INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

MARKOV CHAIN MONTE CARLO SAMPLING FOR BAYESIAN COMPUTATION IN DYNAMIC MULTIVARIATE BINARY TIME SERIES

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

YU-TING CHENG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

October 1996

UNIVERSITY OF MINNESOTA

This is to certify that I have examined this copy of a doctoral thesis by

YU-TING CHENG

and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the final examining committee have been made.

LUKE TIERNEY

Name of Faculty Adviser(s)

Signature of Faculty Adviser(s)

20/10/95

Date

GRADUATE SCHOOL

Abstract

In social science research, it is common to observe multivariate time series data where the outcomes are binary. The classical approach is to fit a categorical response regression model. However, the application of these static regression models often impose unrealistic assumptions that a single simple model adequately represents a particular series at all possible times. These inappropriate static regressions are applied to the observed time series data in order to estimate what are really dynamic effects. In this thesis, a class of multivariate dynamic generalized linear models (MDGLM) is introduced and a Bayesian approach is developed.

This Bayesian approach extends the Gibbs sampling framework for dynamic generalized linear models by introducing more general Markov chain methods. This thesis outlines several basic Markov chain Monte Carlo methods, including Metropolis-Hastings algorithms, adaptive rejection sampling (ARS) algorithms and other variations. In addition, a modified ARS algorithm is derived for efficiently sampling from log-concave distributions. The methods are applied to analyzing two Minnesota Innovation Research Program (MIRP) studies, a research program on the management of innovation at the University of Minnesota Strategic Management Research Center.

Acknowledgements

I like to express my sincere gratitude to my advisor, Dr. Luke Tierney, for his invaluable advice and subsequent patience during the preparation of this thesis. This thesis is largely the result of Dr. Tierney's ideas and guidance. I am grateful for his careful supervision and his help in substantially improving the quality of the writing of this thesis. In addition, I thank Dr. Tierney for a Research Assistantship during the last two years of work in this thesis.

Special thanks to Professors Andrew Van de Ven and Charles Geyer. I have benefited from discussions with them. Thanks to Professors Kinley Larntz and Andrew Van de Ven for their helpful suggestions and comments regarding this thesis. I also wish to thank Professor Gary W. Oehlert and Douglas Polley for serving on my examination committee. My fellow graduate students deserve thanks. Their warm friendship have certainly made my life delightful.

Finally, I would like to thank my parent for their constant love and support. I wish to thank my wife Lan-Chin, son Justin and daughter Serene for their unfaltering encouragement and support. Without their love, warmth and understanding throughout my graduate studies nothing would have been possible. It is to them that I dedicate this thesis.

Contents

Abstract iv

Acknowledgements v

1 Introduction 1

1.1 Motivation 1

1.2 Models for Binary Time Series 3

1.3 Objectives and Thesis Outline 6

2 Literature Review 8

2.1 MDGLM 8

2.2 Statistical Inference for the MDGLM 9

2.2.1 Integration-Based Approaches 10

2.2.2 Posterior Mode Estimation 11

2.2.3 The Gibbs Sampler 15

2.3 Other Estimating Methods for the MDGLM 19

3 Markov Chain Monte Carlo Sampling 21

3.1 Metropolis-Hastings Sampling Algorithms 21

3.1.1 Random Walk Chains 23

3.1.2 Independence Chains 24

3.2 Adaptive Rejection Algorithms 25

3.3 Auxiliary Variable Methods 29

4 MCMC Samplers for Binary Data **32**

4.1 A Modified Version of ARS 32

4.1.1 Modified ARS within the Gibbs Sampler 33

4.1.2 Proof of the Modified Version of ARS 36

4.1.3 Modified ARS and DGLM 37

4.2 Use of M-H Algorithms 38

4.2.1 Random Walk Chains 39

4.2.2 Independence Chains 40

4.2.3 Block-At-A-Time M-H Algorithms 41

4.3 Link Functions 42

4.3.1 Probit Model 43

4.3.2 Mixtures of Normal Distributions 45

4.3.3 Auxiliary Variable Methods 47

4.3.4 Generalizations to a Multinomial Response 48

5 Performance of the MCMC Samplers **51**

5.1 Application to Binary Data 51

5.1.1 Binary Rainfall Data 51

5.1.2 Diagnostics by a Long Run Gibbs Sampler 52

5.1.3 Implementation Issues 56

5.2 Performance of the Gibbs Sampler 58

5.3 Performance of M-H Algorithms 61

5.4 Choice of Link Function 65

5.5 Comparison of the MCMC Samplers 70

5.6 Summary 73

6 Binary Events Analysis of Two MIRP Examples **75**

6.1 Co-Evolution Model 75
6.1.1 Cochlear Implant Evolution 75
6.1.2 Binary Events Analysis of the Co-Evolution Model 78
6.2 Adaptive Learning Model 85
6.2.1 Adaptive Processes of Organizational Learning during the Development of TAP 85
6.2.2 Binary Events Analysis of Learning Model 88
6.3 Discussion 93

7 Conclusions and Future Research **94**

Bibliography **96**

List of Tables

The data generating scheme for the starting points. 53

Estimated posterior means and standard deviations of the parameter β_{173} using rejection sampling and the modified ARS algorithms within the plain and block Gibbs sampler. 59

Estimated posterior means and standard deviations of the parameter β_{173} using M-H algorithms. 65

Estimated posterior means and standard deviations using different link functions. 69

Asymptotic relative efficiency of MCMC samplers compared to $t(7)$. 71

Van de Ven and Garud's (1992) parameter estimates of the time series regression analysis of the co-evolution model. 78

Process of learning. 86

List of Figures

Trajectory of the parameter β_{173}. 54
Empirical autocorrelation curve of the parameter β_{173}. 55
Estimated posterior means and standard deviations of the parameters using rejection sampling within the Gibbs sampler. 62
Rejection rates of random walk and independence chains. 63
Estimated standard errors of random walk and independence chains. . . . 63
Differences between the logistic and the Student t distribution with 5 - 11 degrees of freedom. 67
Plot of logistic quantile against $t(v)$ quantile with $v = 5$ - 11 degrees of freedom and $2ZK$ quantile for probabilities between 0.0001 and 0.9999. 68
Total variation distance between the density functions of the logistic and the Student t with 5 - 11 degrees of freedom. 69
Comparison of MCMC samplers. 72

Empirical autocorrelation curve of the parameters β_1, β_2, β_3, β_4, β_5, and β_6. 81
Odds of variation, selection, and retention events. 84
Empirical autocorrelation curve of the parameters β_1, β_2, and β_3. 90
Odds of continue actions and change actions events. 92

Chapter 1

Introduction

1.1 Motivation

In social science research, it is common to observe multivariate time series data where the outcomes are binary. The purposes of analyzing such data include assessing the association among variables, identifying lead-lag relationships among variables, and regressing one outcome on others as well as on fixed covariates. One example is from the Minnesota Innovation Research Program (MIRP). MIRP is a five-year (1983-1987) research program on the management of innovation. Since 1983, researchers at the University of Minnesota have been engaged in a longitudinal field research program with the objective of developing and testing the process theory of innovation which explains how and why innovation develops over time and what developmental paths may lead to success and failure for different kinds of innovations (Van de Ven and Associates, 1988). Fourteen related studies of a wide variety of innovations were undertaken by different research teams. Recognizing the limited research and theory on innovation processes in the literature, MIRP researchers decided that it might be more productive to undertake a grounded-theory strategy (Glaser and Strauss, 1967). In essence, the grounded-theory strategy is to discover a process theory of innovation from data systematically obtained from longitudinal research, then to test existing theories that were logically deduced from a priori assumptions which often do not fit or are not based on concrete particulars of the phenomena to be explained.

Two studies have made significant contributions to identifying and explaining process patterns in the development of innovation over time. They are the co-evolution model of technological and institutional innovations and the trial-and-error adaptive learning model. The co-evolution model presents a social evolutionary theory of change for explaining how technological and institutional innovations emerge as a continuous process of variation, selection, and retention events (Van de Ven and Garud, 1992). The learning model focuses on relationships between the action and outcome events of an innovation team within the joint venture as it develops the innovation over time, and on the influence environmental events have on the learning process (Van de Ven and Polley, 1992).

Observing the presence or absence of events of the variables in these two models, a time series analysis (Van de Ven and Garud, 1992; Van de Ven and Polley, 1992) was undertaken to estimate the relationships among the variables in the hypothesized models. To apply standard time series analysis methods it was necessary to aggregate the binary event sequence data into fixed temporal intervals. Given the absence of prior research on temporal intervals for events, Van de Ven and Garud (1992) and Van de Ven and Polley (1992) experimented with weekly, monthly, quarterly, and semiannual intervals. A monthly interval was chosen for aggregating events for both models because it provided the most substantively meaningful interpretation of the time series graphs and correlations among the variables. The results from the time series regression analyses provide substantial support for the hypotheses of the co-evolution model that variation, selection, and retention events endogenously co-produce each other over time, but contradict the other hypothesis of the co-evolution model that there is a significant self-reinforcing loop between variation and retention events. Also, the test results clearly contradict the learning model during an initial expansion period, but strongly support the model during a subsequent contraction period.

These empirical findings quite clearly contradict parts of the hypothesized models. Explanations for why these different patterns of innovation occurred over time and a further qualitative analysis of these historical innovation developments was provided by Van de Ven and Garud (1992) and Van de Ven and Polley (1992). Cheng and Van de Ven (1996) applied chaos theory to the contradictory part of these two innovation models to give another possible answer.

Both these qualitative analyses as well as chaos theory try to give an explanation to each model based on the monthly interval aggregation data. However, in some cases aggregation may diminish the direct relationships among the variables. Therefore, it is useful to re-examine the models just using the raw binary event series rather than aggregating events to monthly intervals. To analyze these hypothesized models using the binary event series would take each event into account and might detect the relationships more efficiently. In this dissertation, we will follow Van de Ven and Garud (1992) and Van de Ven and Polley (1992) and develop methodology that can be used to further investigate the hypothesized models by using the raw binary event series.

1.2 Models for Binary Time Series

In the literature, univariate binary time series have been studied in detail (Kedem 1980; Keenan 1982). Models for independent multivariate binary data are also well developed; Cox (1972) provides a review. Fewer methods are available for multivariate binary time series, unlike multivariate models for Gaussian outcomes, which have been studied in detail (e.g., Tiao and Box, 1981).

For multivariate binary time series, we are concerned with the construction of models for the analysis of sequences of binary data which may be serially correlated. Logistic regression models are commonly used (Cox, 1970) to study the relationship

between binary responses and a set of covariates. For example, a class of conditional logistic regression models for clustered binary data is considered in Connolly and Liang (1988); a logistic model for the conditional distributions of each series given the others for multivariate binary time series is proposed by Liang and Zeger (1989). However, the applications of these static regression models often impose unrealistic assumptions when applied to evolutionary processes during the development of technological innovation. In developing a process theory of innovation, the changing structural and technological conditions, individual behavior, and attitude may cause uncertainty to the hypothesized model of the process theory. This dynamic nature of processes and systems demands that we recognize uncertainty due to the passage of time. Further it might be recognized that, at some future time, the whole model or system form may change. Thus, it is necessary to consider the parameters in the hypothesized model as changing with time.

West, Harrison, and Migon (1985) defined the class of dynamic generalized linear models (DGLM) and developed a Bayesian approach to dynamic modeling and forecasting. In the standard dynamic generalized linear models form, univariate observations y_t are related to an unobservable time-varying parameter vector $\boldsymbol{\beta}_t$ by a linear observation equation $E(y_t|\boldsymbol{\beta}_t, y_1, \ldots, y_{t-1}) = h(\boldsymbol{z}_t'\boldsymbol{\beta}_t)$, where h is one of the common link functions, $\boldsymbol{z}_t$ is a function of covariates and, possibly, past responses, together with a linear evolution equation $\boldsymbol{\beta}_t = \boldsymbol{F}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t$ with a Gaussian noise process $\boldsymbol{\xi}_t$. For binary responses, such as the MIRP data, the time-varying parameters $\boldsymbol{\beta}_t$ relate to the response probabilities π_t as specified by $\pi_t = h(\boldsymbol{z}_t'\boldsymbol{\beta}_t)$. Two basic models are commonly used to analyze binary responses: the logit model with h the logistic distribution function, and the probit model with h the standard normal cumulative distribution function. The model has three features:

1. The process y_t is assumed to have a sampling distribution in the exponential family. This class of models includes sampling distributions that may be non-normal or not likely to be adequately modeled using normality, even after transformation.

2. The natural parameter of the sampling distribution is related to the unobserved parameter $\boldsymbol{\beta}_t$ through a link function. This link function provides a transformation from the natural parameter space to that of the linear predictor.

3. A linear (or nonlinear) Gaussian evolution model for the parameter $\boldsymbol{\beta}_t$. These time varying parameters allow one to include change over time into the model.

This Bayesian approach is built on existing practice in the sense that many common models can formulate and provide alternatives to the standard static regression model that do not suffer its drawbacks. It also has several advantages required for operating with little data, accommodating subjective information, on-line monitoring, and estimation of parameters.

Given the observations $y_1, \ldots, y_T$, estimation of $\boldsymbol{\beta}_t$ is a primary goal of inference. West, Harrison, and Migon (1985) discuss Bayesian inference and data analysis for the univariate dynamic generalized linear model. A key feature of the analysis is the use of conjugate prior and posterior distributions for the exponential family parameters and a derivation of an approximate filter for estimation of time-varying parameters. Unfortunately these methods are difficult to extend to the multivariate case; See Fahrmeir (1992) and Fahrmeir and Tutz (1994) for further discussion. Therefore, different estimation methods for multivariate dynamic generalized linear models need to be considered. There are three approaches: a full Bayesian analyses based on numerical integration (e.g., Kitagawa, 1987), posterior mode estimation (Fahrmeir, 1992), and Gibbs sampling (e.g., Carlin, Polson, and Stoffer, 1992). We will review and summarize these estimation methods in Chapter 2.

1.3 Objectives and Thesis Outline

A full Bayesian analysis based on posterior distributions for multivariate binary time series will generally require repeated multidimensional integrations. As an alternative, Monte Carlo techniques are particularly useful for problems in connection with Bayesian inference and can be applied in our situation. For statistical inference, it is necessary to draw a large number of samples from the posterior densities. In general, direct random drawings from the posterior densities are not available. Our main purpose in this dissertation is to develop several sampling procedures for dynamic multivariate binary time series models, to compare them to the known methods, and to apply them to the MIRP example. The rest of the thesis is organized as follows.

In Chapter 2, we briefly review multivariate dynamic generalized linear models. Three approaches to statistical inference based on posterior densities are introduced.

In addition to these three approaches, Markov chain Monte Carlo (MCMC) methods have been widely used in Bayesian inference problems. We outline a number of the basic Markov chain Monte Carlo samplers that are available for statistical inference in multivariate dynamic generalized linear models in Chapter 3.

In Chapter 4, we propose a method for rejection sampling that is useful for Gibbs sampling for multivariate dynamic generalized models. We also discuss issues related to the choice of the candidate generating densities and block sampling strategies when applying Markov chain Monte Carlo algorithms. Two link functions that are approximately equivalent to the logistic link function are introduced.

In Chapter 5, we apply the MCMC algorithms introduced in the Chapter 4 and empirically compare them with posterior mode estimates of the extended Kalman filter and smoother (EKFS) on a binary rainfall time series example. Some issues that arise in the implementation of MCMC algorithms are discussed. The comparisons of different MCMC algorithms are made based on runs of 50,000 observations.

In Chapter 6, we re-analyze the two studies of MIRP data using some of the methodology developed in this thesis. This analysis based on binary time series provides different views of the hypothesized models.

Finally, Chapter 7 provides a discussion of the results and gives directions for further research.

Chapter 2

Literature Review

As a basis, this chapter gives a short review of multivariate dynamic generalized linear models (MDGLM). Three estimation approaches of statistical inference for the MDGLM in the literature will be introduced. More detailed expositions can be found, for example, in West, Harrison, and Migon (1985), Kitagawa (1987), West and Harrison (1989), Goss (1990), Frühwirth-Schnatter (1991), Schnatter (1992), Fahrmeir (1992), Lindsey (1993), and Fahrmeir and Tutz (1994).

2.1 MDGLM

To establish notation, let responses, covariates, and parameters up to t be denoted by

$$\boldsymbol{y}_t^* = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_t')', \quad \boldsymbol{x}_t^* = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_t')', \quad \boldsymbol{\beta}_t^* = (\boldsymbol{\beta}_0', \ldots, \boldsymbol{\beta}_t')'.$$

where $\boldsymbol{y}_0^*$, $\boldsymbol{x}_0^*$ are empty and $\boldsymbol{y}_t$ and $\boldsymbol{\beta}_t$ have dimension q and p respectively.

The conditional density $p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*, \boldsymbol{x}_t^*)$ is assumed to be of q-dimensional exponential family type with conditional mean

$$E(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*, \boldsymbol{x}_t^*) = \boldsymbol{\mu}_t = \boldsymbol{h}(\boldsymbol{Z}_t'\boldsymbol{\beta}_t), \quad t = 1, 2, \ldots,$$

where $\boldsymbol{h}$: $\mathbf{R}^r \to \mathbf{R}^q$ is a two-times continuously differentiable link function, and $\boldsymbol{Z}_t$ is a $p \times r$ matrix depending on covariates and, possibly, on past responses $\boldsymbol{y}_{t-1}^*$.

For parameter transitions, we use a linear Gaussian evolution model:

$$\boldsymbol{\beta}_t = \boldsymbol{F}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t, \quad t = 1, 2, \ldots . \tag{2.1}$$

The error process $\boldsymbol{\xi}_t$ is Gaussian noise, $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t)$ with $\boldsymbol{\xi}_t$ independent of $\boldsymbol{\xi}_{t-1}$, $\ldots$, $\boldsymbol{\xi}_0$, $\boldsymbol{y}^*_{t-1}$, $\boldsymbol{x}^*_t$ for $t \geq 1$ and of $\boldsymbol{\beta}_0 \sim N(\boldsymbol{a}_0, \boldsymbol{Q}_0)$.

In the simplest form, the system matrices $\boldsymbol{Z}_t$, $\boldsymbol{F}_t$, $\boldsymbol{Q}_t$, $\boldsymbol{a}_0$, and $\boldsymbol{Q}_0$ are assumed to be deterministic and known. In many applications, however, the covariance matrix $\boldsymbol{Q}_t$, the initial values $\boldsymbol{a}_0$ and $\boldsymbol{Q}_0$ are unknown. One can assume these $\boldsymbol{Q}_t$, $\boldsymbol{a}_0$, and $\boldsymbol{Q}_0$, contain hyperparameters, say $\boldsymbol{\alpha}$, so that

$$\boldsymbol{Q}_t = \boldsymbol{Q}_t(\boldsymbol{\alpha}), \quad \boldsymbol{a}_0 = \boldsymbol{a}_0(\boldsymbol{\alpha}), \quad \text{and} \quad \boldsymbol{Q}_0 = \boldsymbol{Q}_0(\boldsymbol{\alpha}).$$

We treat the hyperparameters $\boldsymbol{\alpha}$ as unknown constants. Under the normality assumption, maximum likelihood is then a natural choice for estimation (Fahrmeir and Tutz, 1994). Other estimation procedures have been proposed, such as the EM algorithm (Goss, 1990), generalized least squares (Aoki, 1987; Harvey, 1989), and Bayes methods where $\boldsymbol{\alpha}$ is treated as a stochastic parameter with a prior distribution.

2.2 Statistical Inference for the MDGLM

When a model can be written in MDGLM form this provides the key for employing unified methods of statistical inference. Given the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$, estimation of $\boldsymbol{\beta}_t$ is the primary goal. This is termed filtering for $t = T$, smoothing for $t < T$, and prediction for $t > T$. Three approaches for estimating $\boldsymbol{\beta}_t$ have already been mentioned briefly in Section 1.2: a full Bayes analyses based on numerical integration, posterior mode estimation, and Gibbs sampler methods.

2.2.1 Integration-Based Approaches

The integration-based approach was presented by Kitagawa (1987). In the method described there, recursive formulas for filtering and smoothing were derived and they were implemented using numerical computations. A problem with this numerical method is that it requires intensive use of the computer. Due to its complexity and numerical effort it is difficult to apply to models when the dimension of the parameter vector β_t is large. Considerable work has been done on the refinement of the numerical algorithm for low dimensional modeling of the parameter vector β_t. For example, Hodges and Hale (1993) used a computationally more efficient integration algorithm and Tanizaki (1993) used a Monte Carlo random placement of knots method.

A similar approach was proposed by West and Harrison (1989). They suggest the application of Gauss-Hermite quadrature to solve the analytically intractable integrals in the conditional first two moments. However, due to the recursive dependence of the integrals in posterior densities over time, the numerical effort of their approach increases exponentially with time. Therefore, the method is mainly restricted to shorter time series.

A more practicable solution to the prediction and filtering problem in dynamic generalized linear models with linear Gaussian evolution models has been given by Schnatter (1992) and Frühwirth-Schnatter (1991). Schnatter (1992) also gives a simulation-based comparison of approximate filtered posterior means obtained by Gauss-Hermite quadrature and approximate filtered posterior modes obtained by the generalized extended Kalman filter. Their results show that the estimated posterior moments are often nearly identical after a few filtering steps.

2.2.2 Posterior Mode Estimation

To avoid numerical integration, which can become infeasible for higher dimensions, one way to estimate β_t is by posterior modes; that is, by maximization of posterior densities. Numerical maximization of posterior densities can be achieved by various algorithms. Fahrmeir and Kaufmann (1991) develop interative forward-backward Gauss-Newton (Fisher-Scoring) algorithms. Fahrmeir (1992) suggests the generalized extended Kalman filter and smoother as an approximate posterior mode estimator in dynamic generalized linear models, and shows that the EKFS can be considered as a simplified Fisher scoring algorithm. In their experience, it is a good compromise between computational simplicity and numerical accuracy for estimation in commonly used models (e.g., logit or probit models).

Given the data y_t^* and x_t^*, estimation of β_t^* is based on the posterior density $p(\beta_t^*|y_t^*,x_t^*)$. Repeated application of Bayes' theorem yields

$$\begin{aligned} p(\beta_T^*|y_T^*, x_T^*) &= \prod_{t=1}^T p(y_t|\beta_t^*, y_{t-1}^*, x_t^*) \prod_{t=1}^T p(\beta_t|\beta_{t-1}^*, y_{t-1}^*, x_t^*) \\ &\quad \times[\prod_{t=1}^T p(x_t|\beta_{t-1}^*, y_{t-1}^*, x_{t-1}^*)/p(y_T^*, x_T^*)]p(\beta_0). \end{aligned}$$

This can be simplified if the following assumptions hold:

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

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

(A2). Conditional on y_{t-1}^*, x_{t-1}^*, covariates x_t are independent of β_{t-1}^*, i.e.,

$$p(x_t|\beta_{t-1}^*, y_{t-1}^*, x_{t-1}^*) = p(x_t|y_{t-1}^*, x_{t-1}^*), \qquad t = 1, 2, \ldots$$

(A3). The parameter process is Markovian, i.e.,

$$p(\boldsymbol{\beta}_t|\boldsymbol{\beta}^*_{t-1}, \boldsymbol{y}^*_{t-1}, \boldsymbol{x}^*_t) = p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}), \qquad t = 1, 2, \ldots$$

Under assumptions (A1), (A2), and (A3), we obtain

$$p(\boldsymbol{\beta}^*_T|\boldsymbol{y}^*_T, \boldsymbol{x}^*_T) \propto \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}^*_{t-1}, \boldsymbol{x}^*_t) \prod_{t=1}^{T} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}) p(\boldsymbol{\beta}_0). \tag{2.2}$$

To avoid numerical integration, Fahrmeir (1992) proposes to estimate $\boldsymbol{\beta}^*_t$ by posterior modes; that is, by maximization of posterior densities.

The EKFS algorithm can be derived in a straightforward but lengthy way as an approximate posterior mode estimator by extending Sage and Melsa's (1971) arguments for maximum posterior estimation in nonlinear systems from conditionally Gaussian to exponential family observations. The same result can also be obtained by using Hartigan's (1969) linear Bayes arguments or by linearizing the observation equation around the current estimates. The following additional notation will be of use: Stressing dependence on $\boldsymbol{\beta}_t$, we write $\boldsymbol{\mu}_t(\boldsymbol{\beta}_t) = \boldsymbol{h}(\boldsymbol{Z}'_t\boldsymbol{\beta}_t) = E(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}^*_{t-1}, \boldsymbol{x}^*_t)$ for the conditional expectation, $\boldsymbol{\Sigma}_t(\boldsymbol{\beta}_t) = V(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}^*_{t-1}, \boldsymbol{x}^*_t)$ for the conditional covariance matrix, inserting $\boldsymbol{\mu}_t(\boldsymbol{\beta}_t)$ in the variance function of the exponential family, and $\boldsymbol{Z}_t$ is a function depending on past responses $\boldsymbol{y}^*_{t-1}$ and on present and past covariates $\boldsymbol{x}^*_t$. In the following, prediction, filter, and smoother steps $\hat{\boldsymbol{\beta}}_{t|t-1}$, $\hat{\boldsymbol{\beta}}_{t|t}$, $\hat{\boldsymbol{\beta}}_{t|T}$, denote the posterior mode estimators and $\hat{\boldsymbol{V}}_{t|t-1}$, $\hat{\boldsymbol{V}}_{t|t}$, $\hat{\boldsymbol{V}}_{t-1|T}$ are estimated error covariance matrices.

Extended Kalman Filter and Smoother (EKFS) :

1. *Prediction step*

   For $t = 1, \ldots, T$:

$$\hat{\boldsymbol{\beta}}_{t|t-1} = \boldsymbol{F}_t \hat{\boldsymbol{\beta}}_{t-1|t-1}, \quad \hat{\boldsymbol{\beta}}_{0|0} = \boldsymbol{a}_0,$$
$$\hat{\boldsymbol{V}}_{t|t-1} = \boldsymbol{F}_t \hat{\boldsymbol{V}}_{t-1|t-1} \boldsymbol{F}_t' + \boldsymbol{Q}_t, \quad \hat{\boldsymbol{V}}_{0|0} = \boldsymbol{Q}_0.$$

2. *Filter step*

   For $t = 1, \ldots, T$:

$$\hat{\boldsymbol{\beta}}_{t|t} = \hat{\boldsymbol{\beta}}_{t|t-1} + \boldsymbol{K}_t[\boldsymbol{y}_t - \boldsymbol{h}(\boldsymbol{Z}_t' \hat{\boldsymbol{\beta}}_{t|t-1})],$$
$$\hat{\boldsymbol{V}}_{t|t} = (I - \boldsymbol{K}_t \boldsymbol{Z}_t \boldsymbol{D}_t) \hat{\boldsymbol{V}}_{t|t-1},$$

   where $\boldsymbol{K}_t = \hat{\boldsymbol{V}}_{t|t-1} \boldsymbol{Z}_t' \boldsymbol{D}_t [\boldsymbol{Z}_t' \boldsymbol{D}_t \hat{\boldsymbol{V}}_{t|t-1} \boldsymbol{D}_t \boldsymbol{Z}_t' + \boldsymbol{\Sigma}_t]^{-1}$,

$$\boldsymbol{D}_t = \frac{\partial \boldsymbol{h}}{\partial (\boldsymbol{Z}_t' \boldsymbol{\beta}_t)},$$

   and $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$, $\boldsymbol{D}_t$, are evaluated at $\hat{\boldsymbol{\beta}}_{t|t-1}$.

3. *Smoother step*

   For $s = T, \ldots, 1$:

$$\hat{\boldsymbol{\beta}}_{s-1|T} = \hat{\boldsymbol{\beta}}_{s-1|s-1} + \boldsymbol{B}_s (\hat{\boldsymbol{\beta}}_{s|T} - \hat{\boldsymbol{\beta}}_{s|s-1}),$$
$$\hat{\boldsymbol{V}}_{s-1|T} = \hat{\boldsymbol{V}}_{s-1|s-1} + \boldsymbol{B}_s (\hat{\boldsymbol{V}}_{s|T} - \hat{\boldsymbol{V}}_{s|s-1}) \boldsymbol{B}_s',$$

   where $\boldsymbol{B}_s = \hat{\boldsymbol{V}}_{s-1|s-1} \boldsymbol{F}_s' \hat{\boldsymbol{V}}_{s-1|s-1}^{-1}$.

For the situation where the evolution model contains unknown hyperparameters $\boldsymbol{\alpha}$, Fahrmeir (1992) suggests an indirect approach, using a variant of the EM algorithm and substituting posterior modes for posterior expectations. In this context, the complete data set consists of $\boldsymbol{y}_t^*$ and $\boldsymbol{\beta}_t^*$. In its original version, the EM algorithm computes the next iterate $\boldsymbol{\alpha}^{(k+1)}$, given the current iterate $\boldsymbol{\alpha}^{(k)}$ by maximizing the conditional expectation $E[\ln p(\boldsymbol{y}_t^*, \boldsymbol{\beta}_t^* | \boldsymbol{x}_t^*, \boldsymbol{\alpha}^{(k)})]$ of the joint log density of

y_t^*, β_t^* with respect to α. However, exact computation of this conditional expectation would require a large number of p-dimensional integrations. Fahrmeir (1992) suggests that posterior expectations and covariance matrices be replaced by corresponding posterior modes $\hat{\beta}_{t|T}^{(k)}$ and error covariance matrices $\hat{V}_{t|T}^{(k)}$ obtained by the EKFS above, with the hyperparameters α set to $\alpha^{(k)}$. Proceeding in this way, one obtains an EM type algorithm, where the M step can be carried out analytically in many situations of practical interest.

For estimating hyperparameters, Fahrmeir and Goss (1992) consider the case of a univariate dynamic generalized linear model (2.1), with unknown vector of hyperparameters $\alpha = (a_0, Q_0, Q)$, i.e. $Q_t = Q$, Q_t is independent of t. The algorithm requires that initial values a_0, Q_0, and Q of the evolution model are known or given. The resulting iterative algorithm jointly estimates $\beta_{t|T}$, $V_{t|T}$, and α as follows:

1. Choose starting values a_0, Q_0, Q. Iterate the following steps 2 and 3 for $k = 0, 1, 2, \ldots$

2. Smoothing: compute $\beta_{t|T}^{(k)}$, $V_{t|T}^{(k)}$, $t = 1, \ldots, T$, by EKFS algorithm with unknown parameters replaced by their current estimates $a_0^{(k)}$, $Q_0^{(k)}$, $Q^{(k)}$.

3. EM step: Compute $a_0^{(k+1)}$, $Q_0^{(k+1)}$, $Q^{(k+1)}$ by

$$a_0^{(k+1)} = \beta_{0|T}^{(k)}, \quad Q_0^{(k+1)} = V_{0|T}^{(k)},$$

$$Q^{(k+1)} = \frac{1}{T}\sum_{t=1}^{T}[(\beta_{t|T}^{(k)} - \beta_{t-1|T}^{(k)})(\beta_{t|T}^{(k)} - \beta_{t-1|T}^{(k)})' + V_{t|T}^{(k)} + V_{t-1|T}^{(k)} - 2V_{t-1,t|T}^{(k)}].$$

with $V_{t-1,t|T}^{(k)} = V_{t-1|t-1}V_{t|t-1}^{-1}V_{t|T}$.

4. Stop when some termination criterion is reached.

As is common with EM-type algorithms, convergence cannot be assured generally

and can be slow. More details, in particular concerning derivations, implementation and performance, can be found in Goss (1990).

2.2.3 The Gibbs Sampler

Gibbs sampling is a Markov chain Monte Carlo technique for obtaining posterior densities in cases where traditional numerical integration techniques become infeasible. A Gibbs sampling approach to dynamic models with normal mixture error structure has been proposed by Carlin, Polson, and Stoffer (1992). Fahrmeir, Hennevogl and Klemme (1992) and Knorr-Held (1993) adapted their approach to dynamic generalized linear models with linear Gaussian evolution models. To obtain estimates for the marginal posterior densities $p(\beta_t|y_T^*)$ by Gibbs sampling, it is required that conditional posterior densities $p(\beta_t|\beta_{s\neq t}, y_T^*)$ of β_t given all other parameters β_s, $s \neq t$, be available for sampling. That means it has to be possible to draw random observations from these conditional densities. By definition, we have

$$p(\beta_t|\beta_{s\neq t}, y_T^*) = \frac{p(\beta_T^*, y_T^*)}{p(\beta_{s\neq t}, y_T^*)}. \tag{2.3}$$

Repeated application of the model assumptions (A1), (A2), and (A3) yields

$$p(\beta_T^*, y_T^*) = p(\beta_0) \prod_{t=1}^{T} p(\beta_t|\beta_{t-1}) \prod_{t=1}^{T} p(y_t|\beta_t, y_{t-1}^*).$$

Proceeding similarly for the denominator in (2.3), one obtains

$$p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{s\neq t}, \boldsymbol{y}_T^*) = \begin{cases} \frac{p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t)}{p(\boldsymbol{\beta}_{t+1})}, & \text{if } t = 0 \\ \frac{p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*)p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})}{p(\boldsymbol{y}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{y}_{t-1}^*)p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_{t-1})}, & \text{if } t = 1, \ldots, T-1 \\ \frac{p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*)p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})}{p(\boldsymbol{y}_t|\boldsymbol{\beta}_{t-1}, \boldsymbol{y}_{t-1}^*)}, & \text{if } t = T. \end{cases} \tag{2.4}$$

Since the denominators in (2.4) do not depend on $\boldsymbol{\beta}_t$, the following proportionality holds:

$$p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{s\neq t}, \boldsymbol{y}_T^*) \propto \begin{cases} p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t), & \text{if } t = 0 \\ p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*)p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}), & \text{if } t = 1, \ldots, T-1 \\ p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*)p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}), & \text{if } t = T. \end{cases} \tag{2.5}$$

Carlin, Polson, and Stoffer (1992) show that for linear Gaussian evolution models of the form (2.1), the proportionality (2.5) specializes to

$$p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{s\neq t}, \boldsymbol{y}_T^*) \propto \begin{cases} N(\boldsymbol{B}_t\boldsymbol{b}_t, \boldsymbol{B}_t), & \text{if } t = 0 \\ p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*)N(\boldsymbol{B}_t\boldsymbol{b}_t, \boldsymbol{B}_t), & \text{if } t = 1, \ldots, T. \end{cases} \tag{2.6}$$

with

$$\boldsymbol{B}_t^{-1} = \begin{cases} \boldsymbol{Q}_t^{-1} + \boldsymbol{F}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1}, & \text{if } t = 0, \ldots, T-1 \\ \boldsymbol{Q}_t^{-1}, & \text{if } t = T, \end{cases} \tag{2.7}$$

$$b_t' = \begin{cases} a_0' Q_0^{-1} + \beta_{t+1}' Q_{t+1}^{-1} F_{t+1}, & \text{if } t = 0 \\ \beta_{t-1}' F_t' Q_t^{-1} + \beta_{t+1}' Q_{t+1}^{-1} F_{t+1}, & \text{if } t = 1, \ldots, T-1 \\ \beta_{t-1}' F_t' Q_t^{-1}, & \text{if } t = T. \end{cases} \quad (2.8)$$

Note that one has to assume nonsingularity of Q_0 and Q_t. Fahrmeir, Hennevogl and Klemme (1992) use an EM-type algorithm (Section 2.2.2) to estimate the unknown hyperparameters a_0, Q_0, and $Q_t = Q$. Here, Fahrmeir, Hennevogl and Klemme only consider the situation when Q_t is independent of t. Or one can treat hyperparameters as a stochastic parameter with some prior distribution. This approach is illustrated in Chapter 4 and its application is in Chapter 5.

To obtain a random observation from the conditional density $p(\beta_t|\beta_{s\neq t}, y_T^*)$. Fahrmeir, Hennevogl and Klemme (1992) use rejection sampling (Devroye, 1986; Ripley, 1987). In the context of rejection sampling a random observation $\tilde{\beta}_t$ is drawn from a density g and accepted with probability $f(\tilde{\beta}_t)/(g(\tilde{\beta}_t)M_t)$ where $f(\tilde{\beta}_t)$ has to be proportional to $p(\beta_t|\beta_{s\neq t}, y_T^*)$ and the constant M_t has to be chosen so that

$$M_t g(\beta_t) \geq f(\beta_t) \qquad \text{for all } \beta_t. \quad (2.9)$$

In view of (2.6) one can set $f(\beta_t) = p(y_t|\beta_t, y_{t-1}^*)N(B_t b_t, B_t)$, and $g(\beta_t) = N(B_t b_t, B_t)$. Then the condition (2.9) corresponds to

$$M_t \geq p(y_t|\beta_t, y_{t-1}^*) \qquad \text{for all } \beta_t, \quad (2.10)$$

and the $N(B_t b_t, B_t)$-drawing $\tilde{\beta}_t$ is accepted if

$$u \leq p(y_t|\tilde{\beta}_t, y_{t-1}^*)/M_t,$$

where u denotes a uniformly distributed random number on [0,1]. For a multinomial

observation density $p(y_t|\beta_t, y_{t-1}^*)$ condition (2.10) is fulfilled by $M_t = 1$.

With the conditional densities (2.4) and rejection sampling, the Fahrmeir, Hennevogl and Klemme's (1992) Gibbs sampling procedure runs as follows: Given a set of arbitrary starting values $(\beta_t^{(0)})$, $t = 0, \ldots, T$ one has to draw $\beta_0^{(1)}$ from the conditional density $p(\beta_0|\beta_1^{(0)}, \ldots, y_T^{(0)}, y_T^*)$, then $\beta_1^{(1)}$ from $p(\beta_1|\beta_0^{(1)}, \beta_2^{(0)} \ldots, y_T^{(0)}, y_T^*)$, and so on up to $\beta_T^{(1)}$ from $p(\beta_T|\beta_0^{(1)}, \ldots, y_{T-1}^{(1)}, y_T^*)$, to complete one iteration. After K such iterations which define one Gibbs run, the $(T+1)$-tuple $(\beta_0^{(K)}, \beta_1^{(K)} \ldots, \beta_T^{(K)})$ is obtained. Under suitable regularity conditions (Smith and Roberts, 1993), the samples output from the Gibbs sampler can be used to mimic a random sample from the joint posterior density $p(\beta_0, \beta_1, \ldots, \beta_T|y_T^*)$. Carrying out G Gibbs runs yields $g = 1, \ldots, G$ *i.i.d.* $(T+1)$-tuples $(\beta_0^{(K,g)}, \beta_1^{(K,g)}, \ldots, \beta_T^{(K,g)})$. These can be used to estimate the marginal posterior density $p(\beta_t|y_T^*)$ by

$$\hat{p}(\beta_t|y_T^*) = \frac{1}{G}\sum_{g=1}^{G} p(\beta_t|\beta_{s\neq t}^{(K,g)}, y_T^*),$$

as long as the conditional density $p(\beta_t|\beta_{s\neq t}, y_T^*)$ is given in closed form. If no closed form expression is available, the moments $\beta_{t|T}$ and $V_{t|T}$ of the marginal posterior density $p(\beta_t|y_T^*)$ can be estimated by

$$\hat{\beta}_{t|T} = \frac{1}{G}\sum_{g=1}^{G} \beta_t^{(K,g)}, \qquad \hat{V}_{t|T} = \frac{1}{G}\sum_{g=1}^{G} (\beta_t^{(K,g)} - \hat{\beta}_{t|T})(\beta_t^{(K,g)} - \hat{\beta}_{t|T})'.$$

Gelfand and Smith (1990) argued that a slightly preferable estimate of this marginal posterior density is to use the sample path average, i.e.

$$\hat{p}(\beta_t|y_T^*) = \frac{1}{K}\sum_{k=1}^{K} p(\beta_t|\beta_{s\neq t}^{(k)}, y_T^*).$$

However, Fahrmeir, Hennevogl and Klemme (1992) apply the Gibbs sampler to estimate the posterior moments in a rainfall data (will be introduced in Section 5.1) and claim that

> ... it turned out that appropriate choice of starting values has a significant effect on the number of iterations required for one Gibbs run. We recommend using posterior mode smoothing estimates, which are easily obtained by generalized extended Kalman filtering and smoothing. Then, in contrast to arbitrarily chosen starting values, convergence of one Gibbs run takes only 20 to 40 iterations.

In particular, there are typically strong positive correlation between the consecutive Gibbs samples. Estimates of moments of the marginal posterior densities based on short Gibbs sampler runs with high correlations are not accurate (Geweke, 1992). And due to the strong dependency in the sequence the series may appear to have converged to stationarity even earlier. One needs larger samples than would be required if independent sampling were possible. A further analysis of this rainfall data using the MCMC algorithms is in Section 5.1.

2.3 Other Estimating Methods for the MDGLM

In the preceding section, we discussed three different estimation approaches based on posterior densities. Many other estimating methods are also available. Monte Carlo methods are a particularly attractive choice. In recent years, Markov chain Monte Carlo techniques have become very popular as a way of generating a sample from complicated probability distributions, such as posterior distributions in Bayesian inference problem. A generated posterior sample can then be used for virtually any posterior inference.

There are two basic algorithms for Markov chain Monte Carlo. Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) and the Metropolis-Hastings (M-H) algorithm (Metropolis *et al.*, 1953; Hastings, 1970). The Gibbs sampler is actually a special case of the M-H algorithm (Hastings, 1970; Gelman, 1992). Unlike the Gibbs sampler, the M-H algorithm updates several parameters at a time, rather than only one at each step. This is particularly useful for a higher dimensional conditional posterior distribution. Also unlike the Gibbs sampler, the M-H algorithm can sample from any posterior distribution by selecting a proposal density from which it is easy to sample, and is not restricted to posterior distributions where full conditionals are available for generation.

The M-H algorithm is very general, allowing a variety of useful special cases based on the different choices of the proposal density. Tierney (1994) presents a few examples of proposal densities that are useful for examining posterior distributions, such as Random Walk Chains, Independence Chains, Rejection Sampling Chains, Autoregressive Chains, and Grid-based Chains. In addition, the methods outlined here can be combined into hybrid strategies. These methods can be used to construct more efficient algorithms which will be explored in the next chapter.

Chapter 3

Markov Chain Monte Carlo Sampling

Starting with the work of Metropolis *et al.* (1953), Markov chain Monte Carlo methods have been widely used to solve problems in statistical physics and, more recently, Bayesian statistical inference. This chapter presents the basic methodology of Markov chain Monte Carlo methods, emphasizing the calculation of features of posterior distributions. We will introduce several algorithms, including Metropolis-Hastings algorithms, adaptive rejection algorithms and other variations to simulate multivariate distributions. Various Markov chain Monte Carlo methods and their applications can be found, for example, in Geyer (1992), Smith and Roberts (1993), Besag and Green (1993), and Tierney (1994).

3.1 Metropolis-Hastings Sampling Algorithms

The Metropolis-Hastings algorithm shares many features of the Gibbs sampler, but is more generally applicable, as it avoids any need to sample from difficult distributions. It can be applied to any Bayesian problem as long as it is possible to compute the ratio of the probabilities, or probability densities, of two states. Suppose that we wish to sample from the joint distribution for $\boldsymbol{x} = (x_1, \ldots, x_T)$, with respect to a distribution given by some density function, $\pi(\boldsymbol{x})$. Let $q(\boldsymbol{x},\boldsymbol{y})$ denote a candidate generating density, where $\int q(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y} = 1$. This density is to be interpreted as saying that when a

process is at the point $\boldsymbol{x}$ the density generates a value $\boldsymbol{y}$ from $q(\boldsymbol{x},\boldsymbol{y})$. The Metropolis-Hastings algorithm does this by repeatedly considering randomly generated changes to the components of $\boldsymbol{x}$, accepting or rejecting these changes based on how they affect the probability of the state. The algorithm can be described as follows:

1. Choose an arbitrary initial point $\boldsymbol{x}^{(0)} = (x_1^{(0)}, \ldots, x_T^{(0)})$, and set $t = 0$.

2. Generate a candidate $\boldsymbol{y}$ from the proposal distribution $q(\boldsymbol{x}^{(t)}, \cdot)$.

3. Set

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{y} & \text{with the probability } R(\boldsymbol{x}^{(t)}, \boldsymbol{y}), \\ \boldsymbol{x}^{(t)} & \text{otherwise,} \end{cases}$$

where

$$R(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left\{ \frac{\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{x}^{(t)})}{\pi(\boldsymbol{x}^{(t)})q(\boldsymbol{x}^{(t)}, \boldsymbol{y})}, 1 \right\}.$$

4. Set $t = t + 1$, and go to step 2.

Several remarks about this algorithm are as follows: (1) The M-H algorithm is specified by its candidate generating density $q(\boldsymbol{x},\boldsymbol{y})$. (2) If a candidate value is rejected, the current value is taken as the next value in the sequence. (3) The calculation of $R(\boldsymbol{x},\boldsymbol{y})$ does not require knowledge of the normalizing constant of $\pi(\cdot)$, since it appears both in the numerator and denominator. (4) If $\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{x}) \geq \pi(\boldsymbol{x})q(\boldsymbol{x},\boldsymbol{y})$, the chain moves to $\boldsymbol{y}$, otherwise it moves with probability given by $\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{x})/\pi(\boldsymbol{x})q(\boldsymbol{x},\boldsymbol{y})$.

Under mild regularity conditions, the joint distribution of this $\boldsymbol{x}^{(t)}$ can be shown to converge to the joint distribution of $\boldsymbol{x}$. The regularity conditions required are irreducibility and aperiodicity (e.g., Roberts and Smith, 1994). What these mean is that, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are in the domain of $\pi(\cdot)$, it must be possible to move from $\boldsymbol{x}$ to $\boldsymbol{y}$ in a finite number of iterations with nonzero probability and the number of moves required to move from $\boldsymbol{x}$ to $\boldsymbol{y}$ is not required to be a multiple of some integers. These

conditions are usually satisfied if $q(\boldsymbol{x},\boldsymbol{y})$ has a positive density on the same support as that of $\pi(\cdot)$. For more extensive theoretical results see Roberts and Smith (1994). Besag and Green (1993) and the recent overview by Tierney (1994).

These conditions, however, do not determine the rate of convergence (Roberts and Tweedie, 1994), so there is an empirical question of how much of the initial sample should be discarded and how long the sampler should be run. One possibility, due to Gelman and Rubin (1992), is to start multiple chains from dispersed initial values and compare the within and between variation of the sampled draws. This entire area, however, is quite unsettled. For further details see Gelman and Rubin (1992) and Geyer (1992).

In order to implement the M-H algorithm, it is necessary that a suitable proposal distribution be specified. Clearly, different specific choices of $q(\boldsymbol{x},\boldsymbol{y})$ will lead to different specific algorithms. Tierney (1994) provides a systematic taxonomy of the kinds of choice available. Two special cases are random walk chains and independence chains.

3.1.1 Random Walk Chains

One family of proposal distributions, which appears in the work of Metropolis *et al.* (1953), is given by $q(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{y} - \boldsymbol{x})$, where $f(\cdot)$ is a multivariate distribution. The candidate $\boldsymbol{y}$ is thus drawn according to the process $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$, where $\boldsymbol{z}$ is called the increment random variable and follows the distribution $f(\cdot)$. Since the candidate is equal to the current value plus noise, this case is called a random walk chain. Possible choices for $f(\cdot)$ include the multivariate normal distribution and multivariate t distribution. Note that when $f(\cdot)$ is symmetric, the usual circumstance.

$f(z) = f(-z)$. The probability of a move then reduces to

$$R = \min\{\frac{\pi(y)}{\pi(x)}, 1\}.$$

Roberts and Tweedie (1994) give a general theorem on geometric ergodicity of Metropolis samplers on R^d that iterate a single elementary update with a random walk proposal of the form $q(x,y) = f(y - x)$ where q is any density satisfying $f(x) = f(-x)$.

3.1.2 Independence Chains

Candidate steps y can also be chosen from a fixed density f. This option is discussed in Hastings (1970). In this case, $q(x,y) = f(y)$ and the acceptance probability R can be written as

$$R = \min\{\frac{w(y)}{w(x)}, 1\}$$

where $w(x) = \pi(x)/f(x)$. This function w is the importance weight function that would be used in importance sampling using $f(x)$ as an importance distribution. Thus, it is useful to choose f to produce a weight function that is bounded and as close to constant as possible. Possible choices for $f(\cdot)$ include multivariate t distributions, split-t distributions, or other distributions that have been found to be useful as importance distribution. Roberts and Tweedie (1994) show that an independence chain is geometrically ergodic if and only if $w(x)$ is bounded.

Although these are the main two examples that we consider in this thesis. other techniques, such as Grid-Based Chains, Rejection Sampling Chains. etc.. can clearly be applied as well.

3.2 Adaptive Rejection Algorithms

In contrast to the MCMC methods described above, a classical simulation technique to generate independent samples is the rejection sampling method. Although not an MCMC method, it uses some concepts that also appear in the Metropolis-Hastings algorithm and is a useful introduction to the topic.

The rejection algorithm (Devroye, 1986; Ripley, 1987) to sample from the target density $f(x)$ requires finding a envelope density $g(x)$ from which it is easy to sample and for which there exists an M such that $f(x)/g(x) \leq M$ for all x in D (where D denotes the domain of $f(x)$). Then the rejection algorithm proceeds as follows:

1. Sample a value y from $g(x)$ and u independently from uniform(0,1).

2. If $u \leq f(y)/[Mg(y)]$ then accept y and return to step 1. Otherwise reject y and return to step 1.

The rejection algorithm is only useful if it is more efficient or convenient to sample from the envelope density $g(x)$ than from the target density $f(x)$ itself. With a suitable envelope function the rejection algorithm can generate samples from the target density efficiently.

When applying the rejection algorithm within the Gibbs sampler, it may be very inefficient, since the conditional distribution for each parameter changes from iteration to iteration, it is difficult to find a suitable envelope density $g(x)$ and locate the mode of the target density in order to find c. Gilks and Wild (1992) proposed an adaptive version of rejection sampling (ARS), and Gilks (1992) further tailored it to produce a form called derivative free adaptive rejection sampling that both assume the target density $f(x)$ must be log-concave. The ARS proceeds as follows:

(1) Initialization Step

We assume that D is convex and $f(x)$ is continuous and differentiable in D and $h(x) = \ln f(x)$ is concave in D. Suppose that $h(x)$ and $h'(x)$ have been evaluated at k abscissae in D: $x_1 \leq x_2 \leq \ldots \leq x_k$. Let $T_k = \{x_i : i = 1, \ldots, k\}$. For $j = 1, \ldots, k-1$ the tangents at x_j and x_{j+1} intersect at

$$z_j = \frac{h(x_{j+1}) - h(x_j) - x_{j+1}h'(x_{j+1}) + x_j h'(x_j)}{h'(x_j) - h'(x_{j+1})}.$$

We define the rejection envelope on T_k as $\exp[u_k(x)]$, where $u_k(x)$ is a piecewise linear upper hull formed from the tangents to $h(x)$ at the abscissae in T_k. Thus, for $x \in [z_{j-1}, z_j]$ and $j = 1, \ldots, k$, we define

$$u_k(x) = h(x_j) + (x - x_j)h'(x_j) \tag{3.1}$$

where z_0 is the lower bound of D (or -∞ if D is not bounded below) and z_k is the upper bound of D (or $+\infty$ if D is not bound above). We also define the squeezing function on T_k as $\exp[l_k(x)]$, where $l_k(x)$ is a piecewise linear lower hull formed from the chords between adjacent abscissae in T_k. Thus for $x \in [x_j, x_{j+1}]$ and $j = 1, \ldots, k-1$, we define

$$l_k(x) = \frac{(x_{j+1} - x)h(x_j) + (x - x_j)h(x_{j+1})}{x_{j+1} - x_j}. \tag{3.2}$$

For $x < x_1$ or $x > x_k$ we define $l_k(x) =$ - ∞. Thus the rejection envelope and squeezing function are piecewise exponential functions. The concavity of $h(x)$ ensures that $l_k(x) \leq h(x) \leq u_k(x)$ for all x in D. Finally, we define

$$s_k(x) = \frac{\exp u_k(x)}{\int_D \exp u_k(x')dx'}. \tag{3.3}$$

If D is unbounded on the left then choose x_1 such that $h'(x_1) > 0$. If D is unbounded on the right then choose x_k such that $h'(x_k) < 0$. Having defined k starting abscissae, calculate the function $u_k(x)$, $l_k(x)$, and $s_k(x)$ from equations (3.1), (3.2), and (3.3) respectively. To sample m points independently from $f(x)$ by adaptive rejection sampling, perform the following sampling and updating steps alternately until m points have been accepted.

(2) Sampling Step

Sample a value y from $s_k(x)$, and independently u from uniform(0,1). Perform the squeezing test: if

$$u \leq \exp\{l_k(y) - u_k(y)\}$$

then accept y. Otherwise evaluate $h(y)$ and $h'(y)$ and perform the rejection test: if

$$u \leq \exp\{h(y) - u_k(y)\}$$

then accept y; otherwise reject y.

(3) Updating Step

If $h(y)$ and $h'(y)$ were evaluated at the sampling step, including y in T_k to form T_{k+1} and relabeling the elements of T_{k+1} in ascending order. And construct the functions $u_{k+1}(x)$, $l_{k+1}(x)$, and $s_{k+1}(x)$ from equations (3.1), (3.2), and (3.3) respectively on the basis of T_{k+1}. Return to the sampling step if m points have not yet been accepted.

The adaptive rejection sampling algorithm has two important features compared with other existing methods for generating independent observations from a probability density function. First, for generating from log-concave density functions and most universal random variate generating methods, such as rejection sampling or the ratio of uniforms methods, ARS does not require knowledge of the position of

the mode. Second, the rejection probability is decreasing as more random variates are sampled from the envelop function because with the addition of more points the density function is closer to the upper and lower functions used to squeeze it.

Gilks, Best, and Tan (1995) have proposed a further modification, adaptive rejection Metropolis sampling (ARMS) in which the sample y that has been accepted at step (2) is passed through an additional Metropolis-Hastings acceptance step. This algorithm removes the necessity for log-concavity of the target density because the Metropolis step corrects for violations of the envelope.

Now, we apply the ARS approach to Gibbs sampling. Suppose we wish to sample from the joint distribution for $\boldsymbol{x} = (x_1, \ldots, x_T)$. To implement the Gibbs sampler, initial values are assigned to each component, $x_i^{(0)}$, $i = 1, \ldots, T$. Then the Gibbs sampler repeatedly replaces each component with a value picked from its distribution conditional on the current values of all other components $[x_i|x_{j\neq i}]$, where $[\cdot|\cdot]$ denote a conditional distribution function. To apply adaptive rejection sampling to a Gibbs sampler, we require that the full conditional distribution $[x_i|x_{j\neq i}]$ is continuous, differentiable and log-concave with respect to x_i. Most commonly used densities are concave on the logarithmic scale with respect to both random variable and distributional parameters. However, when this is not so, the log-density may be concave with respect to a suitably transformed random variable or parameter (Gilks and Wild, 1992). Therefore, taking logarithms in $[x_i|x_{j\neq i}]$, the $h(x_i) = \ln[x_i|x_{j\neq i}]$ will be concave with respect to x_i. In these circumstances adaptive rejection sampling can be used to sample efficiently from $h(x_i)$.

Furthermore, if $\boldsymbol{x_i}$ is multivariate, then adaptive rejection sampling can still be used. For each component x_{ik} of $\boldsymbol{x_i}$, the univariate full conditional $[x_{ik}|\cdot]$ is proportional to the multivariate full conditional $[\boldsymbol{x_i}|\cdot]$. Therefore, if $[\boldsymbol{x_i}|\cdot]$ is log-concave with respect to $\boldsymbol{x_i}$, then $[x_{ik}|\cdot]$ will be log-concave with respect to x_{ik}. Thus, the Gibbs sampler can be implemented to update each component x_{ik} in turn using adaptive

rejection sampling with $h(x_{ik}) = \ln[x_{ik}|\cdot]$.

3.3 Auxiliary Variable Methods

Suppose that we wish to sample from the joint distribution for $\boldsymbol{x} = (x_1, \ldots, x_T)$ with respect to a distribution given by some density function, $\pi(\boldsymbol{x})$ for $\boldsymbol{x} \in X$. When high correlations among the components of $\boldsymbol{x}$ are present, the Markov chain Monte Carlo algorithms such as the Gibbs sampler that update component-wise using conditional distributions will converge very slowly. Auxiliary variable techniques (Besag and Green, 1993) have been recommended as a method of breaking correlation. These auxiliary variables enable us to design simple chains that make substantial changes to many components at once when these components display strong dependence.

In the method of auxiliary variables, the state variable $\boldsymbol{x}$ is augmented by one or more additional variables $\boldsymbol{u} \in U$. The joint distribution of $\boldsymbol{x}$ and $\boldsymbol{u}$ will be defined by taking the given distribution of interest $\pi(\boldsymbol{x})$ as the marginal for $\boldsymbol{x}$ and specifying the conditional $\pi(\boldsymbol{u}|\boldsymbol{x})$; for the moment this can be chosen quite arbitrarily. We write $\pi(\boldsymbol{x},\boldsymbol{u}) = \pi(\boldsymbol{x})\pi(\boldsymbol{u}|\boldsymbol{x})$, so that $\pi(\boldsymbol{x}|\boldsymbol{u}) \propto \pi(\boldsymbol{x},\boldsymbol{u})$. We now construct a Markov chain on $X \times U$ that alternates between two steps of transition: first, $\boldsymbol{u}$ is drawn from $\pi(\boldsymbol{u}|\boldsymbol{x})$; then, $\boldsymbol{x}$ is from $\pi(\boldsymbol{x}|\boldsymbol{u})$. Such an approach defines a valid MCMC procedure for $\pi(\boldsymbol{x})$ (Besag and Green, 1993).

The purpose of the constructions is effectively the Gibbs sampler applied to $\pi(\boldsymbol{x},\boldsymbol{u})$, with block updating of all of $\boldsymbol{u}$, then all of $\boldsymbol{x}$, alternately. That is, given $\boldsymbol{x}^{(t)}$, we first draw $\boldsymbol{u}^{(t)}$ from $\pi(\boldsymbol{u}|\boldsymbol{x}^{(t)})$, then draw $\boldsymbol{x}^{(t+1)}$ from $\pi(\boldsymbol{x}|\boldsymbol{u}^{(t)})$. When dealing with more complicated models, direct simulation from $\pi(\boldsymbol{x}|\boldsymbol{u})$ is unlikely to be available. The following construction introduced by Edwards and Sokal (1988)

provides an alternative possibility. Suppose that $\pi(\boldsymbol{x})$ can be written in the form

$$\pi(\boldsymbol{x}) \propto \pi_0(\boldsymbol{x}) \prod_k b_k(\boldsymbol{x}),$$

where $\pi_0(\boldsymbol{x})$ is a simple distribution. Then, if we introduce one auxiliary variable u_k for each interaction $b_k(\boldsymbol{x})$, and define $\pi(\boldsymbol{u}|\boldsymbol{x})$ to be the uniform distribution on the rectangle $\Pi_k[0, b_k]$, we have

$$\begin{aligned} \pi(\boldsymbol{x}, \boldsymbol{u}) &= \pi(\boldsymbol{x})\pi(\boldsymbol{u}|\boldsymbol{x}) \\ &= \pi_0(\boldsymbol{x}) \textstyle\prod_k b_k(\boldsymbol{x}) \{I[0 \le u_k \le b_k(\boldsymbol{x})] b_k(\boldsymbol{x})^{-1}\} \\ &= \pi_0(\boldsymbol{x}) I[\textstyle\bigcap_k \{0 \le u_k \le b_k(\boldsymbol{x})\}], \end{aligned}$$

where I is the indicator function. Thus $\pi(\boldsymbol{x}|\boldsymbol{u})$ is simply $\pi_0(\boldsymbol{x})$, conditional on the constraints $\{b_k(\boldsymbol{x}) \ge u_k\}$. That is, given $\boldsymbol{x}^{(t)}$, we first draw $\boldsymbol{u}$ from $\pi(\boldsymbol{u}|\boldsymbol{x}^{(t)})$, then draw $\boldsymbol{x}^{(t+1)}$ from $\pi_0(\boldsymbol{x})$, and impose the conditions $\{b_k(\boldsymbol{x}^{(t+1)}) \ge u_k\}$ by rejection. Moreover, the above equation demonstrates how auxiliary variables help to kill awkward interactions among components of $\boldsymbol{x}$ by introducing one auxiliary variable u_k for each interaction $b_k(\boldsymbol{x})$.

There is an interesting comparison that can be drawn here with three similar approaches. Suppose that we wish to sample from the joint distribution for $\boldsymbol{x} = (x_1, \ldots, x_T)$, with respect to a distribution given by some density function, $\pi(\boldsymbol{x})$.

1. The rejection sampling for $\pi(\boldsymbol{x})$, based on drawing from the envelope density $\pi_0(\boldsymbol{x})$ (Section 3.2), which produce independent samples but requires normalization of both π and π_0.

2. The Metropolis-Hastings algorithm for $\pi(\boldsymbol{x})$, using $\pi_0(\boldsymbol{x})$ as proposal distribution (Section 3.1), which produces dependent samples but does not require

knowledge of distribution π and can sample from any suitable family of distribution.

3. The data augmented algorithm (e.g., Tanner and Wong, 1987; Albert and Chib, 1993) for $\pi(\boldsymbol{x})$, based on sampling from the augmented data posterior $\pi_0(\boldsymbol{x}|\boldsymbol{u})$, where $\boldsymbol{u}$ is a latent variable. The data augmented algorithm exploits the simplicity of the posterior distribution of the parameter given the augmented data.

Chapter 4

MCMC Samplers for Binary Data

In this chapter, we provide a detailed, introductory exposition of the M-H algorithm for MDGLM's. We derive a modified version of adaptive rejection sampling to apply the Gibbs sampler to full conditional densities of MDGLM's. in particular. for lower dimension situations. We also discuss applications of M-H algorithms. random walk chains and independence chains to MDGLM's and discuss issues related to the choice of the candidate generating densities and implementation. A block M-H algorithm is introduced for situations when the acceptance rate is near zero. Three link functions, probit, logistic, and mixtures of normal distributions will be introduced and the connection among them will be explored in an example in the next chapter. Applications of the methods illustrated in this chapter to univariate dynamic generalized linear models will be examined in the Chapter 5; applications to multivariate dynamic generalized linear models will be examined in the Chapter 6.

4.1 A Modified Version of ARS

Gibbs sampling is a MCMC technique for drawing dependent samples from complex distributions. In the Bayesian context, these distributions are usually posterior distributions of the model parameters, and samples produced by the Gibbs sampler can be used straightforwardly for Bayesian inference. At each iteration of the Gibbs sampler. each parameter or set of parameters is updated in turn by sampling a new value from its full conditional distribution. The full conditional distribution of a parameter is its

distribution conditional on the data and on the current values of all the other parameters. Thus, from one iteration to the next, full conditional distributions change as the conditioning parameters change. So methods for constructing full conditional distributions and for sampling from them must be very efficient. Certain full conditionals reduce analytically to well known distributions, for which special methods for efficient random variate generation are available. More usually no analytical reduction is possible. Gilks and Wild (1992) show that in practice full conditional distributions are often log-concave, and proposed the ARS method for efficiently sampling from univariate log-concave distributions. However not all models of practical importance yield log-concave full conditionals. One such example is the non-linear mixed effect model. Gilks *et al.* (1995) extend the ARS to deal with distributions that are not log-concave by appending a Metropolis-Hastings algorithm step.

4.1.1 Modified ARS within the Gibbs Sampler

As shown in equation (2.5), the full conditional density of univariate dynamic generalized linear models ($q = 1$) under the usual assumptions (A1), (A2), and (A3) is proportional to

$$p(\beta_t|\beta_{s\neq t}, y_T^*, x_T^*) \propto p(y_t|\beta_t, y_{t-1}^*, x_t^*)p(\beta_{t+1}|\beta_t)p(\beta_t|\beta_{t-1}).$$

Instead of making the effort to reduce the full conditional densities analytically to well known distributions, for which standard algorithms are available to generate random variates or to figure out the log-concave property of full conditional densities, for which ARS or ARMS is available to generate random variates, we propose a modified version of ARS in the application of Gibbs sampling. The key is provided by the following.

We assume that the target density, the full conditional density $f^*(x)$, can be

written as $f^*(x) \propto f(x)r(x)$, where $f(x)$ is a log-concave function and $r(x)$ is a density function such that efficient random variate generation is possible from $e^{u(x)}r(x)$, where $u(x)$ is a piecewise linear upper hull for $f(x)$ constructed by the ARS method. Then our modified algorithm proceeds as follows:

(1)′ Initialization Step

We assume that D^* is convex, where D^* denotes the domain of $f^*(x)$, and $f(x)$ is continuous and differentiable in D^* and $h(x) = \ln f(x)$ is concave in D^*. Also assume $h^*(x) = \ln f^*(x) = h(x) + \ln r(x)$. Suppose that $h(x)$ and $h'(x)$ have been evaluated at x_1, where $x_1 \in D^*$. Let $T_1 = \{x_1\}$. We define the rejection envelope on T_1 as $\exp[u_1(x)]$, where $u_1(x)$ is a linear upper hull formed from the tangents to $h(x)$ at the abscissae x_1. Thus, for $x \in [z_0, z_1]$ we define

$$u_1(x) = h(x_1) + (x - x_1)h'(x_1), \tag{4.1}$$

where z_0 is the lower bound of D^* (or -∞ if D^* is not bounded below) and z_1 is the upper bound of D^* (or $+\infty$ if D^* is not bound above). Let $u_1^*(x) = u_1(x) + \ln r(x)$. The concavity of $h(x)$ ensures that $h(x) \leq u_1(x)$ for all x in D^*, thus, $h^*(x) = h(x) + \ln r(x) \leq u_1(x) + \ln r(x) = u_1^*(x)$. Finally, we define

$$s_1^*(x) = \frac{\exp u_1^*(x)}{\int_{D^*} \exp u_1^*(x')dx'} \propto e^{u_1(x)}r(x). \tag{4.2}$$

To sample a point independently from $f^*(x)$, first, we sample a value y from $s_1^*(x)$, and independently u from uniform(0,1), then perform the rejection test: if

$$u \leq \exp\{h^*(y) - u_1^*(y)\}$$

then accept y; otherwise perform the following updating and sampling steps until accepted. Set $k = 1$.

(2)′ Updating Step

If y was rejected at the sampling step, including y in T_k to form T_{k+1} and relabeling the elements of T_{k+1} in ascending order (i.e., $x_1 \leq x_2 \leq \ldots \leq x_{k+1}$). For $j = 1, \ldots, k$ the tangents at x_j and x_{j+1} intersect at

$$z_j = \frac{h(x_{j+1}) - h(x_j) - x_{j+1}h'(x_{j+1}) + x_j h'(x_j)}{h'(x_j) - h'(x_{j+1})}.$$

And construct the functions $u^*_{k+1}(x)$, $l^*_{k+1}(x)$, and $s^*_{k+1}(x)$ on the basis of T_{k+1}. Thus for $x \in [z_{j-1}, z_j]$ and $j = 1, \ldots, k+1$, we define

$$u^*_{k+1}(x) = h(x_j) + (x - x_j)h'(x_j) + \ln r(x) = u_{k+1}(x) + \ln r(x).$$

where z_0 is the lower bound of D^* and z_{k+1} is the upper bound of D^*. We also define

$$s^*_{k+1}(x) = \frac{\exp u^*_{k+1}(x)}{\int_{D^*} \exp u^*_{k+1}(x')dx'} \propto e^{u_k(x)} r(x).$$

Finally, for $x \in [x_j, x_{j+1}]$ and $j = 1, \ldots, k$, we define

$$l^*_{k+1}(x) = \frac{(x_{j+1} - x)h(x_j) + (x - x_j)h(x_{j+1})}{x_{j+1} - x_j} + \ln r(x) = l_{k+1}(x) + \ln r(x).$$

For $x < x_1$ or $x > x_{k+1}$ we define $l^*_{k+1}(x) = -\infty$.

(3)′ Sampling Step

Sample a value y from $s^*_{k+1}(x)$, and independently u from uniform(0,1). Perform the squeezing test: if

$$u \leq \exp\{l^*_{k+1}(y) - u^*_{k+1}(y)\}$$

then accept y. Otherwise evaluate $h(y)$ and $h'(y)$ and perform the rejection test: if

$$u \leq \exp\{h^*(y) - u^*_{k+1}(y)\}$$

then accept y; otherwise reject y. Return to the updating step if a point has not yet been accepted.

4.1.2 Proof of the Modified Version of ARS

The proof that modified adaptive rejection sampling leads to independent samples from $f^*(x)$ is in the following. Let y denote the nth sampled value of y^*, whether or not it was accepted or included in T_k. Let

$$\delta_n = \begin{cases} 1 & \text{if } y \text{ was accepted at the squeezing or rejection test.} \\ 2 & \text{if } y \text{ was rejected.} \end{cases} \tag{4.3}$$

Let H_n denote the history of the process, up to and including the processing of y: so $H_n = \{(y, \delta_i); i = 1, \ldots, n\}$. Thus H_n defines the current upper and lower hulls. Let $[\cdot|\cdot]$ denote a conditional probability density function and U be an independent uniform $[0, 1]$ random variable. Then

$$\begin{aligned} P[\delta_{n+1} = 1|H_n] &= P[U \leq \exp\{h^*(x) - u_k^*(x)\}|H_n] \\ &= \int_{D^*} \frac{\exp h^*(x)}{\exp u_k^*(x)} \frac{\exp u_k^*(x)}{\int_{D^*} \exp u_k^*(x')dx'} dx \\ &= \frac{\int_{D^*} \exp h_k^*(x)dx}{\int_{D^*} \exp u_k^*(x')dx'} \end{aligned}$$

and so

$$\begin{aligned} P[(y = y^*)|H_n \cap (\delta_{n+1} = 1)] &= \frac{P[(y=y^*) \cap (\delta_{n+1}=1)|H_n]}{P[\delta_{n+1}=1|H_n]} \\ &= \frac{\exp h_k^*(y^*)/\int_{D^*} \exp u_k^*(x')dx'}{\int_{D^*} \exp h_k^*(x)dx/\int_{D^*} \exp u_k^*(x')dx'} \\ &= \exp h^*(y^*)/\int_{D^*} \exp[h_k^*(x)]dx \\ &= f^*(y^*) \end{aligned}$$

which does not depend on H_n. Thus accepted values of y are drawn independently from $f^*(y)$.

4.1.3 Modified ARS and DGLM

Since the full conditional density of univariate dynamic generalized linear models under the usual assumptions (A1), (A2), and (A3) is proportional to

$$p(\beta_t|\beta_{s\neq t}, y_T^*, x_T^*) \propto p(y_t|\beta_t, y_{t-1}^*, x_t^*)p(\beta_{t+1}|\beta_t)p(\beta_t|\beta_{t-1}),$$

the assumption of the modified version of ARS is usually satisfied. The modified version of ARS we propose, requiring only that the likelihood function, $p(y_t|\beta_t, y_{t-1}^*, x_t^*)$ is log-concave and $p(\beta_{t+1}|\beta_t)p(\beta_t|\beta_{t-1})$ is a density function, is generally applicable to the univariate dynamic generalized linear models. Even if a non-log-concave density is present in the form of the posterior densities, the ARMS can be used to implement this modified ARS, in which the sample y that has been accepted at step (3)′ is passed through an additional Metropolis-Hastings acceptance step. However, when applied to high dimensional observations, it could be very complex and difficult to write a problem-specific code for generating variates from the model.

Compared to the ARS, the modified version of ARS has a number of advantages. First, the modified ARS only needs to start at one point instead of two points in the ARS algorithm and therefore, thus simplifying the ARS algorithm. Second, applying this modified ARS within Gibbs sampling requires simulation mainly from standard distributions such as multivariate normal and therefore is easy to implement in many statistical computer languages. A nice feature of this modified ARS is that if $r(x)$ is from an exponential family, then $e^{u(x)}r(x)$ is also a piecewise exponential family. For linear Gaussian evolution models of the form (2.1), the $r(x)$ is a normal density and the $e^{u(x)}r(x)$ is only a shift the location of $r(x)$ for each piece because

of $u(x)$ is piecewise linear in x. To illustrate this feature, consider the analysis of univariate dynamic generalized models with linear Gaussian evolution models. In the plain Gibbs sampler, a sample $\tilde{\beta}_t$ is generated from $N(\mu, \sigma^2)$ and accepted if $u \leq p(y_t|\tilde{\beta}_t)$, where u denotes a uniformly distributed random number. For applying the Modified ARS algorithm within the plain Gibbs sampler, a sample $\tilde{\beta}_t$ is generated from $N(\mu^*, \sigma^2)$ and accepted if $u \leq \exp\{h^*(\tilde{\beta}_t) - u^*(\tilde{\beta}_t)\}$, where $\mu^* = \mu + \sigma^2 h'(\beta_t)$. h^*, u^*, and h' are as indicated in the above section, and β_t is the accepted sample value of the previous Gibbs run. Finally, in the nice form of $r(x)$, the calculation of $s_k^*(x)$ can be omitted and samples can be generated from $r(x)$ with minor adjustment. This simplification is especially attractive when $s_k^*(x)$ is complex and difficult to be sampled from.

4.2 Use of M-H Algorithms

In order to implement the M-H algorithm, it is necessary that a suitable candidate generating density be specified. Typically, this density is selected from a family of distributions that requires the specification of such tuning parameters as the location and spread. This is an important matter that has implications for the efficiency of the algorithm. The spread of the candidate generating density affects the behavior of the chain in two ways: one is the acceptance rate, and the other is the region of the sample space that is covered by the chain. If the current value is around the mode and the spread is extremely large, the generated candidate will be far from the current value and therefore have a low probability of being accepted. But if the spread is chosen too small, the chain will take longer to traverse the domain of the density, and low probability regions will be under-sampled. Both of these situations are likely to be reflected in high autocorrelations across sample values.

For the MDGLM, the posterior density under assumptions (A1), (A2), and

(A3) is

$$\pi(\boldsymbol{\beta}_T^*) \propto \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{y}_{t-1}^*, \boldsymbol{x}_t^*) \prod_{t=1}^{T} p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}) p(\boldsymbol{\beta}_0).$$

where $\boldsymbol{\beta}_t$ has dimension p. We now illustrate the use of different candidate generating densities in two special cases of M-H algorithms, random walk chains and independence chains, to generate samples.

4.2.1 Random Walk Chains

The candidate $\boldsymbol{y}$ is drawn according to the process $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$, where $\boldsymbol{z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_T')'$ is the increment random variable and follows the candidate generating distribution $f(\cdot)$. If $f(\cdot)$ is symmetric, the probability of a move then reduces to

$$R = \min\{\frac{\pi(\boldsymbol{x}+\boldsymbol{z})}{\pi(\boldsymbol{x})}, 1\}.$$

Several easily generated candidate densities are suggested in the following, for which the parameters need to be adjusted by experimentation to achieve an optimal acceptance rate:

1. The increment random variable $\boldsymbol{z}_i'$ is distributed as multivariate uniform, i.e., the jth component of $\boldsymbol{z}_i'$ is uniform on the interval $(-\delta_{ij}, \delta_{ij})$, $j = 1, \ldots, p$. Note that δ_{ij} controls the spread along the coordinate axis.

2. The increment random variable $\boldsymbol{z}_i'$ is distributed as independent multivariate normal $N_p(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \text{diagonal}(\sigma_{i1}^2, \ldots, \sigma_{ip}^2)$.

3. The increment random variables $\boldsymbol{z}_0'$ and $\boldsymbol{z}_i'$ are distributed as

$$\boldsymbol{z}_0 \sim N_p[\mathbf{0}, \boldsymbol{\Sigma}/(1-\rho^2)], \quad \boldsymbol{z}_i|\boldsymbol{z}_{i-1} \sim N_p[\rho \boldsymbol{z}_{i-1}, \boldsymbol{\Sigma}], \quad i = 1, \ldots, T.$$

where ρ and Σ can be approximated by $\mathrm{Cov}(\boldsymbol{\beta_i},\boldsymbol{\beta_{i-1}})$ and $\mathrm{Var}(\boldsymbol{\beta_i})$, respectively. Although these are the only three examples that we list here, other techniques, for example, Split-t distributions (Geweke, 1989), the Metropolisized hit-and-run sampler (Chen and Schmeiser, 1993), and random walk Metropolis chains with an increment density that is symmetric about the origin (Müller, 1991) can be applied as well.

Recent work by Roberts, Gelman, and Gilks (1994) discusses the spread of the candidate generating density issue in the context of the random walk chain. They show that if the target and candidate generating densities are normal, then the spread of the latter should be tuned so that the acceptance rate is approximately 0.45 in 1-dimension and approximately 0.23 as the number of dimensions approaches infinity, with the optimal acceptance rate being around 0.25 in as low as 6 dimensions. It is important to mention that a chain with the optimal acceptance rate may still display high autocorrelations. The high serial correlations with the random walk chain are not unexpected and stem from the long-memory in the candidate draws. In such circumstances it is usually worth trying a different family of candidate generating densities.

4.2.2 Independence Chains

The candidate $\boldsymbol{y}$ is drawn from a fixed density $q_2(\boldsymbol{y})$. Then the probability of a move reduces to

$$R = \min\{\frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}, 1\},$$

where $w(\boldsymbol{x}) = \pi(\boldsymbol{x})/q_2(\boldsymbol{x})$. As in the random walk chain, we can let q_2 be a multivariate normal or multivariate-t density, but now it is necessary to specify the location of the generating density as well as the spread.

The choice of spread of the proposal density in the independence chain is important to ensure that the tails of the candidate generating density dominate those

of the target density, which is similar to a requirement on the importance sampling function in Monte Carlo integration with importance sampling. Thus, it is useful to chose q_2 to produce a weight function, w, that is bounded and as close to constant as possible. If the weight function is constant, then the chain produces an *i.i.d.* sample from π since the algorithm never rejects candidate steps. As a result, if we can choose the candidate generating density not too different in shape from the posterior distribution, then the independence chain usually performs very well.

One possibility is the independence chain with candidate generated by a multivariate normal distribution with mean and covariance matrix estimated by preliminary samples generated by other M-H algorithms, i.e., $q_2 \sim N(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\Sigma}}$ are the estimated posterior mean and covariance. Another possibility is that $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\Sigma}}$ are the posterior mode and negative inverse Hessian evaluated at $\tilde{\boldsymbol{\theta}}$.

4.2.3 Block-At-A-Time M-H Algorithms

The problem of a low acceptance rate becomes serious when the dimension of the posterior distribution is large. One possible strategy is to apply the M-H algorithm to sub-blocks of the $\boldsymbol{\beta}_T^*$, rather than simultaneously to all components of the vector. This strategy is discussed in Hastings (1970). Suppose that $\boldsymbol{\beta}_T^* = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T')'$, is divided into k blocks $\boldsymbol{\beta}_T^* = (\boldsymbol{\phi}_1', \ldots, \boldsymbol{\phi}_k')'$, $1 < k \leq T$, and the existence of conditional transition kernel $P_{i|j\neq i}(\boldsymbol{\phi}_i|\boldsymbol{\phi}_j, j \neq i)$ with the property that it is invariant to the conditional distributions $\pi_{i|j\neq i}(\boldsymbol{\phi}_i|\boldsymbol{\phi}_j, j \neq i)$. Any kernel with this property will have invariant distribution $\pi(\boldsymbol{\beta}_T^*)$ (e.g., Tierney, 1994; Chib and Greenberg, 1995). In particular, the product of the transition kernels has $\pi(\boldsymbol{\beta}_T^*)$ as its invariant distribution.

With this result, several important features of the M-H algorithm can be performed. A special case is the Gibbs sampler algorithm when the block size is one. This block strategy gives rise to several interesting hybrid algorithms obtained by

combining M-H updates. Several ways to form a hybrid strategy are to use conditioning, mixtures and cycles; see Tierney (1994) for further details. An example is Zeger and Karim (1991). Suppose $\boldsymbol{\beta}$ can be split into two components $(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$, and sampling from $\boldsymbol{\beta_1}|\boldsymbol{\beta_2}$ is available but sampling from $\boldsymbol{\beta_2}|\boldsymbol{\beta_1}$ is not available. Then we can apply the Gibbs sampler to $\boldsymbol{\beta_1}|\boldsymbol{\beta_2}$ and a M-H algorithm to $\boldsymbol{\beta_2}|\boldsymbol{\beta_1}$. Another useful strategy is that in each block the M-H algorithm can take different forms according to the problems need or the characteristics of the conditional densities.

The cost of using blocks lies in the decreased dependence between successively samples although this can increase the iteration time significantly. To obtain an optimal compromise between serial correlation in samples and acceptance rate is not as easy as we might expect. Jensen et al. (1993) present heuristic guidelines to obtain an optimal compromise based on the outcome of their study. However, these guidelines based on their empirical investigation are not as clear as we might wish. To apply to other areas of interest, further investigation is necessary.

4.3 Link Functions

Different choices of the link function h in the dynamic generalized linear models lead to different models. For example, the most popular link function for binary responses is the logit, which corresponds to a choice of a logistic distribution for h. The other commonly used model is the probit model obtained by taking $h = \Phi$, where Φ represents the standard normal cumulative distribution function. Albert and Chib (1993) investigated the sensitivity of the posterior density estimates to the different link functions by using Gibbs sampling in an example and the results suggested that inferences can be sensitive to the choice of link function. Their analysis suggested that the choice of link function can make a difference and that it is worthwhile to consider a variety of link functions. In the following, several link functions will be

introduced. The connection among these link functions will be further explored in the example in Chapter 5.

4.3.1 Probit Model

The probit model is obtained if h is the standard normal cumulative distribution function. To apply MCMC algorithms for computing the posterior distribution, we introduce the data augmentation approach. The key idea is to introduce T independent latent variables $w_1, \ldots, w_T$ into the model. Consider the univariate dynamic generalized linear models,

$$E(y_t|\boldsymbol{\beta}_t, y_1, \ldots, y_{t-1}) = \pi_t = h(\boldsymbol{z}_t'\boldsymbol{\beta}_t), \quad \boldsymbol{\beta}_t = \boldsymbol{F}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t, \quad t = 1, \ldots, T.$$

where $\boldsymbol{\xi}_t$ is Gaussian noise, $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t)$ and $\boldsymbol{\beta}_0 \sim N(\boldsymbol{a}_0, \boldsymbol{Q}_0)$. Introduce T latent variables $w_1, \ldots, w_T$, where the w_t are independent $N(\boldsymbol{z}_t'\boldsymbol{\beta}_t, 1)$ and define

$$y_t = \begin{cases} 1, & \text{if } w_t > 0, \\ 0, & \text{if } w_t \leq 0. \end{cases} \tag{4.4}$$

It can be easily shown that the y_t are independent Bernoulli random variables with $\pi_t = P(y_t = 1) = \Phi(\boldsymbol{z}_t'\boldsymbol{\beta}_t)$.

The joint posterior distribution of the $\boldsymbol{\beta}_T^* = (\boldsymbol{\beta}_0', \ldots, \boldsymbol{\beta}_T')'$ and $\boldsymbol{w} = (w_1, \ldots, w_T)$ given the observations $\boldsymbol{y} = (y_1, \ldots, y_T)$ is complicated. However, the computation of the posterior distribution of $\boldsymbol{\beta}_T^*$ using the Gibbs sampler requires only the posterior distribution of $\boldsymbol{\beta}_t$ conditional on $\boldsymbol{w}$, and $\boldsymbol{\beta}_{s \neq t}$ and the posterior distribution of $\boldsymbol{w}$ conditional on $\boldsymbol{\beta}_T^*$ and $\boldsymbol{y}$. The data augmentation approach leads to simple forms for these full conditional posterior distributions. In the simplest form, the $\boldsymbol{z}_t'$, $\boldsymbol{F}_t$, $\boldsymbol{Q}_t$, $\boldsymbol{a}_0$, and $\boldsymbol{Q}_0$ are assumed to be deterministic and known. Then, the posterior distribution

of w_t conditional on $\boldsymbol{\beta}_T^*$ and $\boldsymbol{y}$ is given by

$$w_t|\boldsymbol{\beta}_T^*, \boldsymbol{y} \propto \begin{cases} N(\boldsymbol{z}_t'\boldsymbol{\beta}_t, 1)I_{(0,\infty)}, & \text{if } y_t = 1, \\ N(\boldsymbol{z}_t'\boldsymbol{\beta}_t, 1)I_{(-\infty,0)}, & \text{if } y_t = 0. \end{cases} \tag{4.5}$$

Carlin, Polson, and Stoffer (1992) show that for the Gaussian noise $\boldsymbol{\xi}_t$ and $\boldsymbol{\beta}_0$ of the univariate dynamic generalized linear models, the posterior distribution of $\boldsymbol{\beta}_t$ conditional on $\boldsymbol{\beta}_{s\neq t}$, and $\boldsymbol{w}$ is given by

$$\boldsymbol{\beta}_t|\boldsymbol{\beta}_{s\neq t}, \boldsymbol{w} \propto N(\boldsymbol{B}_t\boldsymbol{b}_t, \boldsymbol{B}_t), \quad t = 0, \ldots, T, \tag{4.6}$$

with

$$\boldsymbol{B}_t^{-1} = \begin{cases} \boldsymbol{Q}_0^{-1} + \boldsymbol{F}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1}, & \text{if } t = 0 \\ \boldsymbol{Q}_t^{-1} + \boldsymbol{F}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1} + \boldsymbol{z}_t\boldsymbol{z}_t', & \text{if } t = 1, \ldots, T-1 \\ \boldsymbol{Q}_t^{-1} + \boldsymbol{z}_t\boldsymbol{z}_t', & \text{if } t = T, \end{cases}$$

$$\boldsymbol{b}_t' = \begin{cases} \boldsymbol{a}_0'\boldsymbol{Q}_0^{-1} + \boldsymbol{\beta}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1}, & \text{if } t = 0 \\ \boldsymbol{\beta}_{t-1}'\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1} + \boldsymbol{\beta}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1} + w_t\boldsymbol{z}_t', & \text{if } t = 1, \ldots, T-1 \\ \boldsymbol{\beta}_{t-1}'\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1} + w_t\boldsymbol{z}_t', & \text{if } t = T. \end{cases}$$

We can treat the hyperparameters $\boldsymbol{a}_0$, $\boldsymbol{Q}_0$, and $\boldsymbol{Q}_t$ as unknown constants and estimate these values by usual estimation procedures, such as the EM algorithm or generalized least squares. To implement the Gibbs sampler, we start with initial guesses of $\boldsymbol{a}_0$, $\boldsymbol{Q}_0$, and $\boldsymbol{Q}_t$, simulate the w_t from (4.5), and then simulate $\boldsymbol{\beta}_t$ from the distribution (4.6). Or the hyperparameters $\boldsymbol{a}_0$, $\boldsymbol{Q}_0$, and $\boldsymbol{Q}_t$ can be assumed independent with a convenient prior specifications (Section 4.2.4).

4.3.2 Mixtures of Normal Distributions

Since the posterior distribution of $\boldsymbol{\beta}_t$ given $\boldsymbol{w}$, and $\boldsymbol{\beta}_{s\neq t}$ is multivariate normal, it is possible to generalize this model by applying suitable mixtures of normal distributions. This approach was illustrated in Albert and Chib (1993) by consideration of t-link and hierarchical models. By plotting quantiles of the logistic distribution against quantiles of a t distribution for various degrees of freedom, Albert and Chib (1993) found the logistic quantiles are approximately a linear function of the quantiles of t distribution with eight degrees of freedom. Thus, one can view the logistic distribution as an approximate member of the t family. This generalization allows one to investigate the sensitivity of the posterior distribution of the probabilities π_t to the choice of link function. For example, by using mixtures of normals, one can generalize the probit link by choosing h to be the family of t distributions. By inspection of various posterior distributions, one can investigate the sensitivity of the fitted probabilities with respect to various t links and also see which value of the degrees of freedom is best supported by the observations.

We let the w_t be t random variable with locations $\boldsymbol{z}_t'\boldsymbol{\beta}_t$, scale parameter 1, and degrees of freedom v. By introducing the additional random variables λ_t, we can write the distribution of w_t as the following scale mixture of a normal distribution:

$$w_t|\lambda_t \propto N(\boldsymbol{z}_t'\boldsymbol{\beta}_t, \lambda_t^{-1}), \quad \lambda_t \propto \text{ Gamma } (v/2, 2/v), \quad t = 1, \ldots, T.$$

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)$ be the vector of scale parameters and $f(v)$ prior density on v. Similar to the previous section, the full conditional distributions of $\boldsymbol{w}$, $\boldsymbol{\beta}_T^*$, $\boldsymbol{\lambda}$, and v are given below:

1. The full conditional distributions of $w_1, \ldots, w_T$ are independent with

$$w_t|\boldsymbol{\beta}_T^*, \boldsymbol{y}, \boldsymbol{\lambda}, v \propto \begin{cases} N(\boldsymbol{z}_t'\boldsymbol{\beta}_t, \lambda_t^{-1})I_{(0,\infty)}, & \text{if } y_t = 1. \\ N(\boldsymbol{z}_t'\boldsymbol{\beta}_t, \lambda_t^{-1})I_{(-\infty,0)}, & \text{if } y_t = 0. \end{cases} \tag{4.7}$$

2. The fully conditionally distributions of $\boldsymbol{\beta}_t$ are given by

$$p(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{s\neq t}, \boldsymbol{w}, \boldsymbol{\lambda}, v) \propto N(\boldsymbol{B}_t\boldsymbol{b}_t, \boldsymbol{B}_t), \quad t = 0, \ldots, T. \tag{4.8}$$

with

$$\boldsymbol{B}_t^{-1} = \begin{cases} \boldsymbol{Q}_t^{-1} + \boldsymbol{F}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1} + \lambda_t\boldsymbol{z}_t\boldsymbol{z}_t', & \text{if } t = 0, \ldots, T-1 \\ \boldsymbol{Q}_t^{-1} + \lambda_t\boldsymbol{z}_t\boldsymbol{z}_t', & \text{if } t = T. \end{cases}$$

$$\boldsymbol{b}_t' = \begin{cases} \boldsymbol{a}_0'\boldsymbol{Q}_0^{-1} + \boldsymbol{\beta}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1}, & \text{if } t = 0 \\ \boldsymbol{\beta}_{t-1}'\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1} + \boldsymbol{\beta}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1} + w_t\lambda_t\boldsymbol{z}_t', & \text{if } t = 1, \ldots, T-1 \\ \boldsymbol{\beta}_{t-1}'\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1} + w_t\lambda_t\boldsymbol{z}_t', & \text{if } t = T. \end{cases}$$

3. $\lambda_1, \ldots, \lambda_T|\boldsymbol{w}, \boldsymbol{\beta}_T^*, v$ are independent with

$$\lambda_t \propto \text{ Gamma } \left(\frac{v+1}{2}, \frac{2v}{v + (w_t - \boldsymbol{z}_t'\boldsymbol{\beta}_t)^2}\right). \tag{4.9}$$

4. $v|\boldsymbol{w}, \boldsymbol{\beta}_T^*, \boldsymbol{\lambda}$ is distributed according to the density proportional to

$$f(v)\prod_{t=1}^{T}\{[\Gamma(v/2)(2/v)^{(v/2)}]^{-1}\lambda_t^{v/2-1}e^{-v\lambda_t/2}\}. \tag{4.10}$$

To implement the Gibbs sampler, we start with initial guesses of $\boldsymbol{a}_0$, $\boldsymbol{Q}_0$, and $\boldsymbol{Q}_t$, and cycle through the conditional distributions (4.7), (4.8), (4.9), and (4.10) in that order. To simulate from the full conditional distribution in equation (4.8), we can apply the

modified ARS algorithm described in Section 4.1.1. In practice, we are interested in the posterior probabilities for v in a finite set and it is then easy to simulate from the discrete distribution (4.10).

Another suitable mixtures of normal distribution is suggested by Andrews and Mallows (1974). It is established that when K has the asymptotic distribution of the Kolmogorov distance statistic, $2ZK$ is logistic. To use this result we have to be able to generate random variables having the asymptotic Kolmogorov distribution. One possibility is to use the relation found by Watson (1961)

$$2K^2 = \sum_{j=1}^{\infty} W_j/j^2,$$

where $W_1, W_2, \ldots$ are independent unit exponential variables.

4.3.3 Auxiliary Variable Methods

Although the logit model is a popular model for binary responses, the generation of the required samples for statistical inference may not be done directly. Dellaportas and Smith (1993) described the use of Gibbs sampling with an adaptive rejection algorithm to simulate the parameters for a logistic model. The probit model which applies the data augmentation approach requires simulation mainly from standard distributions such as the multivariate normal, therefore, is easy to implement in many statistical computer packages. To utilize this direct sampling advantage to the logit model, the auxiliary variable method described in Section 3.3 provides a possibility.

The auxiliary variable method for $\pi(\boldsymbol{x})$ is restated as follows. Suppose that $\pi(\boldsymbol{x})$ can be written in the form $\pi(\boldsymbol{x}) \propto \pi_0(\boldsymbol{x})b(\boldsymbol{x})$. We introduce one auxiliary variable u and define $\pi(u|\boldsymbol{x})$ to be the uniformly distribution on the interval $[0, b]$. Then we can apply the Gibbs sampler to $\pi(\boldsymbol{x},u)$. That is, we first draw u from $\pi(u|\boldsymbol{x})$.

then draw $\boldsymbol{x}$ from $\pi_0(\boldsymbol{x})$, and impose the condition $\{b(\boldsymbol{x}) \geq u\}$ by rejection. Now, if $\pi(\boldsymbol{x})$ uses the logistic link, then we can let $\pi_0(\boldsymbol{x})$ use the probit link or the family of t distributions and b will be the ratio of density functions for these two link function. Apart from b values near 0, which correspond to the tails, the distribution of logit or probit models is generally quite similar, the rejection rate of the condition $\{b(\boldsymbol{x}) \geq u\}$ should be acceptable.

4.3.4 Generalizations to a Multinomial Response

The Gibbs sampling approach can also be applied to the multinomial probit model (Aitchison and Bennett, 1970; Hausman and Wise, 1978). Suppose the response variables y_t, $t = 1, \ldots, T$, have K possible values, which for simplicity are labeled $1, \ldots, K$. The probabilities are simply connected by $p(y_t = k) = \pi_{tk}$, $k = 1, \ldots, K$. Then we introduce T latent variables $\boldsymbol{w_1}, \ldots, \boldsymbol{w_T}$, where $\boldsymbol{w_t} = (w_{t1}, \ldots, w_{tK})'$, and define

$$w_{tk} = \boldsymbol{z}_{tk}'\boldsymbol{\beta_t} + \varepsilon_{tk}, \quad t = 1, \ldots, T, \quad k = 1, \ldots, K,$$

where $\boldsymbol{\varepsilon_t} = (\varepsilon_{t1}, \ldots, \varepsilon_{tK})'$ is distributed $N_K(\mathbf{0}, \boldsymbol{\Sigma})$. Denote $\boldsymbol{y_T} = (y_1, \ldots, y_T)'$, where $y_t \in \{1, \ldots, K\}$. Let $\boldsymbol{z_t} = (\boldsymbol{z_{t1}}, \ldots, \boldsymbol{z_{tK}})'$, the preceding model can be written as

$$\begin{bmatrix} \boldsymbol{w_1} \\ \vdots \\ \boldsymbol{w_T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z_1\beta_1} \\ \vdots \\ \boldsymbol{z_T\beta_T} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \vdots \\ \boldsymbol{\varepsilon_T} \end{bmatrix} \tag{4.11}$$

or as

$$\boldsymbol{w_T^*} = \boldsymbol{z_T^*\beta_T^*} + \boldsymbol{\varepsilon_T^*}, \quad \boldsymbol{\varepsilon_T^*} \propto N_{TK}(\mathbf{0}, \boldsymbol{\Omega} = \boldsymbol{I_T} \otimes \boldsymbol{\Sigma}),$$

where $\boldsymbol{w_T^*} = [\boldsymbol{w_1}, \ldots, \boldsymbol{w_T}]'$, $\boldsymbol{z_T^*} = [\boldsymbol{z_1}, \ldots, \boldsymbol{z_T}]'$, and $\boldsymbol{\varepsilon_T^*} = (\boldsymbol{\varepsilon_1'}, \ldots, \boldsymbol{\varepsilon_T'})'$. $N_{TK}(\mathbf{0}, \boldsymbol{\Omega} = \boldsymbol{I_T} \otimes \boldsymbol{\Sigma})$. McFadden (1973) has shown that multinomial logit models can be derived

in this setup if and only if the errors ε_T^* are a random sample from a Type I extreme value distribution (Johnson and Kotz, 1970).

Consider the linear Gaussian evolution model, $\boldsymbol{\beta}_t = \boldsymbol{F}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t$, $t = 1, \ldots, T$, where $\boldsymbol{\xi}_t$ is Gaussian noise, $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t)$ and $\boldsymbol{\beta}_0 \sim N(\boldsymbol{a}_0, \boldsymbol{Q}_0)$. For simplicity, the $\boldsymbol{z}_t'$, $\boldsymbol{F}_t$, $\boldsymbol{Q}_t$, $\boldsymbol{a}_0$, and $\boldsymbol{Q}_0$ are assumed to be deterministic and known. Also the Σ is parameterized in terms of a vector $\boldsymbol{\theta}$ and $f(\boldsymbol{\theta})$ is a prior on $\boldsymbol{\theta}$. Similar to Section 4.3.2, the full conditional distributions of $\boldsymbol{w}_T^*$, $\boldsymbol{\beta}_T^*$, and $\boldsymbol{\theta}$ are given below:

1. Given $\boldsymbol{\beta}_T^*$, $\boldsymbol{y}_T$, $\boldsymbol{\theta}$, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ are an independent collection with

$$\boldsymbol{w}_t | \boldsymbol{\beta}_T^*, \boldsymbol{y}_T, \boldsymbol{\theta} \propto N(\boldsymbol{z}_t\boldsymbol{\beta}_t, \Sigma), \quad t = 1, \ldots, T. \tag{4.12}$$

   such that the y_tth component of $\boldsymbol{w}_t$ is the maximum. This can be simulated by drawing a sample $\boldsymbol{w}_t$ from $N(\boldsymbol{z}_t\boldsymbol{\beta}_t, \Sigma)$ and accepting the draw if the condition is satisfied. Another method of performing this drawing is in McCulloch and Rossi (1991).

2. The full conditionally distributions of $\boldsymbol{\beta}_t$ are given by

$$p(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{s \neq t}, \boldsymbol{w}_t^*, \boldsymbol{\theta}) \propto N(\boldsymbol{B}_t\boldsymbol{b}_t, \boldsymbol{B}_t), \quad t = 0, \ldots, T. \tag{4.13}$$

   with

$$\boldsymbol{B}_t^{-1} = \begin{cases} \boldsymbol{Q}_t^{-1} + \boldsymbol{F}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1} + \boldsymbol{z}_t'\Sigma^{-1}\boldsymbol{z}_t, & \text{if } t = 0, \ldots, T-1 \\ \boldsymbol{Q}_t^{-1} + \boldsymbol{z}_t'\Sigma^{-1}\boldsymbol{z}_t, & \text{if } t = T \end{cases}$$

$$\boldsymbol{b}_t' = \begin{cases} \boldsymbol{a}_0'\boldsymbol{Q}_0^{-1} + \boldsymbol{\beta}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1}, & \text{if } t = 0 \\ \boldsymbol{\beta}_{t-1}'\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1} + \boldsymbol{\beta}_{t+1}'\boldsymbol{Q}_{t+1}^{-1}\boldsymbol{F}_{t+1} + \boldsymbol{w}_t'\Sigma^{-1}\boldsymbol{z}_t, & \text{if } t = 1, \ldots, T-1 \\ \boldsymbol{\beta}_{t-1}'\boldsymbol{F}_t'\boldsymbol{Q}_t^{-1} + \boldsymbol{w}_t'\Sigma^{-1}\boldsymbol{z}_t, & \text{if } t = T. \end{cases}$$

3. $\boldsymbol{\theta}|\boldsymbol{w}_T^*, \boldsymbol{\beta}_T^*$ is distributed according to the density proportional to

$$f(\boldsymbol{\theta})|\boldsymbol{\Omega}(\boldsymbol{\theta})|^{-1/2}\exp\{-\frac{1}{2}(\boldsymbol{w}_T^* - \boldsymbol{z}_T^*\boldsymbol{\beta}_T^*)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta})(\boldsymbol{w}_T^* - \boldsymbol{z}_T^*\boldsymbol{\beta}_T^*)\}. \tag{4.14}$$

To implement the Gibbs sampler, we starts with initial guesses of $\boldsymbol{a_0}$, $\boldsymbol{Q_0}$, and $\boldsymbol{Q_t}$, and cycle through the conditional distributions (4.12), (4.13), and (4.14) in that order.

Chapter 5

Performance of the MCMC Samplers

In this chapter, we apply the MCMC algorithms introduced in the preceding chapter and empirically compare them with posterior mode estimates of the EKFS on a binary rainfall time series example. Two link functions that are approximately equivalent to the logistic link function are also considered.

5.1 Application to Binary Data

5.1.1 Binary Rainfall Data

Ishiguro and Sakamoto (1983) present a data set consisting of the number of occurrences of rainfall over 1mm in Tokyo for each day during the years 1983 - 1984. The problem is to estimate the probability π_t of rainfall on a specific calendar day $t = 1, \ldots, 366$, which is believed to be gradually changing with time. Note that February 29 had only one observation. Kitagawa (1987) used the following simple dynamic binary logit model:

$$\begin{aligned} y_t &\sim \begin{cases} \text{Binomial}(1, \pi_t), & t = 60 \quad \text{(February 29)} \\ \text{Binomial}(2, \pi_t), & t \neq 60, \end{cases} \\ \pi_t &= h(\beta_t) = \exp(\beta_t)/(1 + \exp(\beta_t)), \\ \beta_t &= \beta_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma^2), \quad \beta_0 \sim N(a_0, \sigma_0^2). \end{aligned} \tag{5.1}$$

This model contains only a scalar grand mean parameter that follows a random walk model. The unknown hyperparameters a_0, σ_0^2, and σ^2 were estimated by the EM-type algorithm as $\hat{a}_0$ = -1.51, $\hat{\sigma}_0^2$ = 0.0019, and $\hat{\sigma}^2$ = 0.032 (Fahrmeir and Tutz. 1994). The results obtained by Kitagawa's (1987) numerical integration approach and the posterior mode estimates of the EKFS by Fahrmeir (1992) of $\hat{\pi}_t = h(\hat{\beta}_{t|366})$ together with confidence intervals $(\hat{\pi}_t \pm \hat{\sigma})$ are almost identical.

Using the same hyperparameter values, Fahrmeir, Hennevogl. and Klemme (1992) applied the Gibbs sampler to the rainfall data. They found that the estimation of the posterior mean based on 50 Gibbs runs each having 20 iterations and based on posterior mode estimates of the EKFS are more or less identical. Although both estimators are in close agreement, the Gibbs sampler results may be inaccurate as the runs are too short as indicated below.

5.1.2 Diagnostics by a Long Run Gibbs Sampler

To see that the Gibbs runs used by Fahrmeir, Hennevogl, and Klemme (1992) are too short to get accurate results, a Gibbs sampler using the same hyperparameter values based on 50,000 Gibbs runs is applied to the rainfall data. In addition, to insure that the series obtained by the Gibbs samplers do not exhibit any unusual behavior, four parallel chains were run with different starting points $\boldsymbol{\beta}^{(0)} = (\beta_0^{(0)}, \ldots, \beta_{366}^{(0)})$, chosen from above and below the posterior mode smoothing estimates by Fahrmeir et al. (1992). Ideally, the starting points should be overdispersed but not wildly inaccurate. The $N(-4, 1)$, $N(0, 1)$, and $N(4, 1)$ were chosen to ensure that our starting points for the iterative simulation do not entirely miss important regions of the target distribution. The data generating scheme for the starting points for four parallel chains is in the Table 5.1.

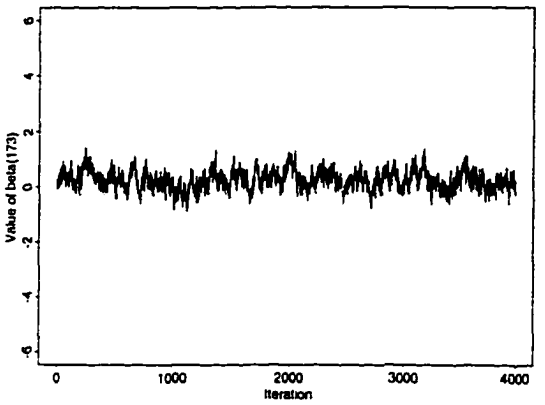
Plots of sample paths and autocorrelations of the parameters are two useful

Starting point set	Data generating scheme
$\boldsymbol{\beta}_1^{(0)}$	posterior mode smoothing estimates
$\boldsymbol{\beta}_2^{(0)}$	i.i.d. $N(4,1)$
$\boldsymbol{\beta}_3^{(0)}$	i.i.d. $N(0,1)$
$\boldsymbol{\beta}_4^{(0)}$	i.i.d. $N(-4,1)$

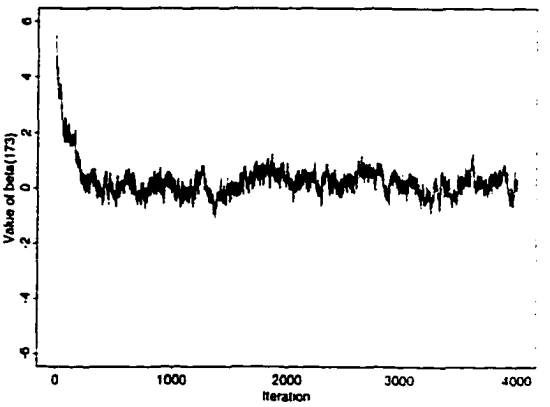
Table 5.1. The data generating scheme for the starting points

tools for monitoring the performance of the samplers. Figure 5.1 (a) - (d) shows the trajectories of the parameter β_{173} (this parameter is the mode of the posterior mean sequence from the posterior mode estimates) based on the Gibbs sampler run of length 50,000 corresponding to the starting values $\boldsymbol{\beta}_i^{(0)}$, i = 1, 2, 3, 4. To observe the initial transient behavior of different starting values, we only plot the first 4,000 steps of the Gibbs samples. The results of these trajectory all seem to have converged to stationarity. The autocorrelation curves for the parameter β_{173} are shown in Figures 5.2 (a) - (d). The autocorrelations show strong serial correlation and are significantly nonzero out to about lag 200 for all four different starting values.

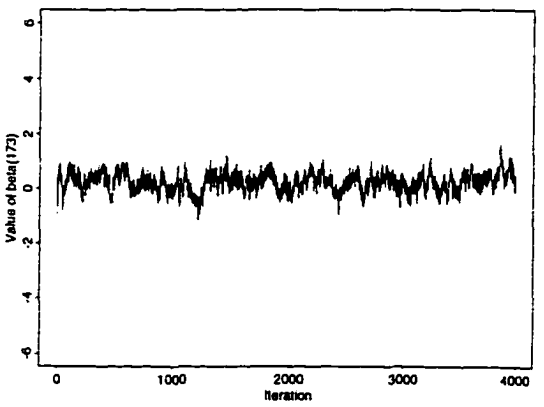
Although all of these four independent series we examined seem to have converged to the same distribution, one may easily get misleading answers when drawing inferences from a short simulated sequence without further diagnostics. Gelman and Rubin (1992) present a simple example to suggest that it is generally impossible to assess convergence of a Gibbs sampler from a single simulated series. The strong correlations also imply that errors are bigger than for *i.i.d.* observations, thus we need larger samples than would be required with *i.i.d.* sampling. In practice, for the Gibbs sampler a run of length 20 would be too short to obtain accurate estimates of the variances of point estimators, including means, and variances. In the next section, several implementation issues will be considered when applying MCMC algorithms.



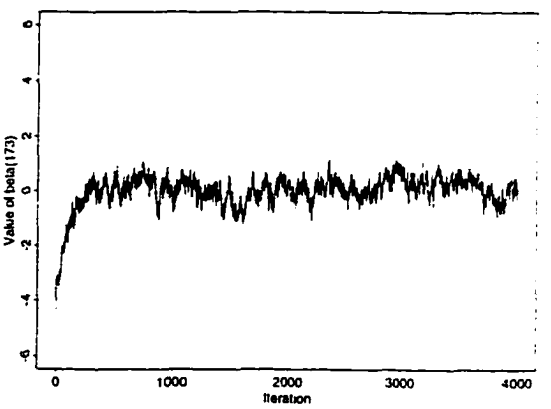
(a) Started at $\beta_1^{(0)}$



(b) Started at $\beta_2^{(0)}$

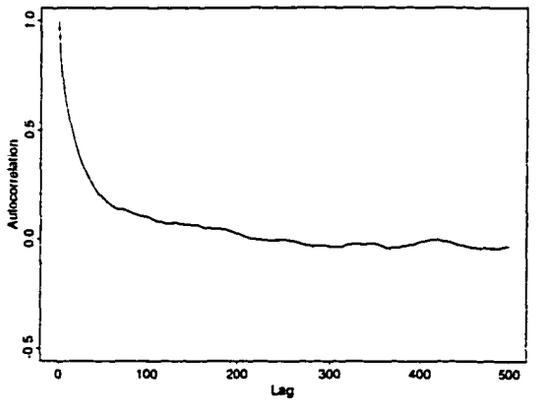


(c) Started at $\beta_3^{(0)}$

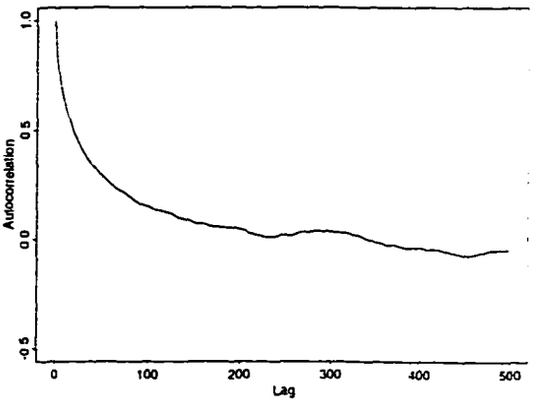


(d) Started at $\beta_4^{(0)}$

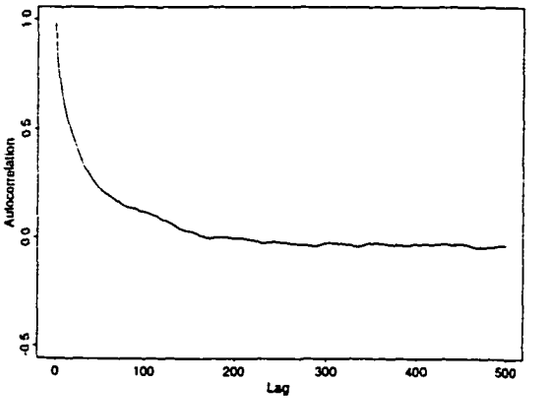
Figure 5.1. Trajectory of the parameter β_{173}.



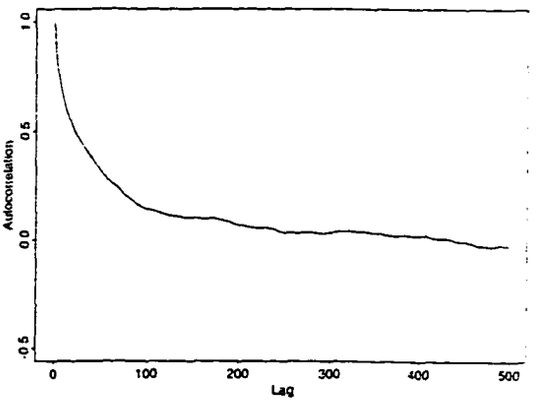
(a) Started at $\beta_1^{(0)}$



(b) Started at $\beta_2^{(0)}$



(c) Started at $\beta_3^{(0)}$



(d) Started at $\beta_4^{(0)}$

Figure 5.2. Empirical autocorrelation curve of the parameter β_{173}.

5.1.3 Implementation Issues

MCMC algorithms can be very useful for iterative simulation techniques. but naive use can give misleading answers. Because a finite number of iterations are used to estimate the target distribution and the simulated random variables are. in general. never from the target distribution, considerable care is required in choosing, implementing and drawing inference from a finite iteration simulation. In this section. several implementation issues will be considered based on the analyzed results in Kitagawa (1987), Fahrmeir (1992), Fahrmeir, Hennevogl, and Klemme (1992), and preliminary diagnostics in the preceding section.

The most basic issue is that valid inference in Markov chain Monte Carlo results from averaging over one long run of the chain or multiple short runs is desirable. It is undeniable that multiple runs have some diagnostic value: if the results of multiple runs completely disagree, then the runs are too short and cannot be used for inference. One long run is also a valuable diagnostic: if the run doesn't seem stationary it is too short, and the longer the run, the better the chance of detection. Based on the preliminary diagnostics in the preceding section, we found: (1) for all different starting values, the results seem to agree with each other. (2) the results are also similar to the results in Kitagawa (1987), Fahrmeir (1992), and Fahrmeir. Hennevogl, and Klemme (1992), (3) the results based on runs of 50.000 in Figure 5.1 show a relatively stable estimated posterior mean. We suggest using a single long run of length 50,000 instead of using multiple separate runs.

As noted in the preceding section, when we use a single run with a Markov chain to sample from a posterior distribution, a complication that arises from the autocorrelation is that variances of estimates are harder to obtain. Three commonly used approaches to computing standard errors that attempt to adjust for correlations are window estimators (e.g., Hastings, 1970; Geyer, 1991; Geweke. 1992: Green and

Han, 1992), the batch means method (e.g., Ripley, 1987), and time series methods (e.g., Priestley, 1981). We will use the batch means method to compute the standard errors since it is the simplest and is easy to program. To calculate standard errors using batch means, the data, $x_1, x_2, \ldots$, are divided into b batches, each of length l. The mean $\hat{\mu}_i$ of each batch i is calculated as

$$\hat{\mu}_i = \frac{1}{l} \sum_{j=(i-1)l+1}^{il} x_j, \quad i = 1, \ldots, b, \tag{5.2}$$

and the standard error of the mean is estimated as

$$\bar{\mu} = \frac{1}{b} \sum_{i=1}^{b} \hat{\mu}_i, \quad \hat{\sigma}_\mu^2 = \frac{1}{b(b-1)} \sum_{i=1}^{b} (\hat{\mu}_i - \bar{\mu})^2. \tag{5.3}$$

For the unknown hyperparameters, we treat the hyperparameters as stochastic parameters with a prior distribution. The priors for the hyperparameters a_0, and σ^2 are specified by $[a_0] \sim N(\mu_{a_0}, \sigma_{a_0}^2)$, and $[\sigma^2] \sim \mathrm{IG}(a_1, b_1)$, where IG denotes the inverse gamma distribution and μ_{a0}, σ_{a0}^2, a_1, and b_1 are assumed known. The value of the hyperparameter σ_0^2, should be similar to σ^2. Let $\sigma_0^2 = c\sigma^2$, where $c > 1$ to reflect variation in the initial guess. Thus one has to add further drawings from the posterior

$[\sigma^2 | \boldsymbol{\beta_T}, \sigma_0^2, a_0, \boldsymbol{y_T}] \sim IG(a_1 + \frac{1}{2}T, b_1 + \frac{1}{2}\sum_{t=1}^{T}(\beta_t - \beta_{t-1})^2)$,

$[\sigma_0^2 | \boldsymbol{\beta_T}, \sigma^2, a_0, \boldsymbol{y_T}] = 2\sigma^2$,

and

$[a_0 | \boldsymbol{\beta_T}, \sigma_0^2, \sigma^2, \boldsymbol{y_T}] \sim N(\frac{\sigma_{a_0}^2 \beta_0 + \mu_{a_0}\sigma_0^2}{\sigma_{a_0}^2 + \sigma_0^2}, \frac{\sigma_{a_0}^2 \sigma_0^2}{\sigma_{a_0}^2 + \sigma_0^2})$,

to the drawings, where $\boldsymbol{\beta_T} = (\beta_0, \ldots, \beta_T)$, and $T = 366$. For the binary rainfall data given above, the hyperparameter prior specification was defined by $\mu_{a_0} = -1.58$, $\sigma_{a_0}^2 = 0.025$, $a_1 = \frac{1}{2}$, $b_1 = 0.016$, and $c = 2$ reflecting rather vague initial information relative to the EM estimates provided by Fahrmeir and Tutz (1994).

All MCMC runs were coded in the XLISP-STAT (version 3) language and run

on a HP 715/100 workstation. We compared run times of the various algorithms. However, run times would have been shorter if one codes in C or Fortran.

5.2 Performance of the Gibbs Sampler

In this section, we will perform a through empirical investigation of the block Gibbs sampler. Using rejection sampling to obtain a random sample from the conditional density of the block Gibbs sampler of block size n, the iteration time will increase quadratically in n. To find the optimal block size, some compromise obviously must be established. A heuristic is to choose the block size as large as possible while not increasing the iteration time significantly. The modified ARS algorithm described in Section 4.1 seems to provide a strategy that decreases the number of iterations significantly. Two investigations are performed: a comparison of block and plain Gibbs, and a comparison of rejection sampling and the modified ARS algorithm within the Gibbs sampler. Note that the constructed linear upper hull is independent in each run of using modified ARS algorithm.

Table 5.2 shows estimates of the posterior moments of the parameter β_{173} for different combinations of block size and rejection method based on Gibbs sampler runs of length 50,000. Because these posterior moments are based on a single run of the Gibbs sampler, three values of the table give summary statistics for this run that are helpful for diagnostic purposes. The lag one correlation gives lag one autocorrelations of the sample, the SE under the posterior mean gives numerical standard errors for the posterior means based on the batch means method described in Section 5.1.3. and the SE under the posterior SD gives standard errors of the estimated posterior standard deviations. Batches of increasing size were collected until the lag one correlation of the batch means was under 0.05. For comparison, the table also shows the posterior mean and standard deviation of the parameter β_{173} computed by applying the extended

Kalman filtering and smoothing method described in Fahrmeir (1992).

Algorithm	Posterior Mean (SE)	Posterior SD (SE)	Lag One Correlation	R	Run Time
EKFS	0.161	0.334			0.5 sec.
Rejection within Gibbs	0.685 (0.0093)	0.547 (0.002)	0.74	0.64	208 min.
Modified ARS within Gibbs	0.706 (0.0095)	0.551 (0.002)	0.75	0.01	132 min.
Rejection within Block Gibbs Size = 2	0.688 (0.0079)	0.549 (0.002)	0.67	0.90	828 min.
Modified ARS within Block Gibbs Size = 2	0.707 (0.0081)	0.555 (0.002)	0.70	0.04	327 min.

Table 5.2. Estimated posterior means and standard deviations of the parameter β_{173} using rejection sampling and the modified ARS algorithms within the plain and block Gibbs sampler. R is the proportion of candidates rejected.

Based on a preliminary sample of 5,000 observations from the block Gibbs sampler with rejection sampling the rejection rate for block size 1, 2, 3, and 4 is about 63%, 93%, 98%, and 99%. Since the run time increases dramatically with the block size, we only presented results from using block size 2 of the block Gibbs sampler. The lag one autocorrelation of the block Gibbs sampler, 0.67, is still very high. The estimated standard errors of posterior means obtained using the block Gibbs sampler are slightly better than for the plain Gibbs sampler. One can get a better improvement by using a larger block size, but the run times could increase enormously. Overall, the result of using block size 2 of the block Gibbs sampler makes only a small improvement compared to the plain Gibbs sampler. This blocking

strategy is not very useful for this binary time series example.

The estimated results obtained by the modified ARS algorithm within the plain and block Gibbs sampler are similar to the results from rejection sampling within the plain and block Gibbs sampler. But comparing the rejection rate 0.01 using the modified ARS algorithm with 0.64 using ordinary rejection sampling on the plain Gibbs sampler and 0.04 using the modified ARS algorithm with 0.9 using ordinary rejection sampling on the block Gibbs sampler, we found that the rejection rate reduction obtained using the modified ARS algorithm within the Gibbs sampler is quite significant. However, the requirement of the initialization, sampling and updating step in the modified ARS algorithm complicated coding and increased execution time and makes the run times of the modified ARS algorithm within the plain and block Gibbs sampler about twice as fast as the rejection sampling within the plain and block Gibbs sampler. Thus using the modified ARS algorithm is worth while in this example.

The estimated posterior mean is about the same for all four algorithms. The value of the estimated standard deviation obtained from all four algorithms is higher than from EKFS. It suggested that the estimated value of this posterior moment could be under-estimated by the EKFS algorithm. Figure 5.3 shows the corresponding estimates $\hat{\beta}_t$, together with $\pm$ one SD, resulting from the plain Gibbs sampler with rejection sampling. For easier reference, the estimates of $\hat{\beta}_t$ obtained by the posterior mode estimates of the EKFS are also imposed in Figure 5.3. In these four algorithms, we treat the hyperparameter σ^2 as a stochastic parameter with a prior distribution to reflect the uncertainty about this parameter. Imposing a fixed value of the hyperparameter σ^2 as is done by the EKFS algorithm seems to make the estimated standard deviations of the β_t parameters too tight and make their posterior means too smooth. This is shown in Figure 5.3 as well as Table 5.2: the local maximum (minimum) values of the posterior means from the EKFS algorithm tend to

have lower (higher) values than from the Gibbs sampler because of smoothing. The estimated posterior standard deviations from the EKFS algorithm are. at least in this binary time series example, also lower than the ones produced by the MCMC methods. The plots of the estimates of parameters with confidence bands for the other three algorithms are similar to Figure 5.3 and are therefore omitted.

5.3 Performance of M-H Algorithms

We next present two examples of the use of the M-H algorithm. The first algorithm was a random walk chain with increment random variables z_0' and z_t', $t = 1, \ldots, T$, distributed as $z_0 \sim N[0, \sigma^2/(1-\rho^2)]$, and $z_t \sim N[\rho z_{t-1}, \sigma^2]$, where ρ and σ^2 can be approximated using $\mathrm{Cov}(\beta_t, \beta_{t-1})$ and $\mathrm{Var}(\beta_t)$. Based on the samples from the plain Gibbs sampler in the preceding section, we let $\rho = 0.88$ and $\sigma^2 = 0.36$ in the random walk chain. The second algorithm was an independence chain with candidate generated by a multivariate normal distribution, $N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are mean and covariance matrix estimated by 5,000 samples generated by the plain Gibbs sampler form the preceding section.

When trying to update all components at a time in the random walk chain and independence chain, the acceptance rate becomes very low. One possible strategy is to apply the M-H algorithm to sub-blocks of the $\boldsymbol{\beta_T}$, rather than simultaneously to all components of the vector as described in Section 4.2.3. The rejection rate and estimated standard error for block size 2 to 10 based on runs of 50,000 observations generating from the candidate distributions mentioned above for random walk chain and independence chain are shown in Figure 5.4 and 5.5. As expected, the rejection rate of the random walk chain increases enormously as block size increases. The rejection rate of the independence chain does not increases as much as for the random walk chain. The estimated standard errors of the random walk and independence

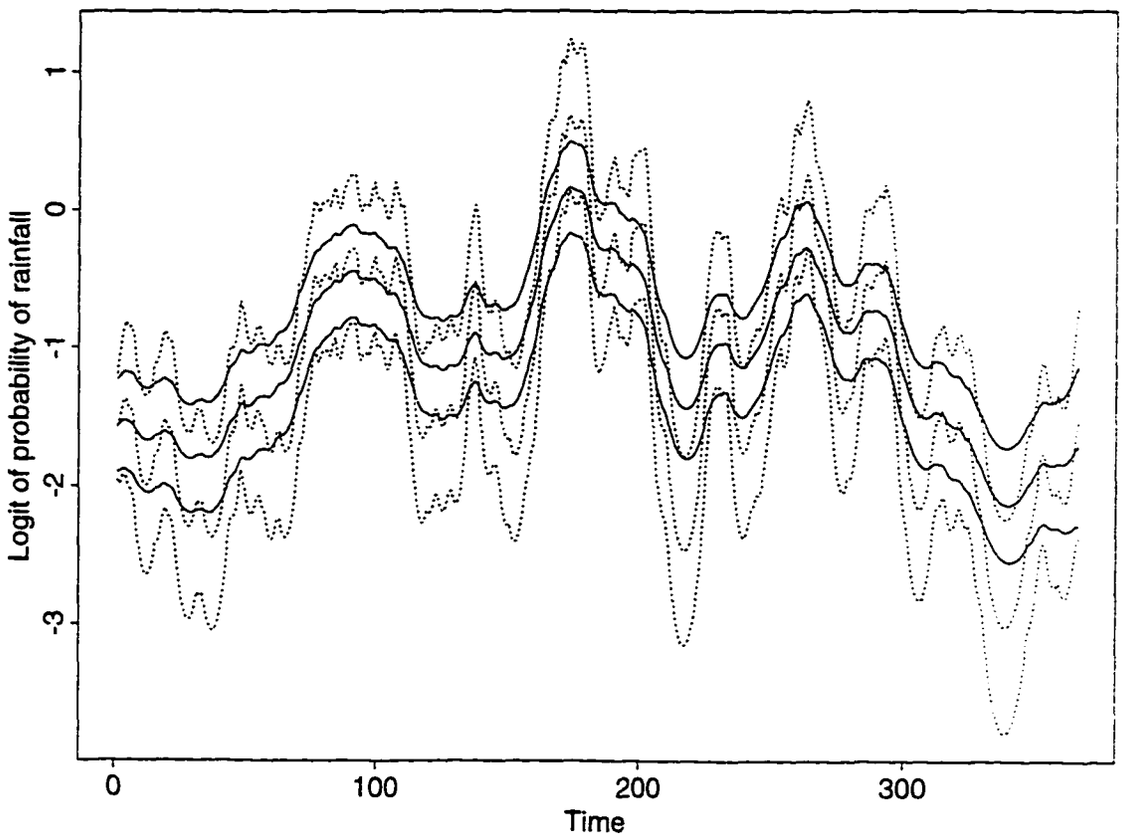
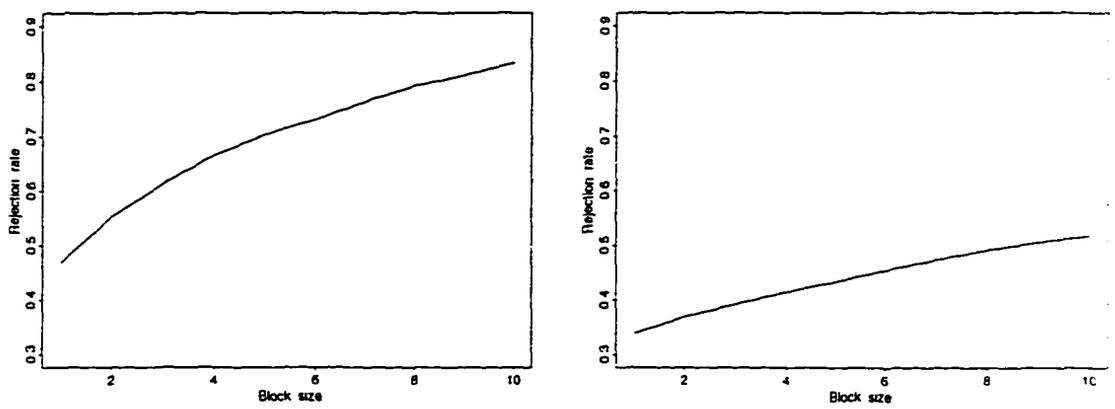
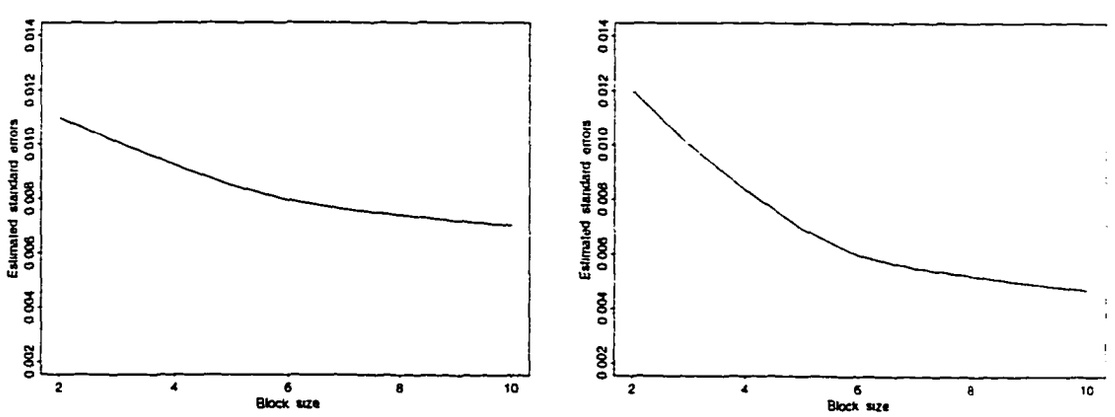


Figure 5.3. Estimated posterior means and standard deviations of logit of probability of rainfall by using rejection sampling within the Gibbs sampler (· · ·) and by EKFS (—).



(a) Random Walk Chain (b) Independence Chain

Figure 5.4. Rejection rates of random walk and independence chains.



(a) Random Walk Chain (b) Independence Chain

Figure 5.5. Estimated standard errors of random walk and independence chains.

chains decrease as block size increases. The block size of 6 is chosen for the random walk chain based on Roberts, Gelman, and Gilks' (1994) suggestion that the optimal acceptance rate is around 0.25. The block size of 6 is also chosen for the independence chain although its acceptance rate is not around 0.25.

The posterior standard moments of the parameter β_{173} based on runs of 50.000 of random walk and independence chains are shown in Table 5.3. Again, the posterior means and standard deviations computed by the extended Kalman filtering and smoothing method are included for comparison. The high serial correlation with the random walk chain is not unexpected and stems from the long-memory in the candidate draws. It is possible that better choices of ρ and σ^2 in the candidate generating distribution of random walk chains will reduce the serial correlation. The independence chain performs somewhat better than the random walk chain, since the candidate generating density choosing in the independence chain is not too different in shape from the posterior distribution. Other useful strategies for independence chains to get better improvements are to increase the block size or to combine with other Markov chain algorithms. For example, the candidate generating distribution can be chosen from a mixture of a multivariate normal distribution $N(\tilde{\mu}.\tilde{\Sigma})$ and a standard normal distribution. The mixing probabilities can be chosen to produce a rejection rate in the candidate generation phase of approximately 0.75.

The estimated posterior mean is about the same for the random walk and independence chains. The values of the estimated standard deviations obtained from these two MCMC algorithms are about the same and similar to the Gibbs sampler values but higher than the EKFS values. It again suggests that the value of the posterior standard deviation could be under-estimated by the EKFS algorithm. The plot of the estimates of parameters with confidence bands is similar to Figure 5.3 and is therefore omitted.

Algorithm	Posterior Mean (SE)	Posterior SD (SE)	Lag One Correlation	R	Run Time
EKFS	0.161	0.334			0.5 sec.
Random Walk Size = 6	0.708 (0.0080)	0.558 (0.002)	0.97	0.73	155 min.
Independence Size = 6	0.710 (0.0074)	0.556 (0.002)	0.56	0.45	235 min.

Table 5.3. Estimated posterior means and standard deviations of the parameter β_{173} using M-H algorithms. R is the proportion of candidates rejected.

5.4 Choice of Link Function

Since the most popular link function for binary data is the logistic link, we first explore the connection between the logistic and t links and the mixture of normal distributions $2ZK$. The logistic distribution function has the simple form $F(x) = (1 + e^{-x})^{-1}$ and the density $f(x) = F(x)\{1 - F(x)\}$. It is well known that the difference of $|F(\beta x) - \Phi(x)|$, where $\Phi(\cdot)$ denotes the standard normal distribution function, is less than 0.023 when $\beta^{-1} = \pi/\sqrt{3}$, the standard deviation of the logistic law (Johnson and Kotz, 1970), and is minimized to 0.009 when $\beta = 0.5875 \simeq (16\sqrt{3})/(15\pi)$ (Birnbaum and Dudman, 1963). The kurtosis of the logistic distribution is 1.2, so that its tails are larger than those of the normal distribution with the same standard deviation. This suggests a possibly closer similarity between the logistic distribution function and the distribution function of a Student t distribution.

The parameter of the t distribution may be obtained in various ways: (1) by minimizing the maximum difference between the distribution functions (Mudholkar and George, 1978), (2) by plotting quantiles of the logistic distribution against quantiles of a t distribution, and (3) by minimizing the absolute value of the difference

between the density functions. In Figure 5.6, we display the differences between the distribution function, $L(x)$ of the logistic random variable with unit standard deviation and the distribution function $t(v,x)$ of a Student t variable with $v = 5, 6, 7, 8, 9, 10$, and 11 degrees of freedom scaled so as to have unit standard deviation, i.e. $L(x) - t(v,x)$. The Student t distributions with 7 and 8 degrees of freedom seem to have smaller difference than others degrees of freedom. Figure 5.7 plots quantiles of the logistic distribution against quantiles of t distributions with v degrees of freedom, $v = 5, 6, 7, 8, 9, 10$, and 11, and against $2ZK$ by using 20 independent exponential variables W_j in the relation $2K^2 = \sum_{j=1}^{\infty} W_j/j^2$. For probabilities between 0.0001 and 0.9999, logistic quantiles are approximately a linear function of $t(7)$ and $t(8)$ quantiles. For using 20 independent exponential variables in generating the asymptotic Kolmogorov distribution, the plot doesn't show an approximately a linear function. In Figure 5.8, we plot the total variation distance between the density functions of the logistic and the Student t with v degrees of freedom, $v = 5$ - 11. The Student t distribution around integer 7 degrees of freedom has the smallest value. Overall, the logistic function appears approximately equivalent to a Student t function with 7 degrees of freedom. This statement is consistent with Mudholkar and George's (1978) result that the logistic distribution has the same kurtosis as a t distribution with nine degrees of freedom and with Albert and Chib's (1993) result that the logistic quantiles are approximately a linear function of $t(8)$ quantiles.

When trying to update all components from the multivariate normal distribution at a time for both link functions we used Cholesky decompositions, but this resulted in large computational time and was therefore not practical. To overcome this problem, Carter and Kohn (1994) and de Jong and Shephard (1995) have suggested an approach which uses the state space structure to draw efficiently from the multivariate posterior distribution of the disturbances of the model. The strategy we

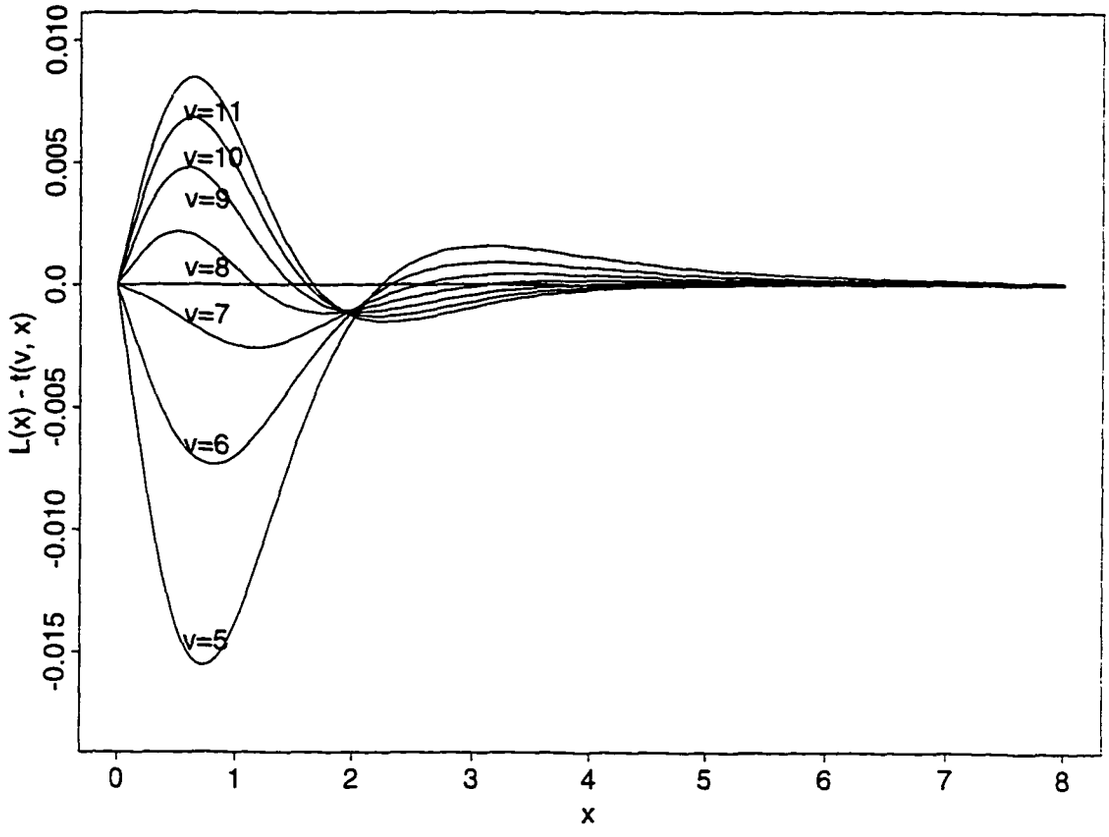


Figure 5.6. Differences between the logistic $L(x)$ and the Student $t(v, x)$ distribution functions with $v = 5$ - 11 degrees of freedom. The logistic and the Student t distribution functions all scaled to unit variance. Seven values of degrees of freedom: 5 (bottom) to 11 (top).

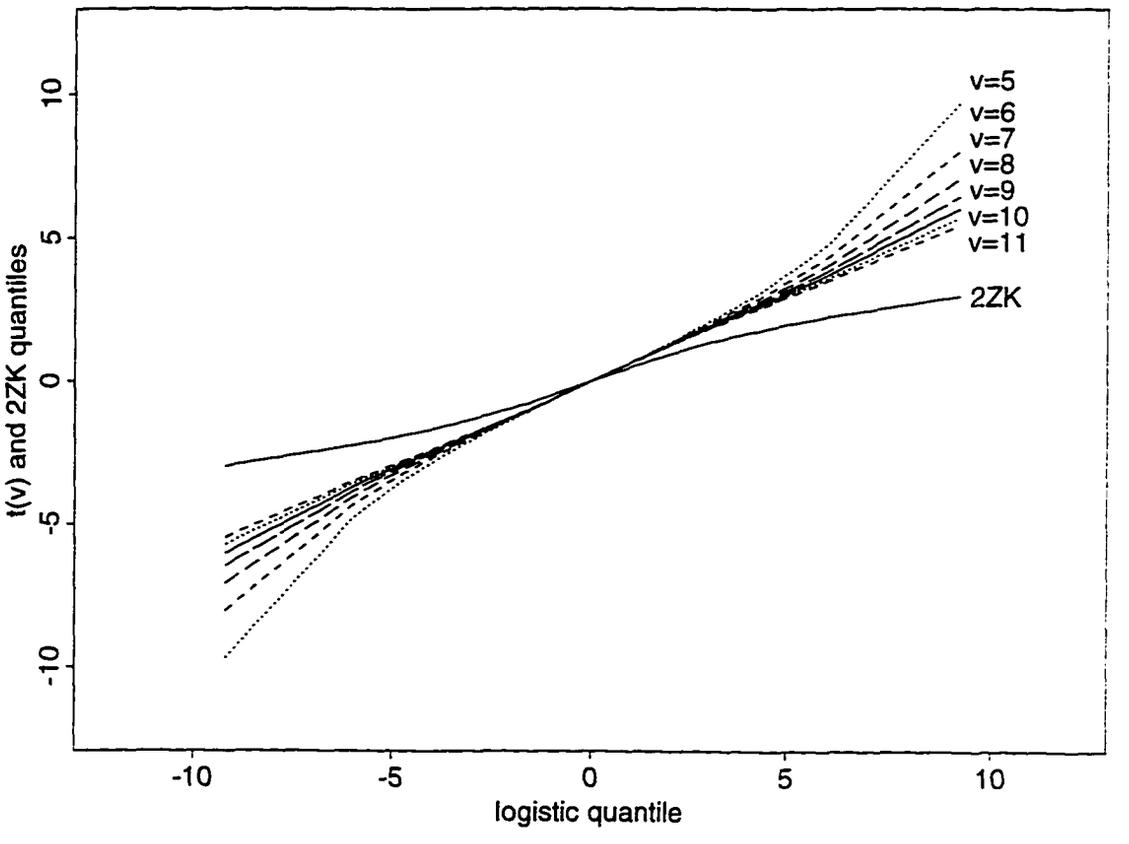


Figure 5.7. Plot of logistic quantile against $t(v)$ quantile with $v = 5$ - 11 degrees of freedom and $2ZK$ quantile for probabilities between 0.0001 and 0.9999.

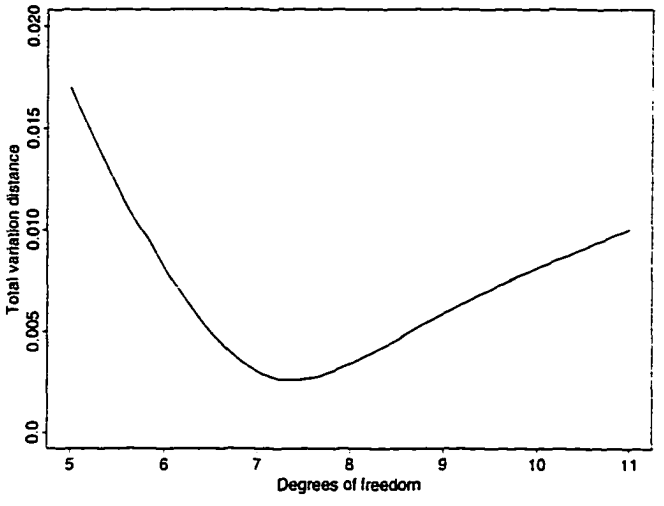


Figure 5.8. Total variation distance between the density functions of the logistic and the Student t with 5 - 11 degrees of freedom.

Algorithm	Posterior Mean (SE)	Posterior SD (SE)	Lag One Correlation	Run Time
EKFS	0.161	0.334		0.5 sec.
t (7) Size = 6	0.690 (0.0057)	0.550 (0.002)	0.22	301 min.
Kolmogorov Size = 6	0.693 (0.0061)	0.554 (0.002)	0.27	725 min.

Table 5.4. Estimated posterior means and standard deviations of the parameter β_{173} using different link functions.

used in this binary rainfall data is to apply Cholesky decomposition method to sub-blocks of components, rather than simultaneously to all components at a time. As in the M-H algorithms, blocks of size 6 are chosen for both link functions, $t(7)$ and $2ZK$. The posterior mean and standard deviation of the parameter β_{173} based on runs of 50,000 are shown in Table 5.4. The posterior means and standard deviations computed by the extended Kalman filtering and smoothing method are included for comparison. The results in Table 5.4 indicate that the blocking strategy is sufficient for this binary rainfall example. These two link functions do not show any high serial correlation. The estimated posterior moments of $t(7)$ and mixtures of normal distributions $2ZK$ are about the same and similar to the results of using the Gibbs sampler. This is not surprising, since these two link functions are approximately equivalent to the logistic distribution. Because the implementation of mixtures of normal distributions $2ZK$ requires one to generate 20 independent exponential variables for the asymptotic Kolmogorov distribution, the run times of mixtures of normal distributions $2ZK$ is slower than from the $t(7)$ link function. Similar to the other algorithm, the value of the estimated standard deviation obtained from these two link functions are higher than from EKFS. The plot of the estimates of parameters with confidence bands is similar to Figure 5.3 and is therefore omitted.

5.5 Comparison of the MCMC Samplers

For comparison of all methods, the two results estimated standard error and run time based on chains of length 50,000 from MCMC samplers we tested are shown in Figure 5.9. The lower the value of the estimated standard error and run time, the better the performance of the algorithm. Clearly, the two link functions have the lowest estimated standard error, the two Gibbs samplers have the largest estimated standard error, and the Metropolis algorithms and the two block Gibbs samplers lie in

between. The Metropolis algorithms and the two block Gibbs samplers have similar estimated standard errors, but the Metropolis algorithms apparently have shorter run times. For further comparison, the asymptotic relative efficiencies (ARE) adjusted for run time are shown in Table 5.5. For a MCMC sampler, M, the standard error σ_M^2 of the posterior mean can be estimated by the batch means method and the cost $C(M)$ for computing M is the cpu time. Then

$$\text{ARE}(M_1 : M_2) \equiv \frac{\sigma_{M_1}^2 C(M_1)}{\sigma_{M_2}^2 C(M_2)}.$$

The estimated standard error of the posterior mean from the link function $t(7)$ will be used as baseline (i.e., M_2 in the above equation) and compared to other MCMC samplers (M_1). The run times may vary when coded in different languages, for example in C language, but the ARE should remain the same. To interpret the results in Table 5.5, for example, to obtain an estimate of a given accuracy the ARS within the Gibbs sampler requires 1.84 times the cpu time as required using the $t(7)$ link function.

M_1	ARS within Gibbs	M. ARS within Gibbs	ARS B. Gibbs Size = 2	M. ARS B. Gibbs Size = 2	R.W. Chain Size = 6	IND. Chain Size = 6	KOL. Size = 6
ARE	1.84	1.22	5.28	2.19	1.01	1.31	2.76

Table 5.5. Asymptotic relative efficiency of MCMC samplers compared to $t(7)$.

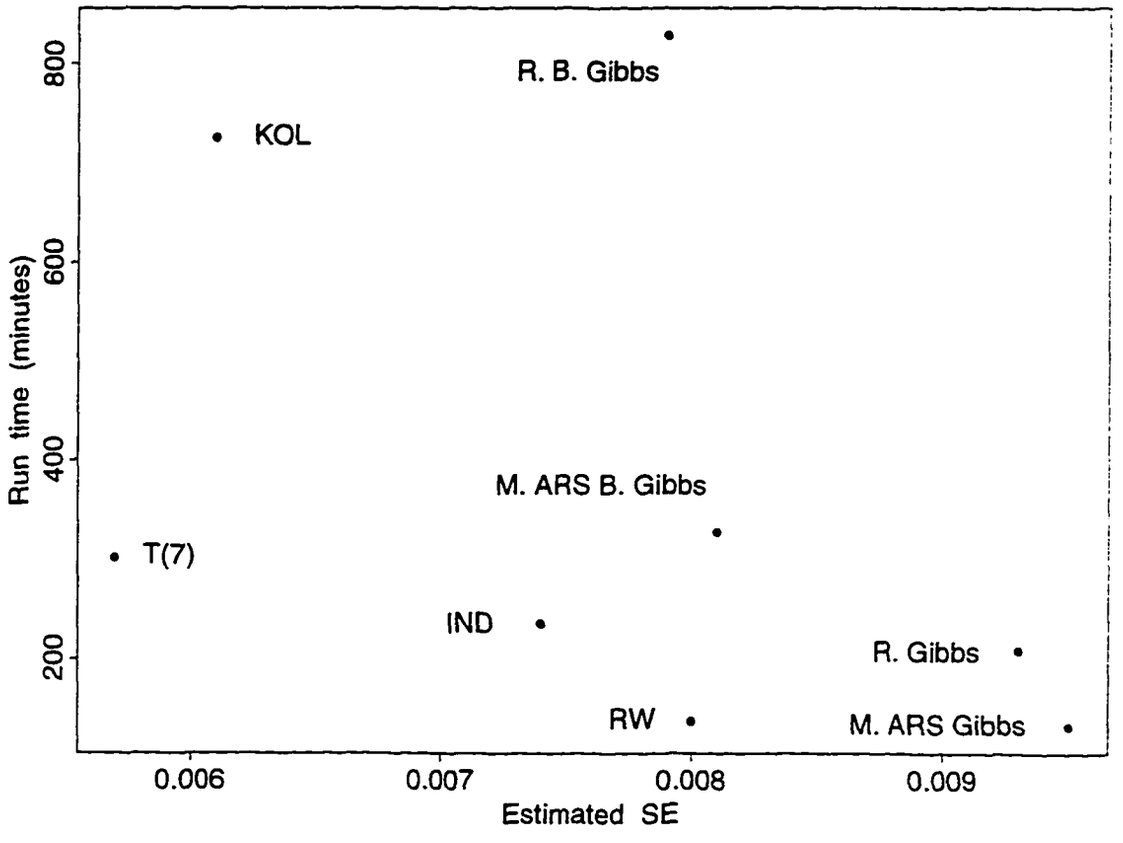


Figure 5.9. Comparison of MCMC samplers.
R. Gibbs - Rejection within Gibbs.
M. ARS Gibbs - Modified ARS within Gibbs.
R. B. Gibbs - Rejection within Block Gibbs, Size = 2.
M. ARS B. Gibbs - Modified ARS within Block Gibbs, Size = 2.
RW - Random Walk Chain, Size = 6.
IND - Independence Chain, Size = 6.
T(7) - Student t Distribution with 7 degrees of freedom, Size = 6.
KOL - Kolmogorov Distribution, Size = 6.

5.6 Summary

The main conclusions of this binary data analysis are as follows:

1. The values of the estimated posterior moments by the modified ARS algorithm within the Gibbs sampler are similar to the Gibbs sampler. but the run time is much faster than the Gibbs sampler with rejection sampling. This modified ARS algorithm has two advantages: (1) decreasing the iteration time and (2) generating samples from the same proposed distribution with minor adjustment. In practice, it is a very efficient algorithm within the Gibbs sampler. A disadvantage is that the simulations are complicated to code for block sizes higher than 2.

2. In general, the Metropolis algorithms perform better than the Gibbs samplers with rejection sampling. In the comparison of chains of length 50.000 the estimated standard errors of the Gibbs sampler with the modified ARS algorithm is about the same as for the Metropolis algorithms, whereas the estimated standard errors of the Metropolis algorithms perform somewhat better than the Gibbs sampler with the modified ARS algorithm. From these results. it is clear that the Metropolis algorithms have quickly and accurately produced a posterior distribution.

3. The run time of the $t(7)$ link function is slightly longer than the Gibbs sampler and Metropolis algorithms, but the estimated standard error is much better than the Gibbs sampler and Metropolis algorithms. Although the mixtures of normal distribution $2ZK$ performs similarly to $t(7)$ link function. the run time is much slower than $t(7)$ link function. More efficient code for generating the asymptotic Kolmogorov distribution is apparently needed. Overall. the two link functions $t(7)$ and Kolmogorov perform better than the Gibbs sampler and

Metropolis algorithms. The data augmented approach seems to be useful in the analysis of dynamic binary logit models based on the comparison of chains of length 50,000.

4. For comparison of all methods, we focus on the two results: estimated standard error and run time. Clearly, in Figure 5.9 the two link functions have the lowest estimated standard error, the two Gibbs samplers have the largest estimated standard error, and the Metropolis algorithms and the two block Gibbs samplers lie in between. The Metropolis algorithms and the two block Gibbs sampler have similar estimated standard errors, but the Metropolis algorithms apparently have shorter run times. To obtain an estimate of a given accuracy the Metropolis algorithms and modified ARS within the block Gibbs sampler require similar but higher cpu times as required using the $t(7)$ link function. The ARS within the block Gibbs sampler requires 5 times the cpu time as required using the $t(7)$ link function to obtain an estimate of a given accuracy.

5. The estimated standard deviations of the parameters using different MCMC algorithms were noticeably different from the estimates obtained by the EKFS algorithm. Also, the means for the parameters from the EKFS algorithm were noticeably smoother than the ones from the MCMC methods. This is shown in Table 5.2, 5.3, 5.4, and Figure 5.3, in which the difference is not within the reasonable accepted range of a standard error of the estimated standard deviations of any of the MCMC algorithms. This suggested that the estimated standard deviations are under-estimated when applying EKFS algorithm.

Chapter 6

Binary Events Analysis of Two MIRP Examples

In this chapter, we re-analyze the two MIRP studies introduced in Chapter 1 using the MCMC algorithms introduced in Chapter 4.

6.1 Co-Evolution Model

6.1.1 Cochlear Implant Evolution

There is growing scholarly interest in a social evolutionary theory of change for explaining how technological and institutional innovations emerge as a continuous process of variation, selection, and retention. Most organizational models of social evolution assume that variations are exogenous shocks that emerge by random chance, and have focused on the selection and retention processes of pre-existing variations. Also, most organization applications of this social evolutionary theory have treated variation, selection, and retention as a sequence of three discrete events. Van de Ven and Garud (1992) argued that variation, selection, and retention processes are misconstrued when viewed as three discrete events; instead, they are better understood as a cumulative progression of numerous interrelated acts of variation, selection, and retention over an extended period of time. Specifically, Van de Ven and Garud (1992) test the hypothesis that variations, selection, and retention events are endogenously related and co-produce each other over time.

The analysis is based on a longitudinal real time study conducted from 1983

to 1989 of 719 events observed in the development and commercialization of cochlear implants, which is a biomedical innovation that provides hearing to profoundly deaf people. The sources of events included: direct field observations and attendance at trade conferences where numerous interviews were conducted with actors from different organizations involved in different functions of cochlear implant development. reviews of trade literature, monthly observations of day-long management meetings of one of the firms involved in this innovation, as well as the administration of standardized questionnaires and interviews every six to twelve months with key actors involved in the innovation. Events were coded according to whether they pertained to novel technical variations, institutional rule making (selection), and institutional rule following (retention) events. The following structural system of three simultaneous time series equations was developed to test the hypothesis that novel technical variations, institutional rule making (selection), and institutional rule following (retention) events endogenously co-evolve over time to develop and commercialize cochlear implants.

The Co – Evolution Model:

$$Y_{1t} = \alpha_1 + \beta_1 Y_{2,t-1} - \beta_2 Y_{3,t-1} + \epsilon_1$$

$$Y_{2t} = \alpha_2 + \beta_3 Y_{1,t-1} + \beta_4 Y_{3,t-1} + \epsilon_2$$

$$Y_{3t} = \alpha_3 - \beta_5 Y_{1,t-1} + \beta_6 Y_{2,t-1} + \epsilon_3$$

where

$Y_1 =$ monthly count of the number of variation events

$Y_2 =$ monthly count of the number of selection events

$Y_3 =$ monthly count of the number of retention events

$\alpha_i =$ constant terms

$\beta_i =$ parameters

$\epsilon_i =$ error terms

The hypothesized relationships in the above equations are consistent with most organizational models of evolution in the following respects. Events representing novel technical variations precipitate selection events of institutional rule making, and the subsequent events of following rules serve as the retention mechanism of those rules that were selected. However, the hypothesized model of endogenous co-evolution departs from most organizational models of evolution in two fundamental respects. First, novel technical variations are endogenous (not exogenous) to the model. The likelihood of novel technical variations not only increases with institutional rule making events, but also decreases with institutional rule following events. Second, retention is commonly argued to suppress the subsequent selection of new variations. This negative feedback from retention to selection most likely operates during periods of incremental refinements of a population or a dominant design.

In order to apply regular time series analysis methods, it was necessary to aggregate the event sequence data into fixed temporal intervals. A monthly interval was chosen for regression analysis of the co-evolution model (Van de Ven and Garud, 1992). The results of the time series regression analyses of the co-evolution model by Van de Ven and Garud (1992) are present in Table 6.1. For each equation the table shows the regression coefficient and its standard error. The results in Table 6.1 is consistent with most organizational formulations of the evolutionary model, technical variation events lead to subsequent institutional rule making selection events, and the latter significantly predict institutional rule following retention events. However, Table 6.1 contains two notable empirical findings that are contrary to most formulations of organizational evolution, but are consistent with the endogenous co-evolution model. First, novel technical variations are not exogenous to the model since they are significantly predicted by prior rule making institutional selection events. Second, there is a significant self-reinforcing loop between institutional rule making selection

events and rule following retention events. While it is commonly expected that institutional selection events lead to rule following retention events, the feedback effects of rule following events on subsequent rule making events are even stronger.

Contrary to the hypothesized co-evolution model as well as most formulations of organizational evolution, Table 6.1 shows there are no negative relationships between technical variation events and institutional rule following events. Thus, rule following retention events did not serve to counteract the self-reinforcing loop between technical variation and institutional selection.

	Dependent Variables					
	(Y1) Variation Events at t		(Y2) Selection Events at t		(Y3) Retention Events at t	
Independent Variables	$\hat{\beta}$	(S.E.)	$\hat{\beta}$	(S.E.)	$\hat{\beta}$	(S.E.)
(Y1) Variations Events at (t-1)	NA		0.55	(0.10)	0.14	(0.09)
(Y2) Selection Events at (t-1)	0.43	(0.05)	NA		0.20	(0.07)
(Y3) Retention Events at (t-1)	0.02	(0.06)	0.42	(0.09)	NA	
Constant	0.04	(0.08)	0.27	(0.12)	0.46	(0.12)

Table 6.1. Van de Ven and Garud's (1992) parameter estimates (standard error in parentheses) of the time series regression analysis of the co-evolution model. NA = Independent variable not included in the regression equation.

6.1.2 Binary Events Analysis of the Co-Evolution Model

As argued in Chapter 1, in some cases aggregation may diminish the direct relationships among variables. Therefore, the above statistical findings will now be examined in terms of binary events instead of monthly events in the period of the development of the cochlear implant technology and industry. Since the collection times of events are irregular, the time intervals of events are not fixed. Their range is from 1 day to

270 days. In the following analysis, we chose to treat the time interval of each event as fixed and concentrate on the effects of the independent variables from previous periods over and above the direct contributions of the dependent variable in the given period. In developing a process theory of innovation, the changing structural and technological conditions, individual behavior, and attitude may cause uncertainty to the hypothesized model of the process theory. For time series data of this kind, application of standard static regression models will not be appropriate. Therefore, we analyze the data with the following dynamic multivariate logistic model (6.1).

If k is the number of categories, responses $\boldsymbol{y}_t$ can be described by a vector $\boldsymbol{y}_t'$ $= (y_{1t}, \ldots, y_{qt})$, with $q = k$ - 1 components. If only one multi-categorical observation is made for each t, then $y_{jt} = 1$ if category j has been observed, and $y_{jt} = 0$ otherwise, $j = 1, \ldots, q$. The models are completely determined by the corresponding response probabilities $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{qt})$, specified by $\boldsymbol{\mu}_t = \boldsymbol{h}(\boldsymbol{\eta}_t) = \boldsymbol{h}(\boldsymbol{Z}_t'\boldsymbol{\beta}_t)$. A dynamic multivariate logistic model is specified by

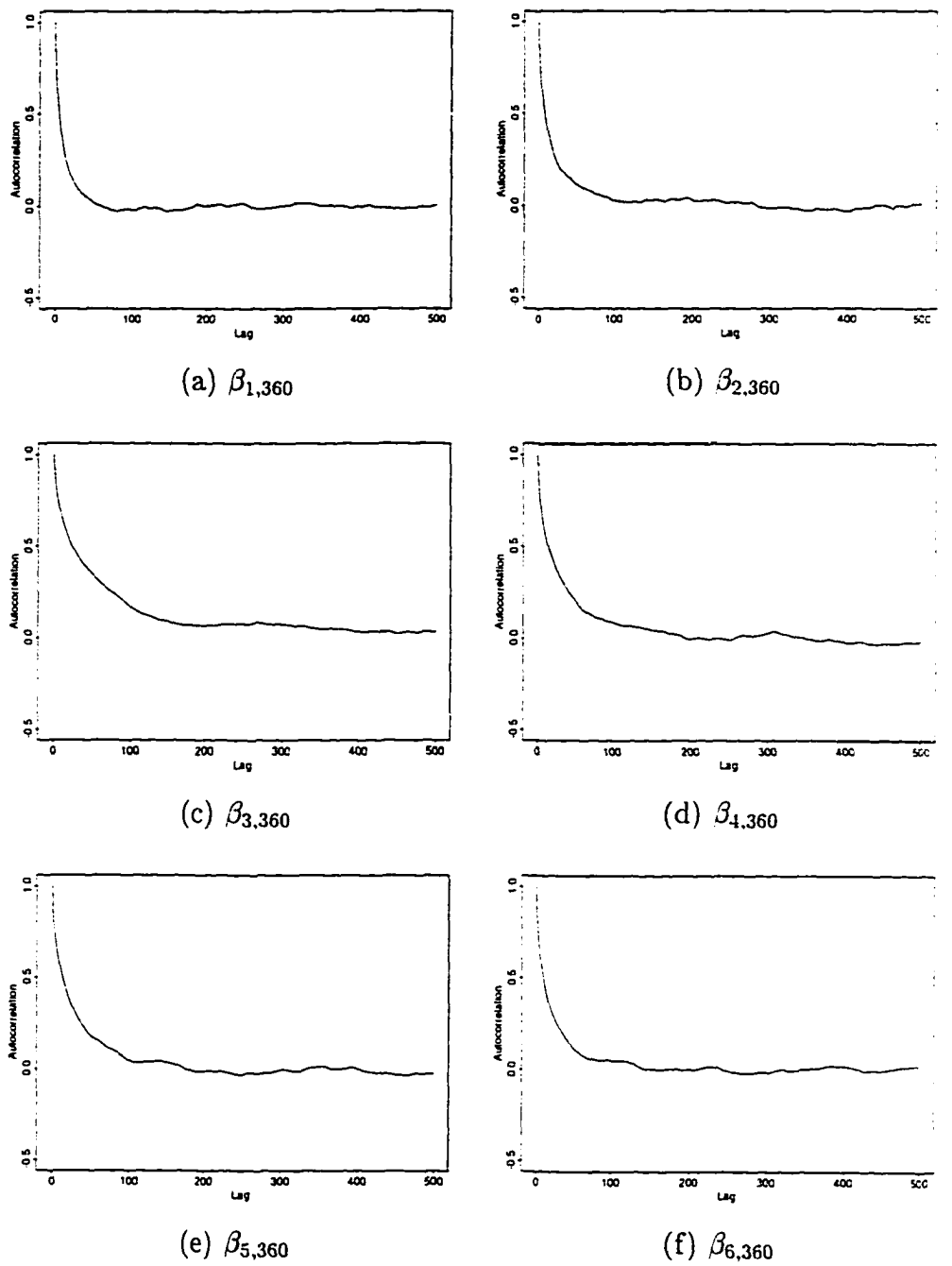
$$\mu_{jt} = \frac{\exp(\eta_{jt})}{1 + \sum_{r=1}^{q} \exp(\eta_{rt})} \tag{6.1}$$

The response variable $\boldsymbol{y}$ has three possible outcomes: variation, selection, retention, and no occurrence which are labeled with 1 to 4, thus having $\boldsymbol{y} \in \{1, 2, 3, 4\}$. Since no multiple events occur for each observation time t, we introduce a multivariate response vector of dummy variables $\boldsymbol{y}_t = (y_{1t}, y_{2t}, y_{3t})$ to take into account the categorical character of $\boldsymbol{y}_t$. Let $y_{1t} = 1$ if the variation event occurred in the tth event, $y_{2t} = 1$ if the selection event occurred in the tth event, and $y_{3t} = 1$ if the retention event occurred in the tth event. Assuming dummy coding, no occurrence of y_1, y_2, and y_3 leads to $\boldsymbol{y}_t = (0, 0, 0)$. We analyzed the cochlear implants data using

a dynamic multivariate logistic model together with a random walk evolution model:

$$
\begin{aligned}
&\boldsymbol{y}_t \sim \text{Multinomial}(1, \boldsymbol{\mu}_t) \quad \text{where} \quad \boldsymbol{\mu}_t = (\mu_{1t}, \mu_{2t}, \mu_{3t}).\\
&\mu_{jt} = \frac{\exp(\eta_{jt})}{1+\sum_{q=1}^{3}\exp(\eta_{qt})}, \quad j = 1, 2, 3,\\
&\boldsymbol{\eta}_t' = (\eta_{1t}, \eta_{2t}, \eta_{3t}) = \boldsymbol{Z}_t'\boldsymbol{\beta}_t\\
&\boldsymbol{Z}_t' = \begin{bmatrix} 1 & y_{2,t-1} & y_{3,t-1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & y_{1,t-1} & y_{3,t-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & y_{1,t-1} & y_{2,t-1} \end{bmatrix}.\\
&\boldsymbol{\beta}_t' = \begin{bmatrix} \alpha_{1t} & \beta_{1t} & \beta_{2t} & \alpha_{2t} & \beta_{3t} & \beta_{4t} & \alpha_{3t} & \beta_{5t} & \beta_{6t} \end{bmatrix},\\
&\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t), \quad \boldsymbol{\beta}_0 \sim N(\boldsymbol{a}_0, \boldsymbol{Q}_0).
\end{aligned}
$$

The priors for the hyperparameters $\boldsymbol{a}_0$, $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_t$ are specified by $[\boldsymbol{a}_0] \sim N(\mu_{a_0}, \sigma^2_{a_0})$. $\boldsymbol{Q}_0 = c\sigma^2 I_9 = c\boldsymbol{Q}_t$, and $[\sigma^2] \sim \text{IG}(a_1, b_1)$, where IG denotes the inverse gamma distribution and μ_{a_0}, $\sigma^2_{a_0}$, c, a_1, and b_1 are assumed known. For the binary event data of the co-evolution model, the hyperparameter prior specification was defined by $\mu_{a_0} = 0$, $\sigma^2_{a_0} = 0.01 I_9$, $a_1 = 2$, $b_1 = 0.0125$, and $c = 2$. Using this initial information, a MCMC sampler independence chain based on 50,000 runs was applied to the data. Based on a preliminary sample of 5,000 observations, the acceptance rate of the block size 4 is 0.23. Therefore the block size of 4 is chosen based on Roberts, Gelman, and Gilk's (1994) suggestion that the optimal acceptance rate is around 0.25. To monitor the performance of the samplers, the autocorrelation curves for the parameters $\beta_{1,360}$, $\beta_{2,360}$, $\beta_{3,360}$, $\beta_{4,360}$, $\beta_{5,360}$, and $\beta_{6,360}$ (the time $t = 360$ is the middle of the posterior mean sequence for each parameter) based on 50,000 runs are shown in Figure 6.1 (a) - (f). The autocorrelation curves are significantly nonzero only out to about lag 150 for all parameters. This suggests that 50,000 runs is a sufficiently large number to obtain accurate estimates of all parameters.



(a) $\beta_{1,360}$ (b) $\beta_{2,360}$

(c) $\beta_{3,360}$ (d) $\beta_{4,360}$

(e) $\beta_{5,360}$ (f) $\beta_{6,360}$

Figure 6.1. Empirical autocorrelation curve of the parameters β_1, β_2, β_3, β_4, β_5, and β_6.

According to (6.1) the model can be written as

$$\log\frac{P(\text{occurrence of variation at } t)}{P(\text{no occurrence})} = \alpha_{1t} + \beta_{1t}y_{2,t-1} + \beta_{2t}y_{3,t-1}. \quad (6.2)$$

$$\log\frac{P(\text{occurrence of selection at } t)}{P(\text{no occurrence})} = \alpha_{2t} + \beta_{3t}y_{1,t-1} + \beta_{4t}y_{3,t-1}. \quad (6.3)$$

$$\log\frac{P(\text{occurrence of retention at } t)}{P(\text{no occurrence})} = \alpha_{3t} + \beta_{5t}y_{1,t-1} + \beta_{6t}y_{2,t-1}. \quad (6.4)$$

In the above equations (6.2), (6.3), and (6.4), "no occurrenc" stands for no occurrence of variation, selection, and retention events. To be consistent with the co-evolution model, the $y_{1,t-1}$, $y_{2,t-1}$, and $y_{3,t-1}$ were not included in the equation (6.2), (6.3), and (6.4) respectively, to adjust for the effect of its prior event. Several hypothesized relationships among variation, selection, and retention events have been proposed by Van de Ven and Garud (1992) and presented in the co-evolution model. To test these hypotheses, the above equations (6.2), (6.3), and (6.4) can be used to examine the hypothesized relationships. In more detail, for example, the odds of variation event y_{1t} occurrence is $\exp(\alpha_{1t}+\beta_{1t})$ times higher with selection event $y_{2,t-1}$ present than with no selection event $y_{2,t-1}$ present. Thus, if the odds (i.e., the value of $\exp(\alpha_{1t}+\beta_{1t})$) is higher than 1, the odds of the variation event occurring at time t increases when a selection event occurred at time t - 1; if the odds is below than 1, the odds of the variation event occurring at time t decreases when a selection event occurred at time t - 1.

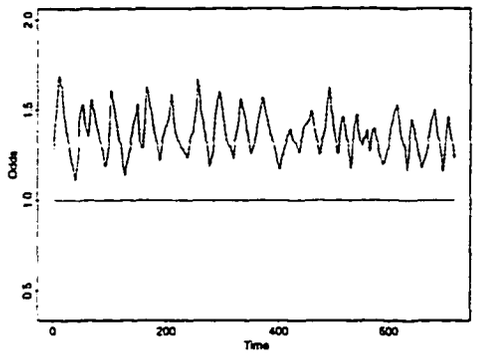
With selection event present the odds, $\exp(\alpha_{1t}+\beta_{1t})$, of variation event for t = 1, ..., 719, is displayed in Figure 6.2(a). Since all the odds are above one, the odds of variation event occurring at time t increases with selection event present at time t

- 1. This result agrees with the co-evolution model that the novel technical variations are not exogenous to the model since the odds of novel technical variations increases with institutional rule making (selection) events.

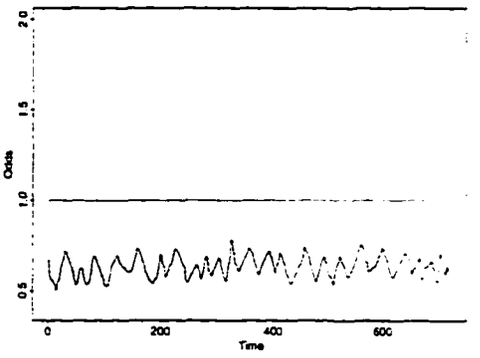
With retention event present the odds of variation event occurring with retention event present and the odds of retention event occurring with variation event present are displayed in Figure 6.2(b) and 6.2(e), respectively. In both figures, all the odds are below one. These results support the hypothesized relationships of the co-evolution model that the odds of variation event occurring at time t decreases with institutional rule following (retention) event present at time t - 1 and the odds of institutional rule following (retention) event occurring at time t decreases with variation event present at time t - 1. This is not shown in the results (Table 6.1) of the time series regression analyses using monthly aggregation events.

Figure 6.2(c) shows the odds of selection event occurring with variation event present. Most of the odds are above one, the odds of selection event occurring at time t increases with the variation event present at time t - 1. In Figure 6.2(f), all the odds of retention event occurring with selection event present are above one. This also indicates that the odds of retention event occurring at time t increases with the selection event present at time t - 1. These two results are consistent with most organizational formulations of the evolutionary model as well as the co-evolution model.

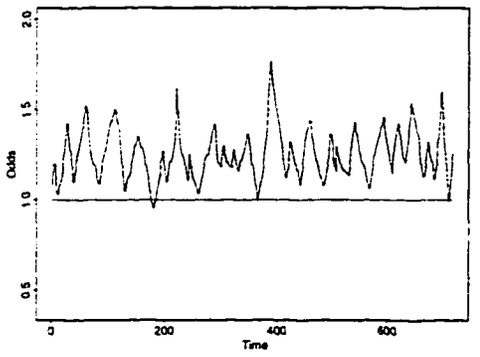
Figure 6.2(d) shows the odds of selection event occurring with retention event present. Contrary to the hypothesized co-evolution model, Figure 6.2(d) does not show that the odds of institutional rule making (selection) event occurring at time t increases with institutional rule following (retention) event present at time t - 1.



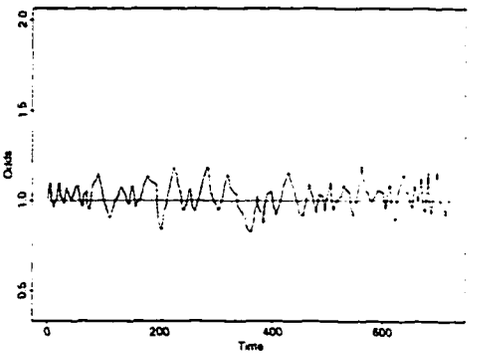
(a) Odds of variation event occurred with selection event present



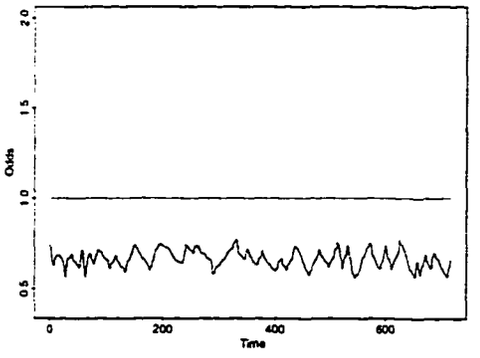
(b) Odds of variation event occurred with retention event present



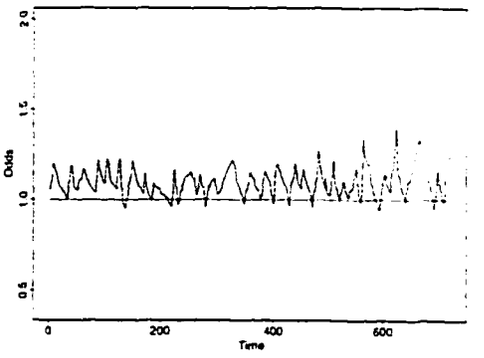
(c) Odds of selection event occurred with variation event present



(d) Odds of selection event occurred with retention event present



(e) Odds of retention event occurred with variation event present



(f) Odds of retention event occurred with selection event present

Figure 6.2. Odds of variation, selection, and retention events.

6.2 Adaptive Learning Model

6.2.1 Adaptive Processes of Organizational Learning during the Development of TAP

A central problem in managing and investing innovations is determining whether and how to continue a developmental effort in the absence of concrete performance information. Adaptive processes of organizational learning have recently gained increasing prominence to address this kind of problem. This adaptive learning model assumes organizations to be target-oriented, routine-based systems which respond to experience by repeating behaviors that have been found to be successful and avoiding those which have not. This basic model has proven quite robust in situations where preferences are clear, alternative courses of action are specified in advance, and outcomes are unambiguous (March, 1972). But very few studies have examined the empirical validity of this model in more ambiguous organizational settings. Van de Ven and Polley (1992) examined this process of adaptive learning during the development of a biomedical innovation and tested the model of adaptive learning in a more ambiguous real-world organizational setting, such as innovation development.

The model of adaptive learning that Van de Ven and Polley (1992) examined in their study of innovation development focused on the relationship between actions and outcomes. Following March (1972), Van de Ven and Polley (1992) assumed that people are adaptively rational. To develop an innovation, entrepreneurs initially choose a course of action (for example A) with the intention of achieving a positive outcome. If a positive outcome is experienced following action course A they will continue with A, and if a negative outcome is experienced they will change or shift to a new course of action (for example B). Subsequently, if positive outcomes are experienced with action course B they will continue with B, but if negative outcomes

are experienced they will change again to another course of action (for example C). which may appear as the next best alternative course at that time. To model this process, Van de Ven and Polley (1992) observed and categorized the actions that entrepreneurs take as either continuing or changing their prior course of action. and also observed if entrepreneurs experienced positive or negative outcomes following their prior actions. When outcome responses follow prior actions. entrepreneurs are disposed to adapt to four possible situations which are described in Table 6.2:

If the action at time t-1 was:	and the outcome of that action was:	then the next action at time t will be:
1. continue prior action	positive	continue prior action again
2. continue prior action	negative	change prior action
3. change prior action	positive	continue prior action
4. change prior action	negative	change prior action again

Table 6.2. Process of learning.

A simple direct effect of prior outcomes on subsequent actions. as suggested by the four situations in the Table 6.2, may not indicate that learning occurred because the outcomes that trigger subsequent actions may not have been caused by the prior actions taken. Thus, Van de Ven and Polley (1992) proposed that adaptive learning is evident when prior actions and outcomes interact to explain subsequent action. This leads to the hypothesis on adaptive learning:

Hypothesis: *Adaptive learning is evident when prior actions and outcomes at time* $t-1$ *have a positive interaction effect on continuing the course of action at* t.

A test of this hypothesis in the learning model is based on a longitudinal study of the development of an innovation, which was undertaken as a joint venture

by three corporations to create a business by developing a new medical technology called therapeutic apheresis. This real-time field study of the therapeutic apheresis program (TAP) was conducted from October 1983 to July 1988. Data collection involved: attendance and recording of proceedings at bimonthly meetings of the TAP strategic business unit (SBU) committee and semiannual administrative reviews of TAP; semiannual interviews with TAP SBU members and questionnaire surveys of all key TAP personnel; annual interviews with top managers of the co-venturing firms: as well as information obtained from company records and industry trade publications. Over the five years of real-time tracking, 258 events were recorded in therapeutic apheresis development. These events are the units of observation for testing the learning model.

A time series analysis of the events was undertaken to estimate the relationships among the variables in the hypothesized learning model. To apply regular time series analysis methods, it was necessary to aggregate the event sequence data into fixed temporal intervals. A monthly interval was chosen for regression analysis of the adaptive learning model (Van de Ven and Polley, 1992). Two temporal periods that have been observed: (1) a startup expansion period from November 1983 to September 1986 (186 events) when the TAP innovation entered the market, followed by (2) an ending contraction period from October 1986 to July 1988 (72 events) when TAP's development was terminated. Very different patterns of correlations were found (Van de Ven and Polley, 1992) between the monthly time series of actions and outcomes during the expansion and contraction periods. Given these differences, Van de Ven and Polley (1992) examined the hypothesized adaptive learning model separately in each period.

The test results in Van de Ven and Polley's (1992) study show little support for the hypothesized learning model during the expansion period but are consistent with hypothesized learning model during the contraction period. During the expansion

period, negative outcomes lead directly to continuing with the prior course of action which do not support the four situations described in Table 6.2. Also, the absence of an interaction effect of prior actions and outcomes on subsequent actions also indicates no learning occured. During the contraction period, adaptive learning is evident by the fact that subsequent actions are explained by a significant positive interaction effect of prior actions and outcomes. The significant positive relationship between continuing prior actions and subsequent outcomes indicates an increase in the propensity of entrepreneurs to select the course of action that was rewarded. This result supports the hypothesis on adaptive learning. Prior outcomes alone have no effect on subsequent actions which also do not support the four situations described in Table 6.2.

6.2.2 Binary Events Analysis of Learning Model

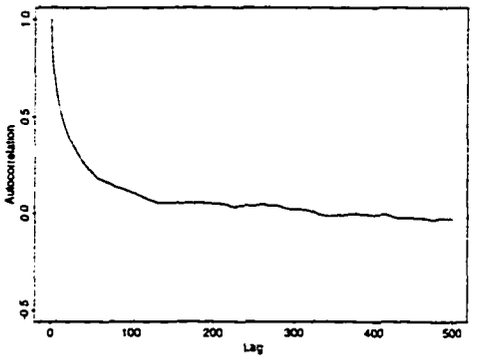
The above statistical findings will now be examined in terms of binary events instead of monthly events in each period of the development of TAP. Since the collection times of events are irregular, the time intervals of events are not fixed. In the following analysis, we again choose to treat the time intervals of each event as fixed and concentrate on the relationships between actions and outcomes for each period. We analyze the data with a dynamic multivariate logistic model (6.1) which is constructed as follows.

The response variable $\boldsymbol{y}$ has three possible outcomes: continue actions, change actions, and no actions which are labeled with 1 to 3, thus having $\boldsymbol{y} \in \{1, 2, 3\}$. We introduce a multivariate response vector of dummy variables $\boldsymbol{y_t} = (y_{1t}, y_{2t})$ to take into account the categorical character of $\boldsymbol{y_t}$. Let $y_{1t} = 1$ if the continue action occurred in the tth event, and $y_{2t} = 1$ if the change action occurred in the tth event. Assuming dummy coding, no occurrence of y_1, and y_2 leads to $\boldsymbol{y_t} = (0, 0)$. There are

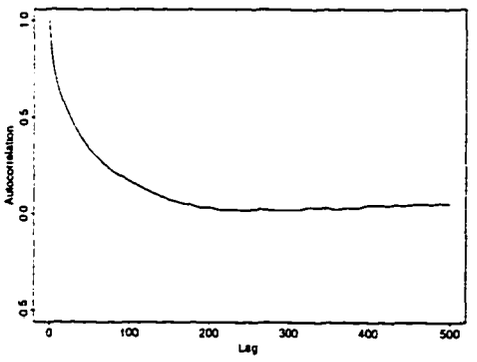
two binary covariates, namely, outcome positive (X_1), and outcome negative (X_2). We analyzed TAP data by a dynamic multivariate logistic model together with a random walk evolution model for the coefficients:

$$
\begin{aligned}
&\boldsymbol{y}_t \sim \text{Multinomial}(1, \boldsymbol{\mu}_t) \quad \text{where } \boldsymbol{\mu}_t = (\mu_{1t}, \mu_{2t}),\\
&\mu_{jt} = \frac{\exp(\eta_{jt})}{1+\sum_{q=1}^{2}\exp(\eta_{qt})}, \quad j = 1, 2,\\
&\boldsymbol{\eta}_t' = (\eta_{1t}, \eta_{2t}) = \boldsymbol{Z}_t'\boldsymbol{\beta}_t\\
&\boldsymbol{Z}_t' = \begin{bmatrix} 1 & x_{1,t-1} & y_{1,t-2} \times x_{1,t-1} & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{2,t-1} \end{bmatrix}.\\
&\boldsymbol{\beta}_t' = \begin{bmatrix} \alpha_{1t} & \beta_{1t} & \beta_{2t} & \alpha_{2t} & \beta_{3t} \end{bmatrix},\\
&\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t), \quad \boldsymbol{\beta}_0 \sim N(\boldsymbol{a}_0, \boldsymbol{Q}_0).
\end{aligned}
$$

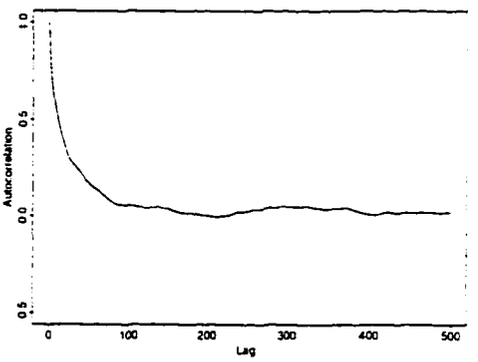
The priors for the hyperparameters $\boldsymbol{a}_0$, $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_t$ are specified by $[\boldsymbol{a}_0] \sim N(\mu_{a_0}, \sigma^2_{a_0})$. $\boldsymbol{Q}_0 = c\sigma^2 I_5 = c\boldsymbol{Q}_t$, and $[\sigma^2] \sim \text{IG}(a_1, b_1)$, where IG denotes the inverse gamma distribution and μ_{a_0}, $\sigma^2_{a_0}$, c, a_1, and b_1 are assumed known. For the binary event data of the co-evolution model, the hyperparameter prior specification was defined by $\mu_{a_0} = 0I_5$, $\sigma^2_{a_0} = 0.02I_5$, $a_1 = 2$, $b_1 = 0.025$, and $c = 2$. Using these initial assumption, a MCMC sampler independence chain based on 50,000 runs was applied to the data. Based on a preliminary sample of 5,000 observations, the acceptance rate of the block size 8 is 0.28. Therefore the block size of 8 is chosen based on Roberts, Gelman, and Gilk's (1994) suggestion that the optimal acceptance rate is around 0.25. To monitor the performance of the samplers, the autocorrelation curves for the parameters $\beta_{1,129}$, $\beta_{2,129}$, and $\beta_{3,129}$ (the time $t = 129$ is the middle of the posterior mean sequence from each parameter estimates) based on 50,000 runs are shown in Figure 6.3 (a) - (c). The autocorrelation curves are significantly nonzero only out to about lag 200 for all parameters. This suggests that 50,000 runs is a sufficiently large number to obtain accurate estimates of all parameters.



(a) $\beta_{1,129}$



(b) $\beta_{2,129}$



(c) $\beta_{3,129}$

Figure 6.3. Empirical autocorrelation curve of the parameters β_1, β_2, and β_3.

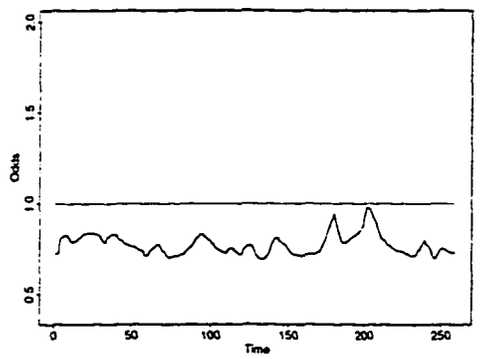
According to (6.1) the model can be written as

$$\log\frac{P(\text{occurrence of continue actions})}{P(\text{no occurrence})} = \alpha_{1t} + \beta_{1t}x_{1,t-1} + \beta_{2t}y_{1,t-2} \times x_{1,t-1}.$$
$$\log\frac{P(\text{occurrence of change actions})}{P(\text{no occurrence})} = \alpha_{2t} + \beta_{3t}x_{2,t-1}.$$

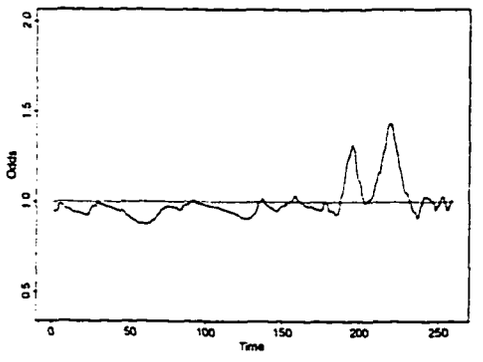
In the study of the adaptive learning model, Van de Ven and Polley (1992) focus on the relationship between actions and outcomes as well as the hypothesis on adaptive learning. To test these relationship and hypothesis, the above equations can be used to examine the hypothesized relationships. Figure 6.4(a) shows the odds of continue actions occurring with outcome positive present. All the odds are below one. the odds of continue actions occurring at time t decreases with outcome positive present at time t - 1. This result is same as the result of the time series regression analyses using monthly aggregation eventsand don't support the four situations of process of learning described in Table 6.2 for expansion and contraction periods.

Figure 6.4(b) shows the odds of continue actions occurred at time t with interaction effect continue actions at time t - 2 and outcome positive at time t - 1 present. This result is similar to the result of the time series regression analyses using monthly aggregation events. During the expansion period (event 1 to 186), the values of odds of continue actions in Figure 6.4(b) show little support for the hypothesized learning model. During the contraction period (event 187 to 258) adaptive learning is evident by the higher value of odds that subsequent continue actions are explained by a positive interaction effect of prior continue actions and outcome positive.

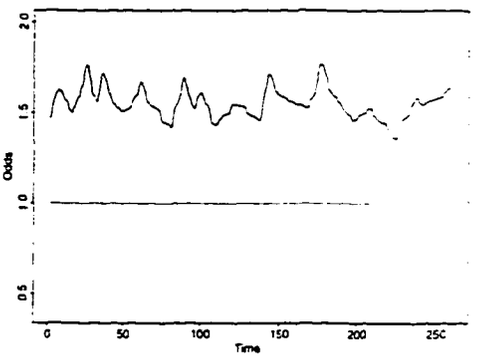
In Figure 6.4(c), the odds of change actions occurred at time t increase with outcome negative present at time t - 1 since all the odds are above one. This result differs from the result of the time series regression analyses using monthly aggregation events, but support the four situations of process of learning described in Table 6.2 for expansion and contraction periods.



(a) Odds of continue actions occurred with outcome positive present



(b) Odds of continue actions occurred with interaction present



(c) Odds of change actions occurred with outcome negative present

Figure 6.4. Odds of continue actions and change actions events.

6.3 Discussion

The results of binary events analyses of the co-evolution model and adaptive learning model refine application of evolutionary theory for understanding the technological and institutional development of cochlear implants and of adaptive processes of organizational learning during the development of a technological innovation of therapeutic apheresis in several important features.

1. There is dynamic nature in developing a process theory of innovation. The graphs in Figure 6.2 and 6.4 suggest that the relative occurrence is not fixed over time and the changing situations such as technological conditions, individual behavior, and attitude may cause uncertainty to the hypothesized model of the process theory.

2. Aggregation may diminish the direct relationships among variables. The results in Figure 6.2(b) and 6.2(e) support the hypothesized relationships of the co-evolution model which are not revealed in the results in Table 6.1 of the time series regression analyses using monthly aggregate data. Also the result in Figure 6.4(c) differs from the result of the time series regression analyses using monthly aggregation events, but support the four situations of process of learning described in Table 6.2 for expansion and contraction periods.

3. Different patterns of parameter effects are reflected in the binary events analysis. One can easily detect different patterns of relative occurrence with a specified event present by observing the plots in Figure 6.2 and Figure 6.4. These empirical findings of different patterns provide an enhanced knowledge for understanding the technological and institutional development of cochlear implants and different learning patterns of therapeutic apheresis in each period of developing an innovation.

Chapter 7

Conclusions and Future Research

This thesis has investigated and applied Markov chain Monte Carlo samplers for dynamic multivariate binary time series. This approach extends the Gibbs sampling framework for dynamic generalized linear models by introducing more general Markov chain methods. This thesis outlines several basic Markov chain Monte Carlo methods, including Metropolis-Hastings algorithms, adaptive rejection algorithms and other variations. From the results on a binary time series example, these algorithms have better performance than the Gibbs sampler. In addition, the basic formulation of the Gibbs sampler is restricted to problems where the complete conditional part of the posterior distribution are available. The generality of MCMC methods remove this restriction. As a result, Markov chain Monte Carlo methods seem to provide a more efficient tool for analyzing multivariate dynamic generalized linear models.

A modified ARS algorithm is derived in this thesis for efficiently sampling from log-concave distributions. In contrast to the ARS algorithm, this modified ARS algorithm only needs to start at one point instead of two points in ARS algorithm. Furthermore, if the target distribution is from an exponential family, the calculation of the rejection envelope and squeezing functions required in ARS algorithm can be omitted. Thus, this modified ARS algorithm within the Gibbs sampler is more efficient and simple than the ARS algorithm. The results on a binary time series example show that the rejection rate reduction obtained using this modified ARS algorithm within the Gibbs sampler is quite significant.

We have also applied Markov chain Monte Carlo samplers to two binary time

series which were previously analyzed by standard time series analysis methods on monthly data which were obtained from aggregating the binary events. The results show that the direct relationships among variables in binary event series are more likely to be detected using the raw event data than in monthly event series because the aggregation may diminish the direct relationships among variables. This work may provide a fundamental statistical tool in analyzing multivariate binary time series data that is commonly observed in the social sciences and is often recognized as having a dynamic nature, especially in developing a process theory of innovation.

Areas for future work include:

1. To incorporate several algorithms to form hybrid algorithms. Tierney (1994) outlines some of the basic Markov chain algorithms that are available and describes several methods and strategies in which the algorithms can be combined to form hybrid algorithms. This can be used to guide the construction of more efficient algorithms.

2. To study the effects of varying the parameters in MCMC samplers. More work is clearly needed to understand the effects of varying the parameters in the M-H and hybrid algorithms and to determine good default values for these parameters. For example, it would be nice to adjust the parameters of the proposal distribution in the M-H algorithm so as to make moves that are as large as possible while maintaining a reasonable acceptance rate.

3. To extend the current work to the case when the observations are not equally spaced in time. In the literature, little material on statistical inference or empirical learning is concerned with the analysis of unequally spaced data. To adjust the time factor into the hypothesized model is clearly a new area needing to be explored.

Bibliography

[1] Aitchison, J., and Bennett, J. A. (1970). Polychotomous Quantal Response by Maximum Indicant. *Biometrika*, 57, 253-262.

[2] Albert, J. H., and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*. 88. 669-679.

[3] Andrews, D. F., and Mallows, C. L. (1974). Scale Mixtures of Normal distributions. *Journal of the Royal Statistical Society, Series B*, 36, 99-102.

[4] Aoki, M. (1987). *State Space Modeling of Time Series*. Heidelberg: Springer.

[5] Besag, J., and Green, P. J. (1993). Spatial Statistics and Bayesian Computation. *Journal of the Royal Statistical Society, Series B*, 55, 25-37.

[6] Birnbaum, A., and Dudman, J. (1963). Logistic Order Statistics. *Annals of Mathematical Statistics*, 34, 658-663.

[7] Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association*, 87, 493-500.

[8] Carter, C. K., and Kohn, R. (1994). On Gibbs Sampling for State Space Models. *Biometrika*, 81, 541-553.

[9] Chen, M. H., and Schmeiser, B. (1993). Performance of the Gibbs, Hit-and Run, and Metropolis Samplers. *Journal of Computational and Graphical Statistics*. 2. 251-272.

[10] Cheng, Y. T., and Van de Ven, A. H. (1996). Learning the Innovation Journey: Order Out of Chaos. *Organization Science,* in press.

[11] Chib, S., and Greenberg, E. (1994). Understanding the Metropolis-Hastings Algorithm. *American Statistician,* 49, 327-335.

[12] Cox, D. R. (1970). *Analysis of Binary Data.* London: Chapman & Hall.

[13] Cox, D. R. (1972). The Analysis of Multivariate Binary Data. *Journal of the Royal Statistical Society, Series C,* 321, 113-120.

[14] Conolly, M. A., and Liang, K. Y. (1988). Conditional Logistic Regression Models for Correlated Binary Data. *Biometrika,* 75, 501-506.

[15] De Jong, P., and Shephard, N. (1995). The Simulation Smoother for Time Series Models. *Biometrika,* 82, 339-350.

[16] Dellaportas, P., and Smith, A. F. M. (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics.* 42, 443-459.

[17] Devroye, L. (1986). *Non-uniform Random Variate Generation.* New York: Springer.

[18] Edwards, R. G., and Sokal, A. D. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang Representation and Monte Carlo Algorithm. *Physical Review,* D, 38, 2009-2012.

[19] Fahrmeir, L. (1992). Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models. *Journal of the American Statistical Association,* 87, 501-509.

[20] Fahrmeir, L., and Goss, M. (1992). On Filtering and Smoothing in Dynamic Models for Categorical Longitudinal Data. In: Heijden, P. v.d., Jansen, W., Francis, B., and Seeber, G. U. H. (Eds.). *Statistical Modelling*, 85-94. Amsterdam: North Holland.

[21] Fahrmeir, L., Hennevogl, W., and Klemme, K. (1992). Smoothing in Dynamic Generalized Linear Models by Gibbs Sampling. In: Fahrmeir, L., Francis, B., Gilchrist, R., and Tutz, G. (Eds.). *Advances in GLIM and Statistical Modelling*. 85-90. Heidelberg: Springer.

[22] Fahrmeir, L., and Kaufmann, H. (1991). On Kalman Filtering. Posterior Mode Estimation and Fisher Scoring in Dynamic Exponential Family Regression. *Metrika*, 38, 37-60.

[23] Fahrmeir, L., and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

[24] Frühwirth-Schnatter, S. (1991). Monitoring von kologischen und biometrischen Prozessen mit statistischen Filtern. In: Minder, Ch., and Seeber, G. (Eds.) *Multivariate Modelle: Neue Ansätze für biometrische Anwendungen*. Heidelberg: Springer Lecture Notes.

[25] Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*. 85, 398-409.

[26] Gelman, A. (1992). Iterative and Non-iterative Simulation Algorithms. In: *Computing Science and Statistics. Proceedings of the 24th symposium on the Interface*. 433-438.

[27] Gelman, A., and Rubin, D. B. (1992a). A Single Series from the Gibbs Sampler Provides a False Sense of Security. In: Berger, J. O., Bernardo, J. M., Dawid, A. P., and Smith, A. F. M. (Eds.). *Bayesian Statistics 4*, 625-631. Oxford: Oxford University Press.

[28] Gelman, A., and Rubin, D. B. (1992b). Inference from Iterative Simulation using Multiple Sequences (with discussion). *Statistical Science*, 7, 457-511.

[29] Geman, S., and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[30] Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In: Berger, J. O., Bernardo, J. M., Dawid, A. P., and Smith, A. F. M. (Eds.). *Bayesian Statistics 4*, 169-193. Oxford: Oxford University Press.

[31] Geweke, J. (1989). Bayesian Inference in Econometric Models using Monte Carlo Integration. *Econometrica*, 57, 1317-1339.

[32] Geyer, C. J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*, 7, 473-511.

[33] Geyer, C. J. (1991). Markov Chain Monte Carlo Maximum Likelihood. In: Keramides, E. M. (Ed.). *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156-163.

[34] Gilks, W. R. (1992). Derivative-free Adaptive Rejection Sampling for Gibbs Sampling. In: Berger, J. O., Bernardo, J. M., Dawid, A. P., and Smith, A. F. M. (Eds.). *Bayesian Statistics 4*, 641-649. Oxford: Oxford University Press.

[35] Gilks, W. R., and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics,* 41, 337-348.

[36] Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Applied Statistics,* 44, 455-472.

[37] Glaser, B. G., and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago: Aldine.

[38] Goss, M. (1990). Schätzung und Identifkation von Struktur- und hyperstrukturparametern in dynamischen generalisierten linearen Modellen. Dissertation. Universität Regensburg.

[39] Green, P. J., and Han, X. L. (1992). Metropolis Methods, Gaussian Proposals, and Antithetic Variables. In: Barone, P., Frigessi, A., and Piccioni, M. (Eds.). *Lecture Notes in Statistics,* 142-164. Berlin: Springer.

[40] Hartigan, J. A. (1969). Linear Bayesian Methods. *Journal of the Royal Statistical Society, Series B,* 31, 446-454.

[41] Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

[42] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika,* 57, 97-109.

[43] Hausman, J. A., and Wise, D. A. (1978). A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica,* 46, 403-426.

[44] Hodges, P. E., and Hale, D. F. (1993). A Computational Method for Estimating Densities of non-Gaussian Nonstationary Univariate Time Series. *Journal of Time Series Analysis*, 14, 163-178.

[45] Ishiguro, M., and Sakamoto, Y. (1983). A Bayesian Approach to Binary Response Curve Estimation. *Annals of the Institute of Statistical Mathematics*. 35-B. 115-137.

[46] Jensen, C. S., Kong, A., and Kjærulef, U. (1993). Blocking Gibbs Sampling in very Large Probabilistic Expert Systems. Technical Report R-93-2031. Institute Electronic Systems, Department of Mathematics and Computer Science. University of Aalborg.

[47] Johnson, N., and Kotz, S. (1970). *Continuous Univariate Distributions-2*. New York: Wiley.

[48] Kedem, B. (1980). *Binary Time Series*, New York: Marcel Dekker.

[49] Keenan, D. M. (1982). A Time Series Analysis of Binary Data. *Journal of the American Statistical Association*, 77, 816-821.

[50] Kitagawa, G. (1987). Non-Gaussian State-Space Modelling of Non-stationary Time Series (with comments). *Journal of the American Statistical Association*. 82, 1032-1063.

[51] Knorr-Held, L. (1993). Schtzen von Zustandsmodellen durch Gibbs Sampling. Diplomarbeit, Institute fr Statistik, Universitt Mnchen.

[52] Liang, K. Y., and Zeger, S. (1989). A Class of Logistic Regression Models for Multivariate Binary Time Series. *Journal of the American Statistical Association*. 84, 447-451.

[53] Lindsey, J. K. (1993). *Models for Repeated Measurements*. Oxford: Oxford University Press.

[54] March, J. G. (1972). The Technology of Foolishness. Reprinted in March. J. G., and Olsen, J. (Eds.). *Ambiguity and Choice in Organizations*. 1976. 69-81. Bergen: Universitetsforlaget.

[55] McCulloch, R., and Rossi, P. E. (1991). An Exact Likelihood Analysis of the Multinomial Probit Model. *Journal of Econometrics*, 64, 207-240.

[56] McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P. (Ed.). *Frontiers in Econometrics*, 105-142. New York: Academic Press.

[57] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller. A. H.. and Teller. E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087-1092.

[58] Mudholkar, G. S., and George, E. O. (1978). A Remark on the Shape of the Logistic Distribution. *Biometrika*, 65, 667-668.

[59] Müller, P. (1991). A Generic Approach to Posterior Integration and Gibbs Sampling. Technical Report 91-09, Department of Statistics. Purdue University.

[60] Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

[61] Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

[62] Roberts, G. O., Gelman, A., and Gilks, W. R. (1994). Weak Convergence and central Optimal Scaling of Random Walk Metropolis Algorithms. Technical Report, University of Cambridge.

[63] Roberts, G. O., and Smith, A. F. M. (1994). Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms. *Stochastic Processes and Their Applications,* 49, 207-216.

[64] Roberts, G. O., and Tweedie, R. L. (1994). Geometric Convergence and central Limit Theorems for Multidimensional Hastings-Metropolis Algorithms. Technical Report 94-09, Department of Statistics, Colorado state University.

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

[66] Schnatter, S. (1992). Integration-Based Kalman-Filtering for a Dynamic Generalized Linear Models. *Computational Statistics and Data Analysis,* 13, 447-459.

[67] Smith, A. F. M., and Roberts, G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B,* 55, 3-23.

[68] Tanizaki, H. (1993). Nonlinear Filters. *Lecture Notes in Economics and Mathematical Systems,* No. 400. New York: Springer.

[69] Tiao, G. C., and Box, G. E. P. (1981). Modeling Multiple Time Series with Applications. *Journal of the American Statistical Association,* 76, 802-816.

[70] Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics,* 22, 4, 1701-1762.

[71] Watson, G. S. (1961). Goodness-of-fit Tests on a Circle. *Biometrika,* 48, 109-114.

[72] West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic Generalised Linear Models and Bayesian Forecasting (with discussion). *Journal of the American Statistical Association,* 80, 73-97.

[73] West, M., and Harrison, P. J. (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer.

[74] Van de Ven, A. H., and Associates (1988). Progress Report on the Minnesota Innovation Research Program. Discussion Paper, Minneapolis: University of Minnesota Strategic Management Research Center.

[75] Van de Ven, A. H., and Garud, R. (1992). The Co-Evolution of Technological and Institutional Innovations. Discussion Paper, Minneapolis: University of Minnesota Strategic Management Research Center.

[76] Van de Ven, A. H., and Polley, D. (1992). Learning While Innovating. *Organization Science,* 3, 1, 92-116.

[77] Zeger, S. L., and Karim, M. R. (1991). Generalized Linear Models with Random Effects: A Gibbs Sampling Approach. *Journal of the American Statistical Association,* 86, 79-86.