcal{F}_t)$. Another well known model for the analysis of binary data is the probit model (Finney (1971)). Here we have

$$p_{t+1} = \Phi(\boldsymbol{\beta}' \mathbf{z}_t)$$

with Φ denoting the cumulative distribution function of a standard normal random variable. This is a non-canonical link case. We will use the logistic model (2.1) to illustrate a connection between control theory and GLM.

The goal is to estimate $\boldsymbol{\beta}$. We use for this purpose partial likelihood, in the event that random time dependent covariates are under consideration (Slud & Kedem (1994)), conditional likelihood (Kaufmann (1987)), or estimating functions approach (Zeger (1988), Zeger & Liang (1991)).

We employ the maximum partial likelihood (PL) estimator (see Slud & Kedem (1994)). Then the partial score under the above general setup and the canonical link assumption is given by (McCullagh & Nelder (1989))

$$S_t(\boldsymbol{\beta}) = \nabla \log \mathrm{PL}(\boldsymbol{\beta}) = \sum_{s=1}^{t} \mathbf{z}_{s-1}(y_s - \mu_s(\boldsymbol{\beta})) \tag{2.2}$$

Furthermore the conditional information matrix is

$$\mathbf{G}_t(\boldsymbol{\beta}) = \sum_{s=1}^{t} \mathrm{Var}[\mathbf{z}_{s-1}(y_s - \mu_s(\boldsymbol{\beta})) \mid \mathcal{F}_{s-1}] = \sum_{s=1}^{t} \mathbf{z}_{s-1}\mathbf{z}'_{s-1} v_s(\boldsymbol{\beta}) \tag{2.3}$$

Note that if a dispersion parameter were present then the above quantities, (2.2)-(2.3), should have been divided by it. Also, it turns out that

$$\mathbf{G}_t(\boldsymbol{\beta}) = \nabla\nabla'(-\log(\mathrm{PL}(\boldsymbol{\beta})))$$

The most widely used method for the solution of $S_t(\boldsymbol{\beta}) = 0$ is Fisher scoring. In the case of canonical link this coincides with the Newton-Raphson method. Details are given in the second chapter of McCullagh & Nelder (1989). A crucial property of this method is that the estimator can

be represented as the solution to a weighted least squares problem, in the limit of the iterations. To see this, define the "working observation vector"

$$\tilde{\mathbf{y}}_t = \mathbf{Z}_t'\boldsymbol{\beta} + \mathbf{V}_t^{-1}\mathbf{s}_t \tag{2.4}$$

with

$$\mathbf{Z}_t = [\mathbf{z}_0, \ldots, \mathbf{z}_{t-1}]$$

and

$$\mathbf{V}_t = \mathrm{diag}(v_s)_{s=1}^t$$

$$\mathbf{s}_t = \mathbf{y}_t - \boldsymbol{\mu}_t$$

where

$$\mathbf{y}_t = (y_1, \ldots, y_t)'$$

$$\boldsymbol{\mu}_t = (\mu_1, \ldots, \mu_t)'$$

Then, we can see that $\boldsymbol{\beta}$ is a solution to a so called Iterative Reweighted Least Squares problem. (For a more general discussion of this phenomenon see Green (1984)). Upon convergence, the estimator is given by

$$\hat{\boldsymbol{\beta}}_t = (\mathbf{Z}_t\mathbf{V}_t\mathbf{Z}_t')^{-1}\mathbf{Z}_t\mathbf{V}_t\tilde{\mathbf{y}}_t \tag{2.5}$$

The above representation will provide an alternative way to obtain the recursive relations that we are about to derive.

3 Recursive Estimation

We will obtain now a recursive estimation method useful when the data are observed sequentially in time, and illustrate it with an example from control theory in the next section.

Denote the estimator of $\boldsymbol{\beta}$, at time t, by $\hat{\boldsymbol{\beta}}_t$. Suppose that a new observation, say $(y_{t+1}, \mathbf{z}_t)$ becomes available. How can we go from $\hat{\boldsymbol{\beta}}_t$ to $\hat{\boldsymbol{\beta}}_{t+1}$? Note that $\hat{\boldsymbol{\beta}}_t$ does not admit a closed form from equation (2.2). However, a recursive formula for $\hat{\boldsymbol{\beta}}_t$ can be obtained and we give two different derivations of this fact. Let $x_{k|l}$ indicate the estimation of a random quantity at time k that depends on the first l observations. It is going to be used throughout the reminder of this paper.

Following Sakrison (1965) (for a comprehensive account of recursive estimation see Nevel'son & Has'minski (1973)) we have by using Taylor expansion and equations (2.2)–(2.3),

$$\mathbf{0} = S_{t+1}(\hat{\boldsymbol{\beta}}_{t+1}) = S_{t+1}(\hat{\boldsymbol{\beta}}_t) - (\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t)\mathbf{G}_{t+1}(\tilde{\boldsymbol{\beta}}) + \circ_p(||\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t||)$$

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

$$\mathbf{0} = \mathbf{z}_t(y_{t+1} - \mu_{t+1}(\hat{\boldsymbol{\beta}}_t)) - (\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t)\mathbf{G}_{t+1}(\tilde{\boldsymbol{\beta}}) + \circ_p(||\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t||)$$

upon noting that

$$S_{t+1}(\hat{\beta}_t) = S_t(\hat{\beta}_t) + \mathbf{z}_t(y_{t+1} - \mu_{t+1}(\hat{\beta}_t))$$

and $S_t(\hat{\beta}_t) = S_{t+1}(\hat{\beta}_{t+1}) = \mathbf{0}$. Therefore we can write :

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \mathbf{G}_{t+1}^{-1}(\tilde{\beta})\mathbf{z}_t(y_{t+1} - \mu_{t+1}(\hat{\beta}_t)) + \circ_p(||\hat{\beta}_{t+1} - \hat{\beta}_t||)$$

provided that the above inverse exists. Now, $\tilde{\beta}$ is lying between $\hat{\beta}_t$ and $\hat{\beta}_{t+1}$, and the conditional information matrix is a continuous function (under some assumptions) of β. This observation indicates the possibility of replacing $\tilde{\beta}$ by $\hat{\beta}_t$. Therefore, the recursive algorithm becomes

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \mathbf{G}_{t+1|t}^{-1}\mathbf{z}_t(y_{t+1} - \mu_{t+1|t}) \tag{3.6}$$

with $\mathbf{G}_{t+1|t} = \mathbf{G}_{t+1}(\hat{\beta}_t)$ and $\mu_{t+1|t} = \mu_{t+1}(\hat{\beta}_t)$.

Equation (3.6) requires the inversion of a $p \times p$ matrix. This can be avoided by utilizing the matrix inversion lemma (Rao (1973)). Writing

$$\mathbf{G}_{t+1|t} = \mathbf{G}_{t|t} + \mathbf{z}_t\mathbf{z}_t'v_{t+1|t}$$

with $v_{t+1|t} = v_{t+1}(\hat{\beta}_t)$ we have that

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \frac{\mathbf{D}_{t|t}\mathbf{z}_t}{1 + \mathbf{z}_t'\mathbf{D}_{t|t}\mathbf{z}_t v_{t+1|t}^{-1}}(y_{t+1} - \mu_{t+1|t}) \tag{3.7}$$

$$\mathbf{D}_{t+1|t} = \mathbf{D}_{t|t} - \frac{\mathbf{D}_{t|t}\mathbf{z}_t\mathbf{z}_t'\mathbf{D}_{t|t}}{v_{t+1|t}^{-1} + \mathbf{z}_t\mathbf{D}_{t|t}\mathbf{z}_t'} \tag{3.8}$$

after some algebra and by letting $\mathbf{G}_{t|t}^{-1} = \mathbf{D}_{t|t}$. These relations now do not call for the inversion of any matrix but require starting value for $\hat{\beta}_t$ and $\mathbf{D}_{t|t}$. A good starting value for $\mathbf{D}_{0|0} = \epsilon\mathbf{I}$ with $\mathbf{I}$ the identity matrix. We need to point out that the above recursions do not necessarily produce the maximum partial likelihood estimator as we do not solve $S_t(\beta) = 0$ for a fixed sample. For a related problem see Kedem (1994)[Ch. 7]. We conjecture however that the difference between the resulting estimator and the maximum partial likelihood estimator tends to zero in probability as the sample size increases. The recursive estimation procedure can be generalized to any link function, provided that the maximum partial likelihood estimator is unique. 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 partial log-likelihood (see equation after (2.3)). Several other remarks follow.

If a dispersion parameter were present then it would have been cancelled out by equation (3.6). Therefore the assumption of a dispersion parameter is not necessary for the proof of the recursive

algorithm. Note also that the recursions (3.7)-(3.8) coincide with the well known recursion for linear time series models (see Kumar & Varaiya (1986)[Ch. 10]). We need to point out that these equations can be used for independent data.

We derive now the relations (3.7)-(3.8) using a different approach. Namely, we are going to use the fitting procedure (2.4) as described in section 2. To be more specific, Suppose we have a fixed number of observations, say t. In the limit of the Fisher-scoring iterations, the solution of the partial score equations, $\hat{\boldsymbol{\beta}}_t$, is given by the weighted least squares estimator (2.5). Since this has a closed form, the recursive formulas for the update of this weighted least squares can be used with weights equal to the inverse of $v_{t+1|t}$. Then we have from (2.5)

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + \mathbf{G}_{t+1|t}^{-1} v_{t+1|t} \mathbf{z}_t (\tilde{y}_{t+1} - \mathbf{z}_t \hat{\boldsymbol{\beta}}_t)$$

leading to

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + \mathbf{G}_{t+1|t}^{-1} \mathbf{z}_t (y_{t+1} - \mu_{t+1|t})$$

upon noting that $v_{t+1|t}\mathbf{z}_t(\tilde{y}_{t+1} - \mathbf{z}_t\hat{\boldsymbol{\beta}}_t) = \mathbf{z}_t(y_{t+1} - \mu_{t+1|t})$ by employing (2.4). This result is identical to (3.6). The rest follows as before. The last argument, just takes advantage of the fitting procedure. If $\{y_t\}$ is a binary time series, and the logistic model is used then (3.7) and (3.8) hold with

$$v_{t+1|t} = p_{t+1|t}(1 - p_{t+1|t})$$

The result is in accordance with the algorithm proposed in Walker & Duncan (1967). Their proof is based implicitly on the representation (2.4).

4 Controlling a Probability

The recursive estimation equations (3.7)-(3.8) will be suitably modified for an application to control theory. There is nowadays well developed theory concerning the adaptive control of linear systems. The books by Goodwin & Sin (1984), Kumar & Varaiya (1986), among others, provide an introduction to this subject. In general both input and output are continuous random variables and the goal of control is either to *regulate* or *track* the output with respect to some cost function. If the parameters of the system are known then we have a standard feedback control law. If the parameters of the system are unknown then we have an adaptive control problem. We say that an adaptive control law is *self-optimizing*, with respect to some cost function, if it yields the same cost as the optimal control law that would have been used if the parameters of the system were completely known. We refer to an adaptive control law as *self-tuning* if it approaches asymptotically the law that would have been used if the system were known. Another area of significant importance is the adaptive control of Markov chains (see Hernandez-Lerma (1989)). Here, we want to make the "best" decision under uncertainty with respect to some cost function. A summary of the recent advances on the adaptive control is given by Kumar (1985).

We use the logistic model (2.1) to model binary (0-1) time series. The purpose of the control is to keep the transition probabilities of the system,$\{p_t\}$, close to some value. For instance, suppose that one wants to keep a machine operational, say 90% of the times. Then it is wise to control the system, in terms say of fuels or maintenance, such that this can be achieved. We will demonstrate our point by controlling the transition probabilities near to 1/2. This implies that in the long run we would expect equal number of occurrences of both states. We call this the *regulation* problem. We are going to see that the control law that we propose posses analogous geometric properties as in Becker et al. (1985). This implies that their results are applicable in our case yielding the same kind of conclusions.

Let us be more specific. Suppose $\{y_t\}$ is a binary time series and let as in (2.1)

$$\begin{aligned} \lambda_{t+1} = \log \frac{p_{t+1}}{1-p_{t+1}} &= \sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=0}^{p} b_i u_{t-i} \\ &= \boldsymbol{\beta}' \mathbf{z}_t \end{aligned} \tag{4.9}$$

with $\boldsymbol{\beta} = (a_0, \ldots, a_p, b_0, \ldots, b_p)' \in R^{2p+2}$, $\mathbf{z}_t = (y_t, \ldots, y_{t-p}, u_t, \ldots, u_{t-p})'$,$b_0 \neq 0$ and $\{u_t\}$ is the control sequence. Clearly (4.9) implies that $\{y_t\}$ is a non-stationary Markov chain of order p. We assume without loss of generality the same order for both past input and control. Our original regulation problem now has been transformed to keeping the logits near zero. 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 (4.9) becomes zero. In other words we would be able to regulate such a system if we choose

$$u_t = -\frac{1}{b_0}[\sum_{i=0}^{p} a_i y_{t-i} + \sum_{i=1}^{p} b_i u_{t-i}] \tag{4.10}$$

in the case that the coefficients are known, i.e. when the system is completely known. Equivalently, this can be written as

$$u_t = -\frac{A(z)}{B(z)} y_t \tag{4.11}$$

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. As in the case of linear systems we need to assume that $B(z)$ has all its root outside the unit circle. That is our system satisfies the minimum phase condition.

Now in conjunction with 3.6, motivated by the stochastic gradient algorithm which is used for the control of linear system we propose the following stochastic approximation type algorithm (Robbins & Monro (1951)):

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + \gamma \frac{\mathbf{z}_t}{r_t}(y_{t+1} - p_{t+1|t}) \tag{4.12}$$

$$r_t = r_{t-1} + \mathbf{z}_t' \mathbf{z}_t q_t; \quad r(0) = 1 \tag{4.13}$$

where γ is constant greater or equal to 1 and $q_t = \exp(-\hat{\beta}_t' \mathbf{z}_t)/(1+\exp(-\hat{\beta}_t' \mathbf{z}_t))^2$. Notice that the only difference between (3.6) and (4.12) is that we replace the conditional information matrix by its trace r_t.

Based on the parameter estimate $\hat{\beta}_t$ available at time t, the control input u_t is chosen to satisfy $\mathbf{z}_t' \hat{\beta}_t = 0$. This is the so called Certainty Equivalence Principle. We replace the parameters by their corresponding estimators in the control law (4.10). In other words [1]

$$u_t = -\frac{1}{\hat{b}_0(t)}[\sum_{i=0}^{p} \hat{a}_i(t) y_{t-i} + \sum_{i=1}^{p} \hat{b}_i(t) u_{t-i}] \tag{4.14}$$

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 the control law is calculated by means of (4.12)-(4.13)-(4.14). This is applied to the system and an output y_{t+1} is obtained. The process continues by calculating a control u_{t+1} and so on.

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

$$C_t = \frac{\sum_{s=1}^{t} y_s}{t}$$

C_t close to $1/2$, is equivalent to self-optimality. Indeed, if the system was completely known and (4.10) had been used then $p_t = 1/2 \, \forall t$. This means that we have a homogeneous stationary Markov chain, so that

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

However, it doesn't depend on control inputs or parameters. An appealing alternative measure of performance is

$$D_t = \frac{\sum_{s=1}^{t} \lambda_s^2}{t}$$

which depends on the unknown parameters, past observations and control actions. Convergence of D_t to 0 indicates again self-optimality. The last criterion can be interpreted as "variance output" in line with the linear models. Another possible measure of performance is

$$\tilde{D}_t = \frac{1}{t} \sum_{s=1}^{t} (p_s - \frac{1}{2})^2$$

Self-optimality means $\tilde{D}_t \to 0$, as $t \to \infty$. We will examine all three measures of performance in the following theorem whose proof is given in an appendix. First, we need to state some assumptions.

Assumption C

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

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

C.2 The true probability measure which governs $\{y_t, \mathbf{z}_t\}$, obeys (4.9) with the true parameter being $\boldsymbol{\beta}$.

C.3 The random variables $\mathbf{z}_t$ 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}

$$\frac{\sum_{s=1}^{t} I_{[\mathbf{z}_t \in A]}}{t} \to \mu(A) \quad \text{a.s.}$$

Theorem 4.1 Consider the system (4.9). Suppose that assumption **(C)** holds. If the stochastic gradient based adaptive control law (4.12)-(4.13)-(4.14) is used then we have

1. The adaptive control law is self-optimizing with respect to all three criteria C_N, D_N and $\tilde{D}_N$.

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

Consider again the recursions (4.12)-(4.13)-(4.14). Then it is easy to check that $\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t$ is orthogonal to $\hat{\boldsymbol{\beta}}_t$. Indeed, since $\mathbf{z}_t'\hat{\boldsymbol{\beta}}_t = 0$, equation (4.12) implies

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

so that $\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t$ is parallel to $\mathbf{z}_t$. However, $\mathbf{z}_t'\hat{\boldsymbol{\beta}}_t = 0$ implies that $\hat{\boldsymbol{\beta}}_t$ is orthogonal to $\mathbf{z}_t$. Therefore , it follows that $\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_t$ are orthogonal.

We point out here that the last observation is a property of the algorithm and has no relation with the system being controlled. The above fact gives the following very simple proposition. Its proof is omited.

Proposition 4.1 Consider again the recursions (4.12)-(4.13)-(4.14). Then the following hold a.e. :

1. $||\hat{\boldsymbol{\beta}}_t||^2 = ||\boldsymbol{\beta}_0||^2 + \sum_{s=1}^{t} ||\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_{s-1}||^2$
2. $||\hat{\boldsymbol{\beta}}_{t+1}||^2 \geq ||\hat{\boldsymbol{\beta}}_t||^2 \quad \text{for } t \geq 0$
3. If $\sup_t ||\hat{\boldsymbol{\beta}}_t|| < \infty$ then $\{||\hat{\boldsymbol{\beta}}_t||\}$ converges a.s.
4. If $\sup_t ||\hat{\boldsymbol{\beta}}_t|| < \infty$ then $||\hat{\boldsymbol{\beta}}_{t+1} - \hat{\boldsymbol{\beta}}_t|| \to 0$ a.s. as $t \to \infty$
5. If $||\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}||$ converges so does $\{||\hat{\boldsymbol{\beta}}_t||\}$

with || || denoting the eucledean norm.

Combining Theorem 4.1 and Proposition 4.1 we see that $\hat{\boldsymbol{\beta}}_t$ converges to the intersection of two random spheres. The one sphere is centered at the origin and the other at $\boldsymbol{\beta}$. 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 $\boldsymbol{\beta}$. Now H consists of only one point if and only if $H \cap L \neq \emptyset$. To show that H and L are indeed tangential, it suffices to show that there exists a subsequence t_l such that

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

This can be proved by using Becker et al. (1985)[Lemma 3]. Using now Becker et al. (1985)[Lemmas 10,13] and the following technical lemma the proof of self-tuning can be worked out along the lines of Becker et al. (1985). The proof of the following lemma is given in the appendix.

Lemma 4.1 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

$$\lim_{N\to\infty} \sum_{t=1}^{N} \frac{1}{N}(\eta_{t-1}y_t + s_{t-1})^2 = 0 \text{ a.s.} \tag{4.15}$$

Then

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

and

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

Now we can state the theorem that shows that our proposed adaptive control law is self-tuning.

Theorem 4.2 Consider the system (4.9). Suppose that the system has no reduced minimum variance controllers, that is the polynomials $A(z)$ and $B(z)$ are irreducible. Assume that **(C)** is true. If the control law (4.12)-(4.13)-(4.14) is used then

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

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

with

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

2. The adaptive control law is self-tuning:

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

For the calculation of the random variable k, note from (1) of Proposition 4.1 that as $t \to \infty$ we have

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

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

5 Simulations

We present now a simulation study that clarifies some important points. 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 being generated by a first order autoregressive process with parameter .3. We have run 15 simulations and then took the average of the results. The first two hundred observations gave a preliminary estimate using the method of partial likelihood. We had $k = 1.0897$, $D_t = 0.0521$, $C_t = .4973$ and $\tilde{D}_t = .00029$. Figure (1) illustrates (a) the norm of the estimators $||\hat{\beta}_t||$, (b) the transition probabilities p_t 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 is an increasing and bounded sequence. The transition probabilities fluctuate around 1/2 and the norm of the difference converges to 0. An unfortunate characteristic is the slow rate of convergence being evident from Figure (1(c)). Although the partial likelihood estimator is satisfactorily close to the true value, we see that convergence towards zero is very slow. This is to be expected however since we have a stochastic approximation scheme.

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 (2) demonstrates the same picture as figure (1). Note that (2(c)) shows that the norm of the difference converges toward zero but again at a very slow rate (The dotted line indicates 0). We see that the recursions are not close to convergence. This is to ensure that if we want to control such a system we would rather obtain a preliminary estimator, like the maximum partial likelihood estimator and then use this as a starting value.

We proceeded by fitting 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 .3. We now use a time series of length 1850. The first two hundred observations gave again 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 (3) summarizes the same results as the previous ones.

6 Concluding Remarks and Further Research

We proved a recursive estimation method (3.6) for estimation of the regression coefficients for time series models based on generalized linear models. This work did not investigate the statistical properties of the algorithm which may be of interest. Furthermore, we believe that these recursions can be extended to the case of non-canonical link as well, under some assumptions. The modification of this algorithm to a stochastic approximation type with the help of a logistic regression model connected the areas of control and the area of generalized linear models. Further research needs to be done in this topic. Finally, we would like to mention that we can, in principle, track the probabilities along a specified trajectory (i.e. $p_t = .2$). Here, we will illustrate this point with Figure (4), which shows how one can control the probabilities around .2 (see figure 4(b)). We used the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ as before. Suppose that we want to control the probabilities around a known number r, $0 < r < 1$. Then the control law should be modified to

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

Note however that this law does not preserve the orthogonality property of the proposed algorithm. Hence some adjustments must be made. Our idea concerning this particular example was not to model the response probabilities but rather the log-odds ratio. In other words we can consider the following model

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

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 from Figure (4) that $k = 1.0239$, $D_N = 0.0264$, $C_N = 0.1971$ and $\tilde{D}_N = 0.0007$. We would like to point out however that to prove self-tuning and self optimality for $p_t \neq 1/2$ is more difficult.

References

Becker, A., Kumar, P. R. & Wei, C. Z. (1985), 'Adaptive control with the stochastic approximation algorithm: geometry and convergence', *IEEE Trans. Autom. Control* **AC-30**, 330–338.

Cox, D. R. & Snell, E. J. (1989), *The Analysis of Binary Data*, 2nd edn, Chapman and Hall, London.

Fahrmeir, L. & Kaufmann, H. (1987), 'Regression models for nonstationary categorical time series', *Journal of Time Series Analysis* **8**, 147–160.

Finney, D. (1971), *Probit Analysis*, 3rd edn, Cambridge University Press, Cambridge.

Goodwin, G. C. & Sin, K. S. (1984), *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ.

Green, P. J. (1984), 'Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives', *Journal of Royal Statistical Society* **B-46**, 149–192.

Hernandez-Lerma, O. (1989), *Adaptive Markov Control Processes*, Springer-Verlag, New York.

Kaufmann, H. (1987), 'Regression models for nonstationary time series: Asymptotic estimation theory', *Annals of Statistics* **15**, 79–98.

Kedem, B. (1994), *Time Series Analysis by Higher Order Crossings*, IEEE Press, New York.

Kumar, P. (1985), 'A survey of some results is stochastic adaptive control', *SIAM Journal of Control and Optimization* **23**, 329–380.

Kumar, P. & Varaiya, P. (1986), *Stochastic Systems:Estimation, Identification and Control*, Prentice-Hall, Englewood Cliffs, NJ.

Lai, T. Z. & Wei, C. Z. (1982), 'Least squares estimation in stochastic regression models with applications to identification and control of dynamic systems', *Annals of Statistics* **10**, 154–166.

Li, W. K. (1994), 'Time series models based on generalized linear models: some further results', *Biometrics* **50**, 506–511.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Nevel'son, M. B. & Has'minski, R. Z. (1973), *Stochastic Approximation and Recursive Estimation*, American Mathematical Society Translations, Providence, Rhode Island.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, 2nd edn, Wiley, New York.

Robbins, H. & Monro, S. (1951), 'A stochastic approximation method', *Annals of Mathematical Statistics* **22**, 400–407.

Robbins, H. & Siegmund, D. (1971), A convergence theorem for non-negative almost supermartingales and some applications, *in* J. S. Rustagi, ed., 'Optimization Methods in Statistics', Academic Press, New York.

Sakrison, D. J. (1965), 'Efficient recursive estimation: application to estimating the parameters of a covariance function', *International Journal of Engineering Science* **3**, 461–483.

Slud, E. & Kedem, B. (1994), 'Partial likelihood analysis of logistic regression and autoregression', *Statistica Sinica* **4**, 89–106.

Stout, W. F. (1974), *Almost Sure Convergence*, Academic Press, New York.

Walker, S. H. & Duncan, D. B. (1967), 'Estimation of the probability of an event as a function of several independent variables', *Biometrika* **54**, 167–179.

Weddeburn, R. W. M. (1976), 'On the existence and uniqueness of the maximum likelihood estimates', *Biometrika* **63**, 27–32.

Zeger, S. L. (1988), 'A regression model for time series of counts', *Biometrika* **75**, 621–629.

Zeger, S. L. & Liang, K.-Y. (1991), 'Feedback models for discrete and continuous time series', *Statistica Sinica* pp. 51–64.

Appendix

Proof of Theorem 4.1

The proof will be based on the following quadratic form

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

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

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

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

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

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

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

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 $\mathbf{z}_t'(\hat{\beta}_t - \beta) = -\mathbf{z}_t'\beta = -\lambda_{t+1}$. Thus

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

Define now the function,

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

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

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

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

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

It follows from Robbins & Siegmund (1971)[Theorem 1] that

I.

$$\sum_t \frac{f(\mathbf{z}_t'\boldsymbol{\beta})}{r_t} < \infty \text{ a.s.}$$

II.

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

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

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

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

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

we get that

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

On the other hand, by assumption C.4

$$\frac{1}{N}\sum_{t=1}^{N} f(\mathbf{z}_t'\boldsymbol{\beta}) \to \int f(\mathbf{z}'\boldsymbol{\beta})\mu(d\mathbf{z})$$

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

$$\mathbf{z}'\boldsymbol{\beta} = 0 \text{ a.s. } \mu$$

Therefore we get that

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

Therefore, asymptotic optimality of the proposed control law with respect to D_N was established. It is also clear that $\tilde{D}_N \to 0$. Now, we examine the behavior of C_N. Upon noting that $y_t - p_t$ is a martingale difference, 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, we can conclude therefore that

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

It follows that

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

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

Proof of Lemma 4.1

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 (4.15) 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

$$\frac{1}{N}\sum_{t=1}^{N}(\eta_{t-1}w_t + v_{t-1})^2 = \frac{1}{N}\sum_{t=1}^{N}\eta_{t-1}^2 w_t^2 + \frac{1}{N}\sum_{t=1}^{N}v_{t-1}^2 + \frac{2}{N}\sum_{t=1}^{N}\eta_{t-1}v_{t-1}w_t \qquad (A.1)$$

Using Lai & Wei (1982)[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 (A.1) and (4.15) that

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

and

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

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 (A.1) we have

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

a.s. on Ω^c. Because of (A.1) 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

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

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 Stout (1974)[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 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. □.

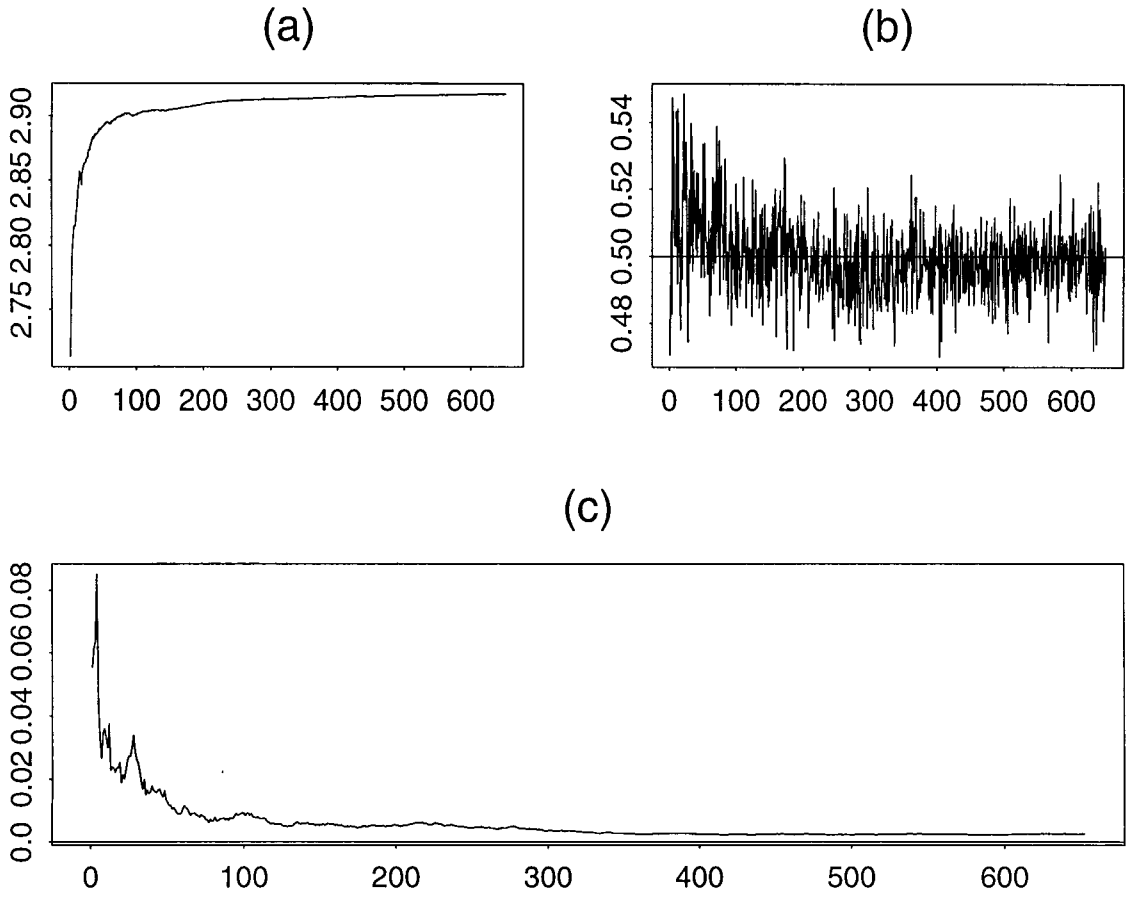


Figure 1: (a) Norm of the estimators $||\hat{\boldsymbol{\beta}}_t||$ (b) Controlled Probabilities around 1/2 (c) Norm of the difference $\hat{\boldsymbol{\beta}}_t - k\boldsymbol{\beta}$ for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = .3u_t + e_t$.

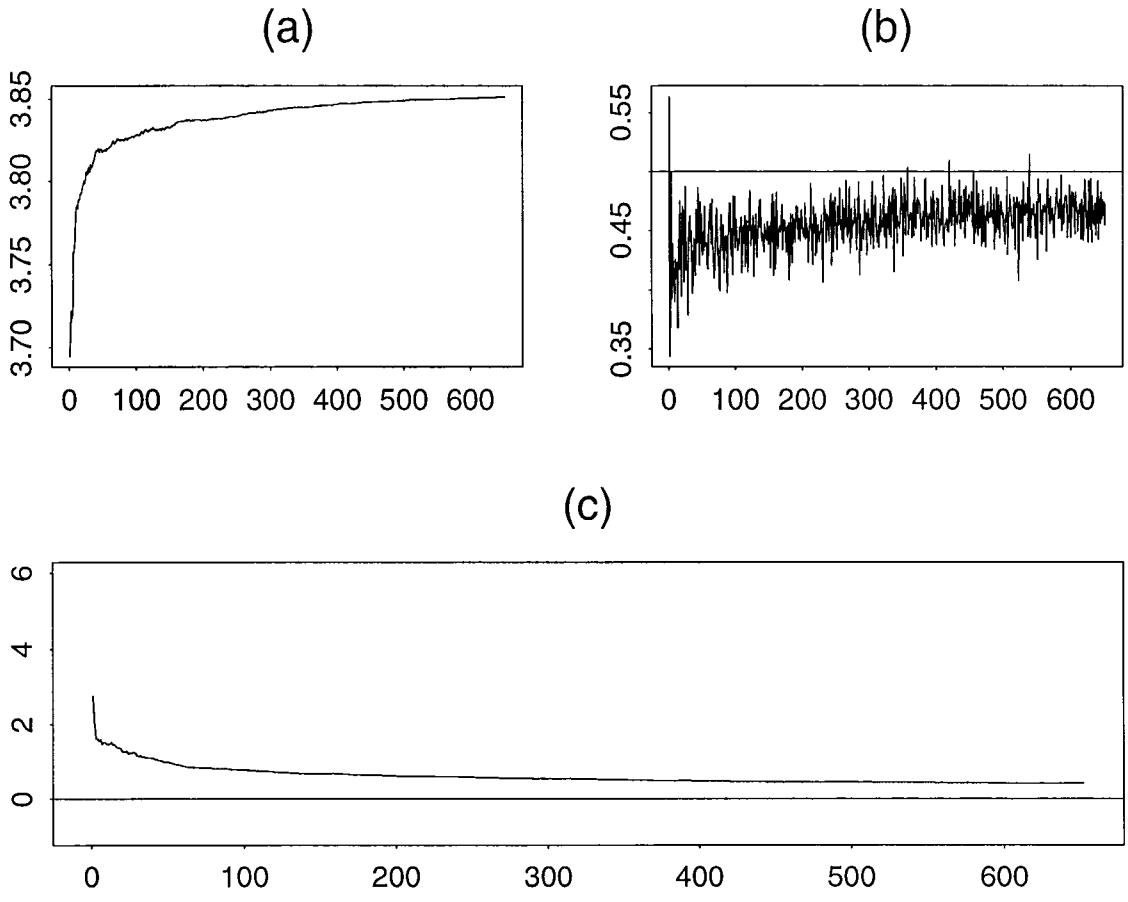


Figure 2: (a) Norm of the estimators $||\hat{\beta}_t||$ (b) Controlled Probabilities around 1/2 (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with starting values (-2,3.1,2) for the parameters and $u_t = .3u_t + e_t$

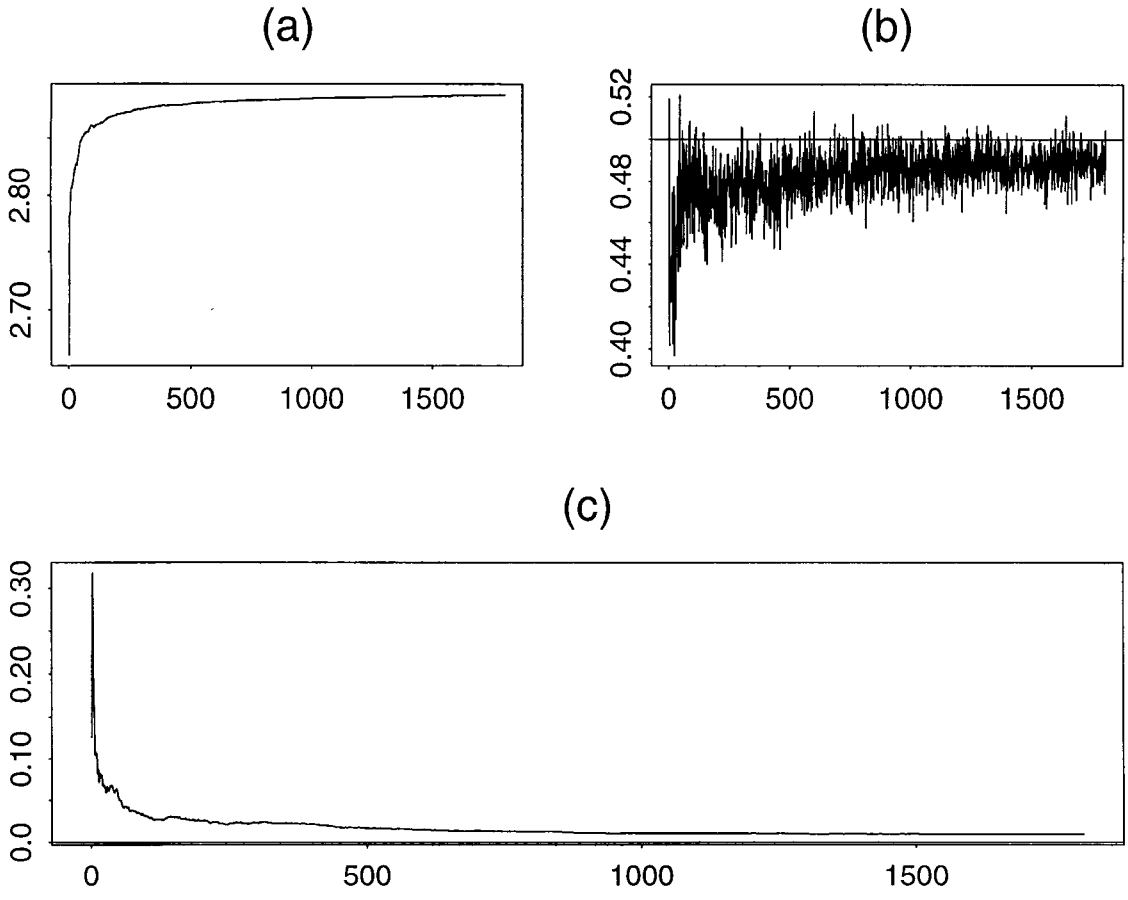


Figure 3: (a) Norm of the estimators $||\hat{\beta}_t||$ (b) Controlled Probabilities around 1/2 (c) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = y_t - .5y_{t-1} - 2u_t + u_{t-1}$ with $u_t = .3u_t + e_t$.

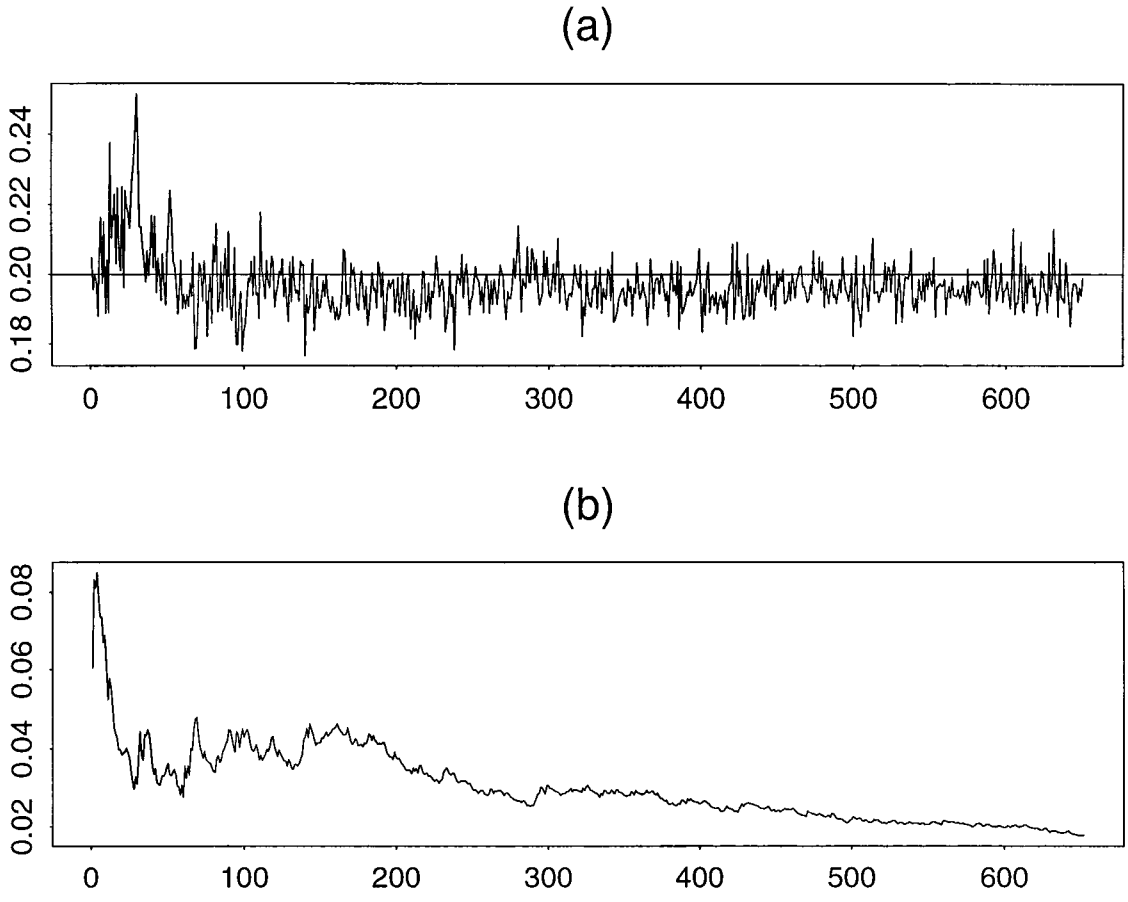


Figure 4: (a) Controlled Probabilities around .2 (b) Norm of the difference $\hat{\beta}_t - k\beta$ for the model $\lambda_{t+1} = -1.5y_t + 2u_t - u_{t-1}$ with $u_t = .3u_t + e_t$.