DISSERTATION

MODELING TIME SERIES OF COUNT DATA

Submitted by

Ying Wang

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Ph. D.

Colorado State University

Fort Collins, Colorado

Fall, 2002

COLORADO STATE UNIVERSITY

October 4, 2002

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY YING WANG ENTITLED " MODELING TIME SERIES OF COUNT DATA " BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF PH. D. .

Committee on Graduate Work

Duane C Boes

Charles W Anderson

Peter J. Brockwell

Richard A. Davis

Adviser

Richard A. Davis

Department Head

ABSTRACT OF DISSERTATION

MODELING TIME SERIES OF COUNT DATA

The focus of this thesis is on modeling time series of count data. We consider an extension of linear Gaussian state space models - parameter driven models in which the mean function of a time series of observed counts $\{Y_t\}$ is specified by a linear predictor modified by a 'latent process'. As in linear regression with correlated errors, there is a need for model diagnostic and identification techniques to decide if it is necessary to include a latent process in the specification of the mean of the Poisson counts and, if so, is there any evidence of autocorrelation in such a process.

For a parameter driven model, the asymptotic distribution of standard generalized linear model estimators is derived for the case that an autocorrelated strong mixing latent process is present. Simple formulas for the effect of the autocovariance of the latent process on standard errors of the regression coefficients are also provided. A method of testing for the existence of a latent process is developed and compared to existing test statistics via simulation. Once the existence of a latent process has been detected, a simple and easily implementable method for estimating the autocovariance of the latent process is given. The standard errors of the estimates are also provided. Methods for adjusting for severe bias in previously proposed estimators of autocovariance are derived and their behavior investigated. A test statistic for testing serial dependence in the latent process is proposed based on the study of the distribution of autocorrelation estimates.

The performance of different test statistics for testing serial dependence have been compared.

Parameter estimation of a parameter driven model is complicated since the latent process is unobservable and the likelihood of the observed data $\boldsymbol{y}$ is an n-fold integral which does not have a simple closed form. Existing estimation methods involve intensive Monte Carlo simulation. A new estimation method that avoids Monte Carlo simulation is developed using an approximation to the likelihood of $\boldsymbol{y}$. Applications of the methods to time series of monthly polio counts in the U.S. is used to illustrate the methods and results. A simulation study has been conducted to compare the performance of the various estimation methods.

Ying Wang
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall, 2002

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my adviser, Dr. Richard Davis, for his guidance, encouragement, support, and patience during the course of this investigation.

Special thanks are given to Dr. William Dunsmuir for his kind help and advice.

I also wish to thank my husband, Jincheng, for being very understanding and supportive.

CONTENTS

1. INTRODUCTION

The focus of this thesis is on modeling time series of count data. Such series are non-negative integer valued and often arise as the number of events occurring in non-overlapping time intervals. We can easily find time series of count data in epidemic studies and analysis of economic behavior, e.g., number of rare disease infections in a given month, number of mortalities in a given day, daily number of discrete price changes on a stock. There are a variety of models and methods that have been developed to analyze count data.

Classical linear models are regression models with the assumption that the observations are normally distributed with a covariance matrix that is known up to a scale constant. Generalized linear models (Nelder and Wedderburn, 1972) extend the classical linear model by allowing for non-normal noise such as the case when the response variable is binary or integer valued. If the response variables are a time series, it is unlikely that neighboring observations are independent and hence serial dependence should be incorporated in the model.

Cox (1981) suggested two classes of models of time-dependent data: parameter driven, and observation driven models. In parameter driven models, dependence is introduced through an unobserved latent process, while in an observation driven model, the conditional distribution of observation is specified as a function of past observations. Both model specifications are becoming increasingly popular because of their ability to handle serial correlation and overdispersion in the data. In this thesis, one type of parameter driven model that combines generalized linear models and time series models is studied in detail.

In this chapter, definition and properties of generalized linear models and generalized state-space models will be introduced and some existing research results on parameter driven models will be reviewed.

1.1 Generalized Linear Models

In a classical linear model, we assume a vector of observations $\boldsymbol{y}$ is a realization of a random vector $\boldsymbol{Y}$ that has a multivariate normal distribution with mean $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and covariance matrix $\sigma^2 I$. The systematic component of the model is the linear predictor $\boldsymbol{\eta}$ given by $\boldsymbol{\eta} = X\boldsymbol{\beta}$, where $X = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ is the design matrix, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_r)^T$ is the vector of parameters. The structure of the systematic part assumes that we know the design matrix that influences the mean and can measure it effectively without error. The link function $g(\cdot)$ between the mean $\boldsymbol{\mu}$ of the random component $\boldsymbol{Y}$ and systematic component $\boldsymbol{\eta}$, defined as $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, is the identity function, i.e., $\boldsymbol{\eta} = \boldsymbol{\mu}$. For this model, least squares can be used to estimate the parameters.

Generalized linear models (GLM) are an extension of the classical linear model. They allow two generalizations from ordinary linear models. First, the distribution of $\boldsymbol{Y}$ may come from an exponential family other than the normal distribution, and second, the link function $g(\cdot)$ between the random and systematic components may be any monotonic differentiable function. We assume that each component of $\boldsymbol{Y}$ has a distribution in the exponential family; its density takes the form

$$f_Y(y; \theta, \phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\} \tag{1.1}$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. If ϕ is known, this is an exponential family model with canonical parameter θ. It may or may not be a two-parameter

exponential family if ϕ is unknown. It is easy to show that

$$\mu := E(Y) = b'(\theta) \quad \text{and} \quad \text{Var}(Y) = b''(\theta)a(\phi).$$

where $b'(\cdot)$ and $b''(\cdot)$ denote the first and second derivatives of $b(\cdot)$ function, respectively.

The function $g(\cdot)$ is called a canonical link if $\eta = g(\mu) = \theta$. Suppose the data $Y_1, \cdots, Y_n$ represent counts which are assumed to be independent and Poisson distributed, that is, $Y_i \sim \text{Poisson}(\mu_i)$, then the log-likelihood of $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ is given by

$$l(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{y}) = \boldsymbol{y}^T \log \boldsymbol{\mu} - \mathbf{1}^T \boldsymbol{\mu} - \mathbf{1}^T \log(\boldsymbol{y}!).$$

Here $\boldsymbol{\theta} = \log \boldsymbol{\mu}$, $b(\boldsymbol{\theta}) = \boldsymbol{\mu}$, $a(\boldsymbol{\phi}) = 1$, and $c(\boldsymbol{y}, \boldsymbol{\phi}) = -\log(\boldsymbol{y}!)$, thus $\boldsymbol{\mu} = \text{E}(\boldsymbol{Y}) = e^{\log \boldsymbol{\mu}} = e^{\boldsymbol{\theta}}$. The canonical link for Poisson data is given by

$$\boldsymbol{\theta} = \log \boldsymbol{\mu} = \boldsymbol{\eta}.$$

For generalized linear models, the inferences can be drawn based on the likelihood function. Maximum likelihood estimates (MLE) of the regression parameters can often be found using iterative weighted least squares. Consistency and asymptotic normality of the MLE for the models with canonical link have been established by Fahrmeir and Kaufmann (1985). More details about fitting generalized linear models can be found in McCullagh and Nelder (1989). Sometimes there is insufficient information to construct a likelihood function. To avoid the complete specification of the underlying distribution, Wedderburn (1974) defined a quasi-likelihood function based on a given relation between the mean and variance of the observations, possibly with an unknown constant of proportionality. The quasi-likelihood function can be used for estimation in the same way as a likelihood function. Suppose that the components of the response vector $\boldsymbol{Y}$ are independent with mean vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^T$ and covariance matrix $\sigma^2 V(\boldsymbol{\mu})$,

where σ^2 may be unknown and $V(\boldsymbol{\mu}) = \text{diag}\{v_1(\boldsymbol{\mu}), \ldots, v_n(\boldsymbol{\mu})\}$ is a diagonal matrix of known functions. The quasi-likelihood for the data is defined as

$$Q(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 v_i(t)} dt. \tag{1.2}$$

Wedderburn (1974) showed that for a one-parameter exponential family the quasi-likelihood is the same as the log-likelihood. The quasi-likelihood is maximized to find the estimates for the model parameters. The asymptotic results of the quasi-likelihood estimators for independent data were given by McCullagh (1983).

1.2 Generalized State-Space Models

State-space models play an important role in time series analysis. A state-space model consists of two equations: the observation equation which specifies the conditional distribution of the observations given the state, and the state equation which specifies the distribution over time of the state. A linear Gaussian state-space model, assuming that both observation and state variables are univariate, can be written as

$$Y_t = G_t S_t + W_t, \quad W_t \sim N(0, H_t), \tag{1.3}$$

$$S_t = F_t S_{t-1} + V_t, \quad V_t \sim N(0, Q_t), \tag{1.4}$$

where Y_t is the observed random variable at time t, and S_t is the state. Equation (1.3) is the observation equation while (1.4) is the state equation. The values of H_t, Q_t, G_t and F_t may depend on an unknown parameter vector $\boldsymbol{\psi}$. Inference about the parameters $\boldsymbol{\psi}$ can then be made based on the likelihood function which can be calculated using the Kalman filter.

Brockwell and Davis (1996) describe the extensions of linear Gaussian state-space models and categorize these extensions as either parameter driven or observation driven. These generalized state-space models are developed for application

in non-normal time series and regression problems. Let $\boldsymbol{Y}^{(t)} = (Y_1, \cdots, Y_t)^T$ with analogous notation for the other variables. Assume that both Y_t and the state variable S_t are univariate. In both parameter driven and observation driven models, we assume that Y_t given $(S_t, \boldsymbol{S}^{(t-1)}, \boldsymbol{Y}^{(t-1)})$ is independent of $(\boldsymbol{S}^{(t-1)}, \boldsymbol{Y}^{(t-1)})$, so the observation equation (1.3) is replaced by a conditional density

$$p(y_t|s_t, \boldsymbol{s}^{(t-1)}, \boldsymbol{y}^{(t-1)}) = p(y_t|s_t). \quad t = 1, 2, \ldots. \tag{1.5}$$

This observation equation is the same for both parameter and observation driven models.

For the parameter driven model, we assume that S_{t+1} given $(S_t, \boldsymbol{S}^{(t-1)}, \boldsymbol{Y}^{(t)})$ is independent of $(\boldsymbol{S}^{(t-1)}, \boldsymbol{Y}^{(t)})$, i.e., the conditional density

$$p(s_{t+1}|s_t, \boldsymbol{s}^{(t-1)}, \boldsymbol{y}^{(t)}) = p(s_{t+1}|s_t). \quad t = 1, 2, \ldots. \tag{1.6}$$

specifies the state equation. Based on (1.5) and (1.6), it can be shown that

$$p(y_1, \ldots, y_n|s_1, \ldots, s_n) = \prod_{j=1}^{n} p(y_j|s_j). \tag{1.7}$$

and hence $Y_1, \ldots, Y_n$ are conditionally independent given the state variables $S_1, \ldots, S_n$. The sequence of state variables $\{S_t\}$ is often referred to as the latent process associated with the observed process.

In an observation driven model, the state equation is specified by the conditional density

$$p(s_{t+1}|\boldsymbol{y}^{(t)}) = p_{S_{t+1}|\boldsymbol{Y}^{(t)}}(s_{t+1}|\boldsymbol{y}^{(t)}). \quad t = 0, 1, \ldots. \tag{1.8}$$

for some prespecified initial density $p(s_1|y^{(0)}) := p_1(s_1)$.

For generalized state-space models, forecasting of future values of the observations is an important problem. Using the observation equation (1.5) and Bayes's Theorem, we have the filtering density

$$\begin{aligned} p(s_t|\boldsymbol{y}^{(t)}) &= \frac{p(s_t, \boldsymbol{y}^{(t)})}{p(\boldsymbol{y}^{(t-1)})} \frac{p(\boldsymbol{y}^{(t-1)})}{p(\boldsymbol{y}^{(t)})} \\ &= \frac{p(s_t, \boldsymbol{y}^{(t-1)}, y_t)}{p(s_t, \boldsymbol{y}^{(t-1)})} \frac{p(s_t, \boldsymbol{y}^{(t-1)})}{p(\boldsymbol{y}^{(t-1)})} / p(y_t|\boldsymbol{y}^{(t-1)}) \\ &= p(y_t|s_t) p(s_t|\boldsymbol{y}^{(t-1)}) / p(y_t|\boldsymbol{y}^{(t-1)}) \end{aligned} \tag{1.9}$$

and one-step-ahead prediction density

$$\begin{aligned} p(s_{t+1}|\boldsymbol{y}^{(t)}) &= \int p(s_{t+1}, s_t|\boldsymbol{y}^{(t)})ds_t \\ &= \int p(s_t|\boldsymbol{y}^{(t)})p(s_{t+1}|s_t)ds_t. \end{aligned} \tag{1.10}$$

Since $\int p(s_t|\boldsymbol{y}^{(t)})ds_t = 1$, from (1.9) we obtain the forecast density function

$$p(y_{t+1}|\boldsymbol{y}^{(t)}) = \int p(y_{t+1}|s_{t+1})p(s_{t+1}|\boldsymbol{y}^{(t)})ds_{t+1}. \tag{1.11}$$

With the setup of the observation driven models, obtaining $p(y_{t+1}|\boldsymbol{y}^{(t)})$ and the calculation of the joint density function $p(y_1, \cdots, y_n) = \prod_{t=1}^{n} p(y_t|\boldsymbol{y}^{(t-1)})$ are straightforward. Forecasting and estimation for the model are easy to carry out. On the other hand, the stochastic mechanism governing the transition of S_{t-1} to S_t is defined implicitly. This makes it difficult to establish stability properties, such as stationarity and ergodicity.

For parameter driven models, the forecast function $p(y_{t+1}|\boldsymbol{y}^{(t)})$ can be obtained by recursively updating the densities $p(s_t|\boldsymbol{y}^{(t)})$ and $p(s_{t+1}|\boldsymbol{y}^{(t)})$ using (1.9) and (1.10). The fact that the sequence $\{S_t\}$ is unobservable makes forecasting and estimation difficult. Since the dependence structure of $\{Y_t\}$ is inherited from that of the state process $\{S_t\}$, the assumption of a stationary $\{S_t\}$ will usually ensure the stationarity of the $\{Y_t\}$ process for linear state-space models.

1.3 Literature Review

Recently, much research effort has been devoted to the theoretical development and refinement of parameter driven models. We give a brief review of modeling time series of count data in the perspective of model fitting, hypothesis testing, estimation and applications.

In parameter driven models, a latent process is generating the serial correlation of the observations, which can be viewed as representing the unmeasured effects on the observations. For Poisson count data such models have been considered by Zeger (1988), Brannas and Johansson (1994) and Jorgensen et al. (1995).

The main assumption in Zeger (1988) and Jorgensen et al. (1995) is that the counts (observed data) are conditionally independent given the latent process. Zeger (1988) assumes that the latent process is stationary and known up to the first two moments. Model parameter estimation in Zeger (1988) is based on a quasi-likelihood estimating equation, so it was not necessary to fully specify the distribution of the latent process. In Jorgensen et al. (1995), the latent process is fully known as a non-stationary gamma Markov process. Their long-term covariates enter the model via the latent process, and they emphasized the need for checking both the observed and unobserved parts of the model by means of residual analysis, which might be difficult for partially specified models. Burnett et al (1991) introduced a general regression model for correlated count data to incorporate Zeger's (1988) model and some longitudinal count data models as special cases.

Zeger (1988) established the asymptotic normality of the quasi-likelihood estimator for the parameter driven model in which the observations follow a log linear model and the latent process is stationary. Blais et al. (2000) extended Zeger's asymptotic normality results to a general exponential family data and a stationary strong mixing latent process. Gourieroux et al. (1984a, 1984b) showed strong consistency and asymptotic normality of the MLE ignoring the presence of an IID latent process. Brannas and Johansson (1994) stated that the Poisson maximum likelihood estimator is consistent even if the serial correlation is not accounted for, but the conventional covariance matrix of the MLE is inconsistent.

Model diagnostic and identification techniques are critical for deciding if it is necessary to include a latent process in the specification of the mean of the observed counts. Brannas and Johansson (1994) reviewed a statistic S which was derived by several authors and used as a test for the existence of a latent process. Dean and Lawless (1989) introduced another test statistic S_a to improve on the

small sample performance of the test. A Monte Carlo study revealed that the S_a statistic has better size properties in small samples (Dean and Lawless, 1989). Brannas and Johansson (1994) compared power functions of the Box-Pierce and the Ljung-Box test statistics for three types of residuals: Pearson, Anscombe, and one of their own. These tests are for testing the hypothesis of white noise or no serial correlation in the latent process. They concluded that the test sizes are significantly higher than the nominal levels, more so for the Ljung-Box statistic than for the Box-Pierce statistic.

There have been many applications of parameter driven models for analyzing count data. Zeger (1988) modeled polio counts in the United States; Campbell (1994) investigated the relationship between sudden infant death syndrome (SIDS) and environmental temperature; Jorgensen et al. (1996) studied the relationship between respiratory morbidity and air pollution; Brannas and Johansson (1994) modeled road accident counts as a function of some meteorological variables such as snow depth, etc.

Several methods have been developed for the estimation of the model parameters. Quasi-likelihood procedures are often used for the case when the latent process in the model is not explicitly specified. Zeger (1988) and Burnett et al. (1991) are two examples using quasi-likelihood. Zeger (1988) uses estimating equations to estimate the regression parameters and method of moments to estimate the covariance parameters.

The likelihood-based estimation method requires a model assumption about the distribution of the state process $\{S_t\}$ (latent process). Since $\{S_t\}$ is unobservable, the likelihood of $\boldsymbol{Y}^{(t)}$ cannot be written in a closed form. To calculate the likelihood numerically, the values of the state S_t are needed. For the models with observations from the exponential family, Fahrmeir (1992) estimated $\boldsymbol{S}^{(t)}$ by the mode of the posterior density $p(\boldsymbol{s}^{(t)}|\boldsymbol{y}^{(t)})$ which is proportional to the joint

density of $\boldsymbol{S}^{(t)}$ and $\boldsymbol{Y}^{(t)}$. Chan and Ledolter (1995) proposed Monte Carlo EM (MCEM) algorithm that converges at a linear rate. As a viable alternative to the MCEM algorithm, the Monte Carlo implementation of the Newton-Raphson algorithm was suggested by Kuk and Cheng (1997). It is computationally faster than the MCEM algorithm as it converges at a faster quadratic rate. Both algorithms require simulation from $\boldsymbol{S}$ given $\boldsymbol{Y}$ with the aid of methods like Gibbs sampling and rejective sampling.

Durbin and Koopman (1997) developed Monte Carlo approximations to the log-likelihood of the observed data $\boldsymbol{Y}$ for the models with conditional density $p(\boldsymbol{y}|\boldsymbol{s})$ and a linear Gaussian state transition equation. The log-likelihood is then maximized numerically. This is a "many samples" method according to the definition of Geyer (1996). Inspired by Geyer's (1994 and 1996) "one sample" Monte Carlo method, Kuk (1997) proposed a method using relative likelihood and extended it to a ratio of two different likelihood functions. All these estimation methods involve simulation from the distribution $p(\boldsymbol{s}|\boldsymbol{y})$ which complicates the algorithms. To date, the relative performance of these estimation methods has not been evaluated and this is one of the main goals of this thesis.

1.4 Outline of Thesis

There has been considerable effort in recent years devoted to the development of methods to fit efficiently all the parameters in a parameter driven model. However, existing techniques rely on the specification of a suitable model for the correlation structure in the latent process. As in linear regression with correlated errors, there is a need for model diagnostic and identification techniques to decide if it is necessary to include a latent process in the specification of the mean of the Poisson counts, and if so, to see if there is any evidence of autocorrelation in

such a process. We follow the approaches that have proved successful in linear regression in the next chapter.

In chapter 2, we show the consistency and asymptotic normality of the Poisson maximum likelihood estimator of the regression parameters in a parameter driven model without considering serial correlation. The proof is given when the latent process is strongly mixing. A method of testing for the existence of a latent process is developed and compared to existing test statistics via simulation. Once the existence of a latent process has been detected, a simple and easily implementable method for estimating the autocovariances of the latent process is given. The standard errors of the estimates are also provided. The performance of various estimators of autocovariances is compared in a simulation study. A test statistic for testing serial dependence in the latent process is proposed based on the study of the distribution of autocorrelation estimates. These test statistics are also compared via simulation.

Chapter 3 reviews two existing likelihood-based estimation methods for the parameter driven models. A new estimation method is developed based on the approximation to the likelihood. Connections between our approximate likelihood and the importance density from one of the existing estimation methods are derived. Parameter estimation results of three methods on a real data set are given. Our estimation method is also compared with the other two via simulations.

Chapter 4 summarizes methods, approaches and results of previous chapters. Limitations of the work and open questions are also provided. Finally, the appendix contains consistency and asymptotic normality results for special cases of Poisson regression models.

2. PARAMETER DRIVEN MODELS

The Poisson distribution has a characteristic property that the expected value and the variance are equal, and the Poisson regression model is a basic member in the class of count data models. However, the variance often exceeds the mean (overdispersion) in the count data in Poisson model framework and sometimes, serial correlation is present in the count data. In this chapter, we consider a class of parameter driven models whose "observation equation" is governed by a Poisson distribution. Overdispersion and serial correlation will be incorporated in this class of models. Denote the time series of counts by $Y_1, \ldots, Y_n$ and suppose that for each t, $\boldsymbol{x}_{nt}$ is a r-vector of observed regressors whose first component is 1. In some cases $\boldsymbol{x}_{nt}$ may depend on the sample size n and form a triangular array. We use $\{\epsilon_t\}$ to denote the latent process. In this setting, the state-variable S_t described in Section 1.2 can be set to either the multivariate vector consisting of $(\boldsymbol{x}_{nt}^T, \epsilon_t)^T$ or just ϵ_t.

The conditional distribution of Y_t given $\boldsymbol{x}_{nt}$ and ϵ_t corresponding to the observation equation of (1.5) is assumed to be Poisson with mean $u_t = \epsilon_t \exp\{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}\}$ denoted by

$$Y_t | \epsilon_t, \boldsymbol{x}_{nt} \sim \text{Poisson}(\epsilon_t \exp\{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}\}). \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^T$ is a vector of regression coefficients. The analogue of the state equation in this context is a specification of the distributional properties of the latent process $\{\epsilon_t\}$. We assume that x_{nt} is fixed and $\{\epsilon_t\}$ is a non-negative strictly stationary time series with mean 1 and autocovariance function (ACVF)

$$\gamma_\epsilon(h) = E[(\epsilon_{t+h} - 1)(\epsilon_t - 1)]. \tag{2.2}$$

The assumption of non-negativity of ϵ_t is clear in order to ensure that the conditional mean of Y_t is non-negative. The condition that $E(\epsilon_t) = 1$ is imposed for identifiability reasons; otherwise, if $c = E(\epsilon_t) \neq 1$, then c can be absorbed into the intercept term in the exponent of u_t. (That is, one would replace ϵ_t with ϵ_t/c and β_1 with $\beta_1 + \log c$.)

To meet the non-negativity constraint on ϵ_t, it is often convenient to model the logarithms of ϵ_t. Letting $\alpha_t = \log \epsilon_t$, i.e., $\epsilon_t = \exp(\alpha_t)$, then the conditional mean of Y_t can be written as

$$u_t = \exp\{\boldsymbol{x}_{nt}^T \boldsymbol{\beta} + \alpha_t\}.$$

In order for the corresponding ϵ_t to have mean 1, we must assume $E[\exp(\alpha_t)] = 1$. There is not an explicit relationship between the ACVF's of $\{\epsilon_t\}$ and $\{\alpha_t\}$ unless $\{\alpha_t\}$ is a stationary Gaussian process. Suppose that $\{\epsilon_t\}$ is a stationary log-normal process, i.e., $\{\alpha_t\}$ is a stationary Gaussian process with ACVF $\gamma_\alpha(\cdot)$. To satisfy the identifiability requirement that $E(\epsilon_t) = E[\exp(\alpha_t)] = 1$, it is required that $\alpha_t \sim N(-\sigma_\alpha^2/2, \sigma_\alpha^2)$, where σ_α^2 is defined as $\gamma_\alpha(0)$, the variance of $\{\alpha_t\}$ process. So the mean of $\{\alpha_t\}$ is -0.5 times its variance. With this choice of mean and variance in the log-normal distribution, there is a nice connection between the ACVF's, $\gamma_\epsilon(h) = E[\exp(\alpha_{t+h} + \alpha_t)] - 1 = e^{\gamma_\alpha(h)} - 1$ for all h.

Figure 2.1 displays a sample path of a Poisson time series generated from model (2.1) with $\boldsymbol{x}_{nt}^T \boldsymbol{\beta} = (1, t/n)(0.1, 1)^T$, $n = 200$, $\epsilon_t = e^{\alpha_t}$ and $\alpha_t = 0.5\alpha_{t-1} + z_t$, where $z_t \sim \text{IID}N(-1/4, 3/4)$.

2.1 Means, Variances, and Autocovariances of Y_t

In this section various key facts about the moments of the observed count process Y_t are derived and relationships between the first and second moments of the latent process ϵ_t are provided. Throughout, expectations, variances and

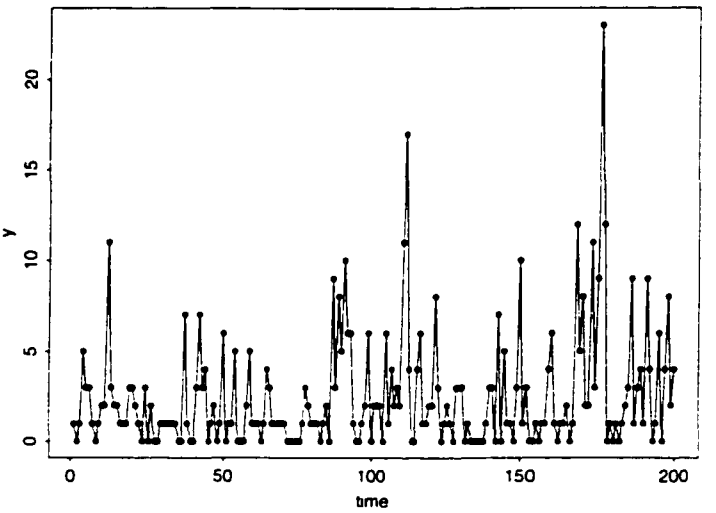

Figure 2.1: Sample path of a parameter driven model.

covariances are conditional upon the regressors $\boldsymbol{x}_{nt}$ (and this will not be explicitly noted), but not on the latent process unless otherwise indicated in the usual way. Here, $\sigma_\epsilon^2 := \mathrm{Var}(\epsilon_t)$.

Mean of Y_t:

$$\mu_t := E(Y_t) = E[E(Y_t|\epsilon_t)] = \exp(\boldsymbol{x}_{nt}^T\boldsymbol{\beta}).$$

Variance of Y_t:

$$\mathrm{Var}(Y_t) = E[\mathrm{Var}(Y_t|\epsilon_t)] + \mathrm{Var}[E(Y_t|\epsilon_t)] = \mu_t + \sigma_\epsilon^2\mu_t^2.$$

Autocovariance function of Y_t:

$$\mathrm{Cov}(Y_{t+h}, Y_t) = \mu_t\mu_{t+h}[E(\epsilon_t\epsilon_{t+h}) - 1] = \mu_t\mu_{t+h}\gamma_\epsilon(h).$$

Autocorrelation function (ACF) of Y_t:

$$\begin{aligned}\mathrm{Corr}(Y_s, Y_t) &= \frac{\mu_s\mu_t\gamma_\epsilon(s-t)}{\sqrt{(\mu_s + \sigma_\epsilon^2\mu_s^2)(\mu_t + \sigma_\epsilon^2\mu_t^2)}} \\ &= \frac{\rho_\epsilon(s-t)}{\sqrt{[1 + (\sigma_\epsilon^2\mu_s)^{-1}][1 + (\sigma_\epsilon^2\mu_t)^{-1}]}}.\end{aligned}$$

where $\rho_\epsilon(h) := \mathrm{Corr}(\epsilon_t, \epsilon_{t+h})$. The ACF of Y_t is not free of the regressors $\boldsymbol{x}_{nt}$, as is to be expected. In the case when there are no regressor terms other than a constant, i.e., $r = 1$, the process $\{Y_t\}$ is then stationary with the ACF given by

$$\rho_Y(h) := \mathrm{Corr}(Y_t, Y_{t+h}) = \frac{\gamma_\epsilon(h)}{\mu^{-1} + \sigma_\epsilon^2},$$

where $\mu = e^{\beta_1}$. Since $\mu > 0$ we see that

$$|\rho_Y(h)| \leq |\rho_\epsilon(h)|.$$

Consider Figure 2.2 in which the ACF for the $\{Y_t\}$ process is shown along with

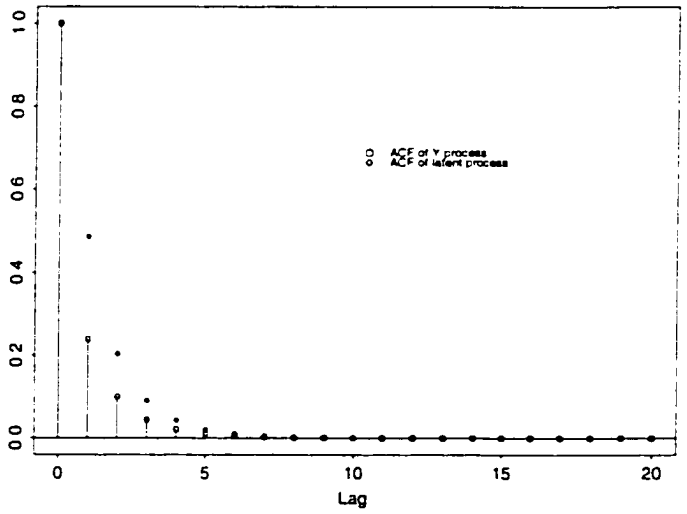

Figure 2.2: Autocorrelation functions of the $\{Y_t\}$ and latent processes

that for the latent process. Clearly, even when the mean of Y_t is stationary, the ACF of the observed count process tends to underestimate that of the latent process. This illustrates the difficulty in detecting dependence within the latent process. Little or no correlation in the Y_t may mask significant correlation in the latent process and demonstrates the futility of developing identification procedures based solely on second-order properties of the data. Because of this, methods are

required to estimate the underlying correlation and to test whether it is zero. Such methods also need to be applicable when the regression terms are present.

2.2 Asymptotic Properties of the GLM Estimates

In a linear model with time series errors, e.g.,

$$Y_t = \boldsymbol{x}_t^T \boldsymbol{\beta} + W_t,$$

the first step in fitting such models is to determine the autocovariance structure of the time series of errors $\{W_t\}$. Assuming that $\{W_t\}$ is a linear process such as an ARMA, the parameter $\boldsymbol{\beta}$ is estimated using ordinary least squares (OLS) by regressing the data vector $(Y_1, \ldots, Y_n)^T$ onto the $\boldsymbol{x}_t$. While this estimate ignores the dependence structure of the $\{W_t\}$, the OLS estimate has the same asymptotic efficiency as the MLE of $\boldsymbol{\beta}$ under a wide class of models for the W_t process (Hannan, 1970). The asymptotic covariance matrix of the OLS (and MLE) estimate does depend on the covariance structure of the W_t. Once a consistent estimator of $\boldsymbol{\beta}$ has been found, then the ACVF of the W_t can be consistently estimated from the sample ACVF of the residuals defined by, $W_t = Y_t - \boldsymbol{x}_t^T \hat{\boldsymbol{\beta}}_{OLS}$. A model is then selected for $\{W_t\}$ and the regression parameter $\boldsymbol{\beta}$ and the parameters of the model for the W_t can be re-estimated using MLE.

We consider carrying out an analogous procedure applied to the parameter driven models. The first step is to estimate $\boldsymbol{\beta}$ using GLM or Poisson regression.

2.2.1 Consistency and Asymptotic Normality of the GLM Estimates

The GLM estimate of the parameters in a parameter driven model are obtained using Poisson regression. The presence of a latent process in the model is ignored. The Poisson model here is a misspecified model (White, 1982) for the parameter driven model. If one does not assume that the probability model is

correctly specified, it is natural to ask what happens to the properties of the maximum likelihood estimators (MLE). We will show below that the GLM estimates of the parameters ignoring the presence of a latent process in the parameter driven models are still consistent and asymptotically normal.

Let $Y_1, \cdots, Y_n$ be observations from model (2.1) with true value $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The GLM estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood

$$l(\boldsymbol{\beta}) = -\sum_{t=1}^{n} e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}} + \sum_{t=1}^{n} Y_t \boldsymbol{x}_{nt}^T \boldsymbol{\beta} - \log[\prod_{t=1}^{n} (Y_t!)] \tag{2.3}$$

which ignores the latent process in the model. In order to derive the asymptotic behavior of $\hat{\boldsymbol{\beta}}$, we need to make some assumptions about the regressors $\boldsymbol{x}_{nt}$. Otherwise, the usual asymptotics may not apply. For example if a linear time trend is included in the model, then it is necessary to divide time by the sample size n, e.g., $\boldsymbol{x}_{nt} = (1, t/n)^T$. Without dividing by n, the trend coefficient cannot be estimated consistently for negative values of the coefficient (the Poisson mean will eventually become arbitrarily close to zero). See Appendices A and B for details.

Assume there exists a sequence of nonsingular matrices M_n such that the regressors obey the following conditions:

$$M_n^T \left(\sum_{t=1}^{n} \boldsymbol{x}_{nt} \boldsymbol{x}_{nt}^T e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} \right) M_n \to \Omega_l(\boldsymbol{\beta}_0). \tag{2.4}$$

$$M_n^T \left(\sum_{t=1}^{n} \boldsymbol{x}_{nt} \boldsymbol{x}_{n,t+h}^T e^{(\boldsymbol{x}_{nt}^T + \boldsymbol{x}_{n,t+h}^T)\boldsymbol{\beta}_0} \right) M_n \to \Omega_h(\boldsymbol{\beta}_0) \text{ (uniformly in } h). \tag{2.5}$$

$$\sup_{1 \le t \le n} |M_n^T \boldsymbol{x}_{nt}| e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} \to 0. \tag{2.6}$$

and for $h \le 0$,

$$M_n^T \sum_{t=1}^{1-h} \boldsymbol{x}_{nt} \boldsymbol{x}_{n,t+h}^T e^{(\boldsymbol{x}_{nt}^T + \boldsymbol{x}_{n,t+h}^T)\boldsymbol{\beta}_0} M_n \to 0. \tag{2.7}$$

and is uniformly bounded in h as $n \to \infty$. Similarly, for $h > 0$,

$$M_n^T \sum_{t=n-h}^{n} \boldsymbol{x}_{nt} \boldsymbol{x}_{n,t+h}^T e^{(\boldsymbol{x}_{nt}^T + \boldsymbol{x}_{n,t+h}^T)\boldsymbol{\beta}_0} M_n \to 0 \tag{2.8}$$

and is uniformly bounded in h as $n \to \infty$. Define

$$\begin{aligned}
\Omega_n(\boldsymbol{\beta}_0) &:= \mathrm{Cov}[M_n^T \sum_{t=1}^{n} \boldsymbol{x}_{nt}(Y_t - \mu_t)] \\
&= M_n^T \sum_{t=1}^{n} \boldsymbol{x}_{nt}\boldsymbol{x}_{nt}^T \mathrm{Var}(Y_t) M_n + M_n^T \sum_{t,s=1, t\neq s}^{n} \boldsymbol{x}_{nt}\boldsymbol{x}_{ns}^T \mathrm{Cov}(Y_t, Y_s) M_n \\
&= M_n^T \sum_{t=1}^{n} \boldsymbol{x}_{nt}\boldsymbol{x}_{nt}^T (\mu_t + \sigma_\epsilon^2 \mu_t^2) M_n + M_n^T \sum_{t,s=1, t\neq s}^{n} \boldsymbol{x}_{nt}\boldsymbol{x}_{ns}^T \mu_t \mu_s \gamma_\epsilon(s-t) M_n \\
&= M_n^T \sum_{t=1}^{n} \boldsymbol{x}_{nt}\boldsymbol{x}_{nt}^T \mu_t M_n + M_n^T \sum_{t=1}^{n}\sum_{s=1}^{n} \boldsymbol{x}_{nt}\boldsymbol{x}_{ns}^T \mu_t \mu_s \gamma_\epsilon(s-t) M_n.
\end{aligned}$$

Theorem 2.1 Let $\hat{\boldsymbol{\beta}}$ be the GLM estimate of $\boldsymbol{\beta}$ obtained by maximizing $l(\boldsymbol{\beta})$ in (2.3) for the parameter driven model (2.1). Further assume that the $\{\boldsymbol{x}_{nt}\}$ satisfy the above conditions (2.4)-(2.8), and $\sum_{h=0}^{\infty} |\gamma_\epsilon(h)| < \infty$. Then,

$$\Omega_n(\boldsymbol{\beta}_0) \to \Omega_I(\boldsymbol{\beta}_0) + \Omega_{II}(\boldsymbol{\beta}_0) \tag{2.9}$$

and

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0. \tag{2.10}$$

where

$$\Omega_{II}(\boldsymbol{\beta}_0) = \sum_{h=-\infty}^{\infty} \Omega_h(\boldsymbol{\beta}_0)\gamma_\epsilon(h).$$

Moreover, if

$$M_n^T \sum_{t=1}^{n} \boldsymbol{x}_{nt} e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0}(\epsilon_t - 1) \xrightarrow{d} N(\mathbf{0}, \Omega_{II}(\boldsymbol{\beta}_0)). \tag{2.11}$$

then

$$M_n^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \Omega_I^{-1}(\boldsymbol{\beta}_0) + \Omega_I^{-1}(\boldsymbol{\beta}_0)\Omega_{II}(\boldsymbol{\beta}_0)\Omega_I^{-1}(\boldsymbol{\beta}_0)) \tag{2.12}$$

as $n \to \infty$.

Remark 2.1 The matrix, $M_n^T \Omega_I^{-1}(\boldsymbol{\beta}_0) M_n$, is the asymptotic covariance matrix associated with the standard GLM estimate, while $M_n^T \Omega_I^{-1}(\boldsymbol{\beta}_0)\Omega_{II}(\boldsymbol{\beta}_0)\Omega_I^{-1}(\boldsymbol{\beta}_0) M_n$ is the additional contribution to the asymptotic covariance caused by the existence of the latent process. Clearly if the latent process has small covariance then this

second term will not contribute very much. The form of the covariance matrix given above shows explicitly how autocorrelation inflates the true asymptotic covariance matrix.

Remark 2.2 The convergence in (2.11) holds under a variety of conditions on the latent process. For example, if $\{\alpha_t = \log \epsilon_t\}$ is a linear Gaussian process (See Davis, Dunsmuir and Wang (2000) for the proof) or if $\{\epsilon_t\}$ is strongly mixing with a suitable rate (the proof is given in Section 2.2.2) then the central limit theorem in (2.11) holds.

Remark 2.3 A wide range of regression functions satisfy the conditions of Theorem 2.1. These include trend functions, harmonic functions and stationary processes.

2.2.2 Proof of Theorem 2.1 When $\{\epsilon_t\}$ Is Strongly Mixing

We will show that if a latent process $\{\epsilon_t\}$ satisfies the assumptions (A1) - (A3) listed below, then (2.11) is valid. The following definition and proposition are needed for the proof.

Definition 2.1 (Strongly Mixing) A stationary process $\{X_t\}$ is said to be strongly mixing (α-mixing) if

$$\alpha(n) := \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{n}^{\infty}} |P(AB) - P(A)P(B)| \to 0 \text{ as } n \to \infty,$$

where P(.) is the probability, $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{n}^{\infty}$ are the σ-fields generated by $\{X_t, t \leq 0\}$ and $\{X_t, t \geq n\}$, respectively, and $\alpha(n)$ is called mixing coefficient.

Proposition 2.1 (Davidson, 1992) Let $\{X_{nt},\ t = 1, \cdots, n,\ n \geq 1\}$ denote a triangular array of random variables defined on the probability space $(\Omega, \mathcal{F}, P)$ such that

1. $E(X_{nt}) = 0$ and $E(\sum_{t=1}^{n} X_{nt})^2 = 1$;

2. There exists a constant $\gamma > 2$ and a triangular array of positive constants $\{d_{nt}\}$ such that $\sup_n\{n\,(\max_{1\le t\le n} d_{nt})^2\} < \infty$ and $\left(E(\frac{X_{nt}}{d_{nt}})^\gamma\right)^{\frac{1}{\gamma}}$ is uniformly bounded in t and n; and

3. For each n, the sequence $\{X_{nt}\}$ is strongly mixing with mixing coefficient $\alpha(m)$ such that $\sum_m \alpha(m)^{\frac{\gamma}{\gamma-2}} < \infty$.

Then the central limit theorem holds:

$$\sum_{t=1}^{n} X_{nt} \xrightarrow{d} N(0,1).$$

In addition to the conditions (2.4), (2.5), (2.7) and (2.8), we need the following assumptions on the latent process $\{\epsilon_t\}$ and the regressors $\boldsymbol{x}_{nt}$ in model (2.1).

Assumptions

(A1). There exist a positive constants δ such that $E(\epsilon_t - 1)^{\delta+2} < \infty$:

(A2). The latent process $\{\epsilon_t\}$ is strongly mixing with mixing coefficient $\alpha(m)$ such that $\sum_m \alpha(m)^{\frac{\delta+2}{\delta}} < \infty$: and

(A3). There exists a sequence of nonsingular matrices M_n such that $\sup_{1\le t\le n} |M_n^T \boldsymbol{x}_{nt}| e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} = O(\frac{1}{\sqrt{n}})$.

Proof of Theorem 2.1 The GLM estimate of $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood function $l(\boldsymbol{\beta})$ of $(Y_1, \ldots, Y_n)$ given in (2.3). Maximizing $l(\boldsymbol{\beta})$ is equivalent to maximizing

$$l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_0) = -\sum_{t=1}^{n} e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0}[e^{\boldsymbol{x}_{nt}^T M_n \boldsymbol{u}} - 1 - \boldsymbol{x}_{nt}^T M_n \boldsymbol{u}] + \sum_{t=1}^{n}(Y_t - e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0})\boldsymbol{x}_{nt}^T M_n \boldsymbol{u},$$

where

$$\boldsymbol{u} = M_n^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

We can rewrite the above expression as

$$g_n(\boldsymbol{u}) := l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_0) = l(M_n \boldsymbol{u} + \boldsymbol{\beta}_0) - l(\boldsymbol{\beta}_0) = -B_n(\boldsymbol{u}) + A_n(\boldsymbol{u}).$$

where

$$B_n(\boldsymbol{u}) := \sum_{t=1}^{n} e^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}[e^{\boldsymbol{x}_{nt}^T M_n \boldsymbol{u}} - 1 - \boldsymbol{x}_{nt}^T M_n \boldsymbol{u}].$$

and

$$A_n(\boldsymbol{u}) := \sum_{t=1}^{n}[Y_t - e^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}]\boldsymbol{x}_{nt}^T M_n \boldsymbol{u}.$$

Note that $g_n(\boldsymbol{u})$ is a convex function of $\boldsymbol{u}$ and so we can apply a standard result for functional limit theorems.

First note that

$$\begin{aligned} B_n(\boldsymbol{u}) &= \sum_{t=1}^{n} e^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}\frac{(\boldsymbol{x}_{nt}^T M_n \boldsymbol{u})^2}{2} + E_n(\boldsymbol{u}) \\ &= \frac{1}{2}\boldsymbol{u}^T M_n^T[\sum_{t=1}^{n}\boldsymbol{x}_{nt}\boldsymbol{x}_{nt}^T e^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}]M_n\boldsymbol{u} + E_n(\boldsymbol{u}). \end{aligned}$$

where

$$E_n(\boldsymbol{u}) = \sum_{t=1}^{n} e^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}[\sum_{i=3}^{\infty}\frac{1}{i!}(\boldsymbol{x}_{nt}^T M_n \boldsymbol{u})^i].$$

Assumption (A3) assures that $E_n(\boldsymbol{u}) \to 0$. and it follows by assumption (2.4) that

$$B_n(\boldsymbol{u}) \longrightarrow \frac{1}{2}\boldsymbol{u}^T\Omega_I(\boldsymbol{\beta}_0)\boldsymbol{u}$$

for any fixed $\boldsymbol{u}$.

Next consider

$$A_n(\boldsymbol{u}) = \boldsymbol{U}_n^T\boldsymbol{u}.$$

where

$$\boldsymbol{U}_n := \sum_{t=1}^{n}[Y_t - e^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}]M_n^T\boldsymbol{x}_{nt}.$$

We show below that

$$\boldsymbol{U}_n \xrightarrow{d} N(\boldsymbol{0}, \Omega_I(\boldsymbol{\beta}_0) + \Omega_{II}(\boldsymbol{\beta}_0)).$$

From the above, $g_n(\boldsymbol{u}) \to g(\boldsymbol{u}) := -\frac{1}{2}\boldsymbol{u}^T\Omega_I(\boldsymbol{\beta}_0)\boldsymbol{u} + N(\boldsymbol{0}, \Omega_I(\boldsymbol{\beta}_0) + \Omega_{II}(\boldsymbol{\beta}_0))\boldsymbol{u}$ for every $\boldsymbol{u}$ and since $g_n(\cdot)$ is convex, the distribution of $\hat{\boldsymbol{u}}_n$ maximizing $g_n(\boldsymbol{u})$ converges in distribution to $\hat{\boldsymbol{u}} = N(\boldsymbol{0}, \Omega_I^{-1}(\boldsymbol{\beta}_0) + \Omega_I^{-1}(\boldsymbol{\beta}_0)\Omega_{II}(\boldsymbol{\beta}_0)\Omega_I^{-1}(\boldsymbol{\beta}_0))$ which

maximizes $g(\boldsymbol{u})$. See Rockafellar (1970) and Pollard (1991) for the theoretical results required for this proof.

We now show that $\boldsymbol{U}_n \xrightarrow{d} N(\boldsymbol{0}, \Omega = \Omega_I(\boldsymbol{\beta}_0) + \Omega_{II}(\boldsymbol{\beta}_0))$. It suffices to show that

$$E e^{i\boldsymbol{s}^T \boldsymbol{U}_n} \to e^{-\frac{1}{2}\boldsymbol{s}^T \Omega \boldsymbol{s}}$$

for every real vector $\boldsymbol{s}$. Since the characteristic function of a Poisson random variable with mean λ is $e^{(e^{it}-1)\lambda}$, we have, by conditioning on $\{\epsilon_t\}$, that

$$E e^{i\boldsymbol{s}^T \boldsymbol{U}_n} = \exp\{\sum_{t=1}^{n} [(e^{i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}} - 1)\epsilon_t e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} - i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0}]\}.$$

Now,

$$\begin{aligned}
&\sum_{t=1}^{n} \left[\left(e^{i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}} - 1 \right) \epsilon_t e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} - i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} \right] \\
&= \sum_{t=1}^{n} \left(e^{i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}} - 1 - i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} \right) e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} \\
&\quad + \sum_{t=1}^{n} \left(e^{i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}} - 1 \right) e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} (\epsilon_t - 1) \\
&:= D_n + F_n.
\end{aligned}$$

From assumption (A3) we know that $\sup_{1 \le t \le n} |M_n^T \boldsymbol{x}_{nt}| e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} \to 0$. It follows easily that

$$D_n + \frac{1}{2} \sum_{t=1}^{n} (\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt})^2 e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} \to 0,$$

so that

$$D_n \to -\frac{1}{2} \boldsymbol{s}^T \Omega_I(\boldsymbol{\beta}_0) \boldsymbol{s}. \tag{2.13}$$

Turning to F_n,

$$\begin{aligned}
&\left| F_n - \textstyle\sum_{t=1}^{n} i\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} (\epsilon_t - 1) \right| \\
&\le \left| \tfrac{1}{2} \textstyle\sum_{t=1}^{n} \left(\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} \right)^2 e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} (\epsilon_t - 1) \right| \\
&\quad + const \textstyle\sum_{t=1}^{n} \left| \boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} \right|^3 e^{\boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0} |\epsilon_t - 1|.
\end{aligned} \tag{2.14}$$

The variance of the term inside the mod signs converges to 0 as $n \to \infty$, and hence this term converges to its mean 0 in probability. More specifically, we have

$$\begin{aligned}
&\mathrm{Var}\left[\sum_{t=1}^{n}\left(\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right)^2 \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}(\epsilon_t - 1)\right] \\
&= \sum_{t=1}^{n}\sum_{j=1}^{n}\left(\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right)^2 \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}\left(\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nj}\right)^2 \epsilon^{\boldsymbol{x}_{nj}^T\boldsymbol{\beta}_0}\mathrm{Cov}(\epsilon_t, \epsilon_j) \\
&\le \left[\sup_{1\le t\le n}\left|\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right|^2 \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}\right]^2 n\sum_{|h|<n}|\gamma_\epsilon(h)| \\
&\to 0 \quad \text{as} \quad n \to \infty
\end{aligned}$$

by assumptions $\sup_{1\le t\le n}\left[\left|\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right| \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}\right] = O(\frac{1}{\sqrt{n}})$ and $\sum_h |\gamma_\epsilon(h)| < \infty$.

The second term in the right-hand-side of equation (2.14) also converges to 0 in probability since its expectation converges:

$$E\left[\sum_{t=1}^{n}\left|\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right|^3 \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}|\epsilon_t - 1|\right] = \sum_{t=1}^{n}\left|\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right|^3 \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0} E|\epsilon_t - 1| \to 0,$$

because $E|\epsilon_t - 1|$ is bounded by the assumptions $E(\epsilon_t - 1)^{\delta+2} < \infty$ for $\delta > 0$ and $\sum_{t=1}^{n}\left|\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\right|^3 \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0} = O(\frac{1}{n\sqrt{n}})$.

Now if we can show that

$$C_n(\boldsymbol{s}) := \sum_{t=1}^{n} \boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} \epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}(\epsilon_t - 1) \xrightarrow{d} V \sim N(0, \boldsymbol{s}^T\Omega_{II}\boldsymbol{s}). \tag{2.15}$$

then

$$D_n + F_n \xrightarrow{d} -\frac{1}{2}\boldsymbol{s}^T\Omega_I\boldsymbol{s} + iV,$$

so that

$$E\epsilon^{i\boldsymbol{s}^T U_n} \to \epsilon^{-\frac{1}{2}\boldsymbol{s}^T\Omega_I\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T\Omega_{II}\boldsymbol{s}} = \epsilon^{-\frac{1}{2}\boldsymbol{s}^T\Omega\boldsymbol{s}}$$

as asserted.

$$\begin{aligned}
\text{Let} \quad \sigma_n^2(\boldsymbol{s}) &:= \mathrm{Var}(C_n(\boldsymbol{s})) = E\left(\sum_{t=1}^{n}\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}(\epsilon_t - 1)\right)^2 \\
&= \boldsymbol{s}^T M_n^T\left[\sum_{t=1}^{n}\sum_{j=1}^{n}\boldsymbol{x}_{nt}\boldsymbol{x}_{nj}^T\epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}\epsilon^{\boldsymbol{x}_{nj}^T\boldsymbol{\beta}_0}\gamma_\epsilon(t-j)\right] M_n\boldsymbol{s}.
\end{aligned}$$

Note $EC_n = 0$ and $\sigma_n^2(\boldsymbol{s}) \to \boldsymbol{s}^T\Omega_{II}\boldsymbol{s}$. Let $Z_{nt} = \boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt}\epsilon^{\boldsymbol{x}_{nt}^T\boldsymbol{\beta}_0}(\epsilon_t - 1)/\sigma_n(\boldsymbol{s})$, we want to show that $\sum_{t=1}^{n} Z_{nt} \xrightarrow{d} N(0,1)$. We now verify that $\{Z_{nt}\}$ satisfies the assumptions in Proposition 2.1.

It is clear that

$$E(Z_{nt}) = 0 \quad \text{and} \quad E(\sum_{t=1}^{n} Z_{nt})^2 = E\left[\frac{C_n(\boldsymbol{s})}{\sigma_n(\boldsymbol{s})}\right]^2 = 1.$$

Take $d_{nt} = 1/\sqrt{n}$, then $\sup_n\{n(\max_{1\le t\le n} d_{nt})^2\} = 1 < \infty$.

Now, for $\delta > 0$,

$$\begin{aligned}\left[E(\tfrac{Z_{nt}}{d_{nt}})^{\delta+2}\right]^{\frac{1}{\delta+2}} &= \left[E(\tfrac{\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} e \boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0 (\epsilon_t - 1)}{\sigma_n(\boldsymbol{s}) d_{nt}})^{\delta+2}\right]^{\frac{1}{\delta+2}} \\ &= \tfrac{\boldsymbol{s}^T M_n^T \boldsymbol{x}_{nt} e \boldsymbol{x}_{nt}^T \boldsymbol{\beta}_0}{\sigma_n(\boldsymbol{s}) d_{nt}} \left[E(\epsilon_t - 1)^{\delta+2}\right]^{\frac{1}{\delta+2}}\end{aligned}$$

which is bounded uniformly in t and n by assumptions (A1) and (A3). The assumption 3 of Proposition 2.1 is assured by the assumption (A2). Apply Proposition to $\{Z_{nt}\}$, we have $\sum_{t=1}^{n} Z_{nt} \xrightarrow{d} N(0, 1)$. Thus,

$$C_n(\boldsymbol{s}) \xrightarrow{d} N(0, \boldsymbol{s}^T \Omega_{II} \boldsymbol{s}).$$

This finishes the proof.

2.2.3 Application to the Polio Data

The polio data set consists of the monthly number of cases of poliomyelitis in the U.S. for the years 1970–1983 as reported by the Center for Disease Control and Prevention. This data was originally modeled by Zeger (1988) and has become a standard example in the field. The data plotted in Figure 2.3 reveals some seasonality and the possibility of a slight decreasing trend. Detection of the decreasing trend is one of the main objectives in modeling this data.

We apply the results of Theorem 2.1 to the polio data and use the same regression variables as in Zeger (1988) consisting of an intercept term, a linear trend, and harmonics at periods of 6 and 12 months. Specifically,

$$\boldsymbol{x}_t = (1, t'/1000, \cos(2\pi t'/12), \sin(2\pi t'/12), \cos(2\pi t'/6), \sin(2\pi t'/6))^T,$$

where $t' = (t - 73)$ is used to locate the intercept term at January 1976 as in Zeger's analysis.

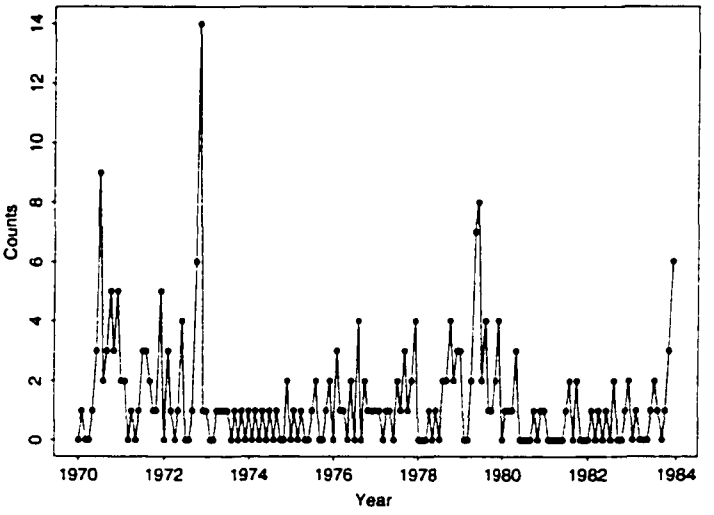

Figure 2.3: Monthly number of cases of poliomyelitis in the U.S. (1970-1983)

Table 2.1 summarizes Zeger's estimates (based on quasi-likelihood estimating equation approach) of $\boldsymbol{\beta}$ and GLM estimates arising from a standard GLM fit. The asymptotic standard errors for the GLM estimates given in the Theorem 2.1 are estimated using the values of $\hat{\sigma}_\epsilon^2 = 0.77$ and $\hat{\rho}_\epsilon(1) = 0.77$ reported in Table 3 of Zeger (1988). These were obtained using the formulas $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{GLM}) = \hat{\Omega}_{I,n}^{-1} + \hat{\Omega}_{I,n}^{-1}\hat{\Omega}_{II,n}\hat{\Omega}_{I,n}^{-1}$ with $\hat{\Omega}_{I,n} = \sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t^T e^{\boldsymbol{x}_t^T \hat{\boldsymbol{\beta}}_{GLM}}$ and $\hat{\Omega}_{II,n} = \sum_{t=1}^{n}\sum_{s=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_s^T e^{(\boldsymbol{x}_t^T + \boldsymbol{x}_s^T)\hat{\boldsymbol{\beta}}_{GLM}} \hat{\gamma}_\epsilon(s-t)$. Use of the correct standard errors for the trend term would lead to the conclusion that the trend is not significant whereas use of the standard errors produced by the GLM analysis would lead to declaring the trend to be significant.

The final two columns of Table 2.1 report the results of 5000 simulations of a time series of length $n = 168$ using the GLM fitted values, $\hat{\boldsymbol{\beta}}_{GLM}$, as the true values. The latent process in this simulation is assumed to be a log-normal autoregression of order 1 with mean $-\sigma_\alpha^2/2 = -0.285$, lag-one correlation $\phi = 0.82$, and variance $\sigma_\alpha^2 = 0.57$. These values of the parameters were chosen in

order to match the second order properties of the fitted $\{\epsilon_t\}$ process of Zeger (1988). The average over the 5000 simulated values of the intercept estimate is observed as 0.145 which is biased downwards from the true value of 0.207 used in the simulations. The other parameters appear to be estimated without substantial bias. The standard deviation of the GLM estimates observed over the 5000 replications are reported in the last column. These are in good agreement with the standard errors obtained from the asymptotic theory.

Table 2.1 Parameter estimates and their standard errors for Polio data.

	Zeger		GLM		Asym	Simulations	
	$\hat{\boldsymbol{\beta}}_Z$	s.e.	$\hat{\boldsymbol{\beta}}_{GLM}$	s.e.	s.e.$(\hat{\boldsymbol{\beta}}_{GLM})$	ave$(\hat{\boldsymbol{\beta}}_{GLM})$	s.d.$(\hat{\boldsymbol{\beta}}_{GLM})$
Intercept	0.17	0.13	0.207	0.075	0.205	0.145	0.214
$t' \times 10^{-3}$	-4.35	2.68	-4.799	1.403	4.115	-4.871	4.261
$\cos(2\pi t'/12)$	-0.11	0.16	-0.149	0.097	0.157	-0.150	0.152
$\sin(2\pi t'/12)$	-0.48	0.17	-0.532	0.109	0.168	-0.536	0.166
$\cos(2\pi t'/6)$	0.20	0.14	0.169	0.098	0.122	0.171	0.123
$\sin(2\pi t'/6)$	-0.41	0.14	-0.432	0.101	0.125	-0.438	0.127

2.3 Testing for the Existence of a Latent Process

Prior to the estimation of autocovariances it is suggested that a test for the existence of a latent process be performed. Brannas and Johansson (1994) reviewed the following statistic

$$S = \frac{\sum_{t=1}^{n}[(Y_t - \hat{\mu}_t)^2 - Y_t]}{[2\sum_{t=1}^{n}\hat{\mu}_t^2]^{1/2}}$$

derived by several authors and based on a local alternative hypothesis or the Lagrange multiplier test of the Poisson distribution against a negative binomial or more general Katz distribution. A variant, introduced by Dean and Lawless (1989) "to improve on the small sample performance of the test", also considered by them is

$$S_a = \frac{\sum_{t=1}^{n}[(Y_t - \hat{\mu}_t)^2 - Y_t + \hat{h}_t\hat{\mu}_t]}{[2\sum_{t=1}^{n}\hat{\mu}_t^2]^{1/2}}.$$

where h_t is the tth diagonal value of the "hat" matrix. The "hat" matrix for generalized linear models extends that for linear regression and is defined in Fahrmeir and Tutz (1994, p.127), for example, as $H = \Lambda^{1/2}X(X^T\Lambda X)^{-1}X^T\Lambda^{1/2}$, where $\Lambda = \text{diag}(\mu_1, \cdots, \mu_n)$ and $X = (\boldsymbol{x}_{n1}^T, \cdots, \boldsymbol{x}_{nn}^T)^T$ is the design matrix. The sum of these h_t values is r (the dimension of $\boldsymbol{\beta}$). Both of these test statistics are one-sided and asymptotically distributed as an $N(0, 1)$ variate. Monte Carlo work reviewed in Brannas and Johansson (1994) suggests that S_a has better size properties in small samples.

We introduce an alternative test specifically designed for overdispersion due to the existence of a latent process in a Poisson observation process. This test uses higher moment properties of Poisson observations. Under the null hypothesis that there is no latent process (i.e., $\epsilon_t \equiv 1$), the Pearson residuals

$$\epsilon_t = \frac{Y_t - \hat{\mu}_t}{\sqrt{\hat{\mu}_t}}$$

have approximately zero mean and unit variance. Hence the statistic

$$Q = (\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2 - 1)/\hat{\sigma}_Q,$$

where

$$\hat{\sigma}_Q^2 = \frac{1}{n}(\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\hat{\mu}_t} + 2),$$

could be used to test for the presence of a latent process. The expression for $\hat{\sigma}_Q^2$ is easily derived using the fact that a Poisson random variable Y_t with mean μ_t has fourth central moment $E(Y_t - \mu_t)^4 = \mu_t + 3\mu_t^2$. Under the hypothesis that the variance of the latent process is zero,

$$Q \sim N(0, 1)$$

approximately.

Simulation results, based on the GLM fitted values from the analysis of the polio data, are used to assess how well the distribution of Q is approximated by the

asymptotic $N(0,1)$ distribution under the null hypothesis. 10000 replications of a time series of length $n = 168$ are obtained from simulating independent Poisson variates (i.e., with no latent process present) with mean $\mu_t = \hat{\mu}_t$, where $\hat{\mu}_t$ is the GLM fit to the polio data considered in Zeger (1988). Simulated type I errors using Q are obtained and given in Table 2.2. For each replicate in the simulation a model of the same form is fitted using the GLM procedures to the simulated data.

Table 2.2 Type I error for statistic Q.

α	0.100	0.050	0.025	0.010
Empirical $P(Q > z_{1-\alpha})$	0.031	0.012	0.005	0.002

These clearly indicate that Q as defined above does not achieve adequate nominal type I error probabilities and will be quite conservative for testing for the existence of a latent process. The mean and standard deviation of the Q were observed to be -0.27 and 0.78, respectively. The main source of the poor coverage is due to the negative bias in Q.

Alternate estimates of residuals which adjust for bias due to model fitting could be used. For example one could use the divisor $n - p$ instead of n in the numerator of Q. The resulting statistic had appreciably better performance than Q but was still on the conservative side. A second approach that is reasonably simple to implement would be to use standardized Pearson residuals defined as

$$\tilde{e}_t = e_t/(1 - h_t)^{0.5}.$$

Using these standardized residuals we define

$$\tilde{Q} = (\frac{1}{n}\sum_{t=1}^{n} \tilde{e}_t^2 - 1)/\hat{\sigma}_Q$$

and simulation results using 10000 replications based on this statistic are summarized in Table 2.3. The mean and standard deviation of $\tilde{Q}$ are -0.004 and

0.809, respectively and clearly this test statistic shows a clear improvement over the previous ones.

Table 2.3 Type I error for statistic $\tilde{Q}$.

α	0.100	0.050	0.025	0.010
Empirical $P(\tilde{Q} > z_{1-\alpha})$	0.063	0.028	0.013	0.005

Further work is required to thoroughly investigate the sampling distribution of $\tilde{Q}$. In the meantime we would recommend the use of the standardized Pearson residuals as in $\tilde{Q}$ but recognizing that the test for the existence of a latent process might be conservative. If this is not satisfactory simulation could be used to determine significance points for the distribution of $\tilde{Q}$. For example the significance points in Table 2.4 were obtained using the same simulation as in Table 2.3 and these values are at most 20% smaller than the values predicted from the asymptotic normal distribution.

Table 2.4 Significance points for statistics $\tilde{Q}$.

α	0.100	0.050	0.025	0.010
$z^*_{1-\alpha}$ s.t. $P(\tilde{Q} > z^*_{1-\alpha}) = 1-\alpha$	1.04	1.40	1.69	2.06

An alternative modification to Q which adjusts for the use of estimates $\hat{\mu}_t$ is

$$Q^* = \{\frac{1}{n}\sum_{t=1}^{n} \hat{\mu}_t^{-1} e^{\boldsymbol{x}_{nt}^T \hat{\Omega}_I^{-1} \boldsymbol{x}_{nt}/2}[(Y_t - \hat{\mu}_t)^2 + \hat{\mu}_t^2 \boldsymbol{x}_{nt}^T \hat{\Omega}_I^{-1} \boldsymbol{x}_{nt}] - 1\}/\hat{\sigma}_{Q^*},$$

where

$$\hat{\sigma}_{Q^*} = \sqrt{\frac{1}{n}(\frac{1}{n}\sum_{t=1}^{n} \hat{\mu}_t^{-1} e^{\boldsymbol{x}_{nt}^T \hat{\Omega}_I^{-1} \boldsymbol{x}_{nt}/2} + 2)}$$

and

$$\Omega_I = \sum_{t=1}^{n} \boldsymbol{x}_{nt} \boldsymbol{x}_{nt}^T e^{\boldsymbol{x}_{nt} \hat{\boldsymbol{\beta}}_{GLM}}.$$

Statistic Q^* showed marked improvement over Q.

We next compare the size and size adjusted power of the four statistics $Q, \tilde{Q}, Q^*$ and S_a. First the size properties are listed in Tables 2.5 and 2.6 for a simulation of 10000 replicates assuming no latent process is present.

Table 2.5 Type I errors for statistics $Q, \tilde{Q}, Q^*$ and S_a using model (2.1) with linear regression $\boldsymbol{x}_{nt}^T\boldsymbol{\beta} = 1 + 1t/100$ for $t = 1, \ldots, 100$.

α	0.100	0.050	0.025	0.010
Empirical $P(Q > z_{1-\alpha})$	0.075	0.039	0.019	0.010
Empirical $P(\tilde{Q} > z_{1-\alpha})$	0.097	0.051	0.030	0.014
Empirical $P(Q^* > z_{1-\alpha})$	0.097	0.050	0.029	0.013
Empirical $P(S_a > z_{1-\alpha})$	0.104	0.058	0.033	0.017

Table 2.6 Type I errors for statistics $Q, \tilde{Q}, Q^*$ and S_a using model (2.1) with cosine regression $\boldsymbol{x}_{nt}^T\boldsymbol{\beta} = 1 + 1\cos(2\pi t/12)$ for $t = 1, \ldots, 100$.

α	0.100	0.050	0.025	0.010
Empirical $P(Q > z_{1-\alpha})$	0.062	0.032	0.015	0.008
Empirical $P(\tilde{Q} > z_{1-\alpha})$	0.084	0.042	0.023	0.011
Empirical $P(Q^* > z_{1-\alpha})$	0.085	0.042	0.022	0.011
Empirical $P(S_a > z_{1-\alpha})$	0.096	0.058	0.035	0.018

In addition the power of the test to detect departures from the null hypothesis are investigated using 10000 replicates and the same mean models. Simulation results are given in Table 2.7. The latent process is generated using a lognormal distribution. The latent process has variance $\sigma_\tau^2 = 0.05$ chosen to give a small deviation from the null hypothesis. The autocovariance is simulated using an autoregressive process with $o = 0$ and $o = 0.9$. The results are for a size 0.05 test, based on 10000 replications.

Table 2.7 Power of the four statistics $Q, \tilde{Q}, Q^*$ and S_a for both linear and cosine regression cases.

	Linear regression $1 + 1t/100$		Cosine regression $1 + 1\cos(2\pi t/12)$	
	$o = 0$	$o = 0.9$	$o = 0$	$o = 0.9$
Power of Q	0.327	0.200	0.212	0.160
Power of $\tilde{Q}$	0.382	0.242	0.259	0.201
Power of Q^*	0.379	0.239	0.258	0.203
Power of S_a	0.439	0.232	0.376	0.271

On the basis of this limited simulation it would appear as though the S_a statistic has better type I error rates (i.e., closer to normal distribution) and

larger power. Note also that if the latent process has positive autocorrelation ($\phi = 0.9$), the power of all four statistics is reduced relative to the case of white noise latent process ($\phi = 0$). The test statistic S_a appeared to perform best and will be adopted in the remainder of this chapter. Further research is required to demonstrate that S_a is preferred in all circumstances.

2.4 Estimating the Variance and Autocovariances of the Latent Process

We now assume that the test based on S_a rejects the hypothesis of no latent process. Various estimates of the autocovariances have been suggested in the literature. We first review these, then suggest general weighted estimates from which "optimal" estimates are derived. Later in this section, adjustments for serious bias in autocovariance estimates will be suggested. Simulation results for comparing the estimates are provided.

2.4.1 Previous Estimates

Zeger (1988) proposed the following estimate

$$\hat{\sigma}^2_{\epsilon,Z} = \hat{\gamma}_{\epsilon,Z}(0) = \frac{\sum_{t=1}^{n}[(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t]}{\sum_{t=1}^{n}\hat{\mu}_t^2}$$

of σ_ϵ^2 which is exactly unbiased if $\hat{\mu}_t$ were replaced by the true value μ_t and, therefore, one might expect $\hat{\sigma}^2_{\epsilon,Z}$ to be approximately unbiased. He also suggested a method-of-moments type estimator for the ACVF and ACF of the latent process which are defined by

$$\hat{\gamma}_{\epsilon,Z}(\tau) = \sum_{t=\tau+1}^{n}\{(Y_t - \hat{\mu}_t)(Y_{t-\tau} - \hat{\mu}_{t-\tau})\} \Big/ \sum_{t=\tau+1}^{n}\hat{\mu}_t\hat{\mu}_{t-\tau}$$

and

$$\hat{\rho}_{\epsilon,Z}(\tau) = \hat{\gamma}_{\epsilon,Z}(\tau)/\hat{\sigma}^2_{\epsilon,Z}. \tag{2.16}$$

respectively.

Brannas and Johansson (1994) took a different approach to estimation of the ACVF. Their estimate,

$$\hat{\sigma}^2_{\epsilon,BJ} = \frac{\sum_{t=1}^{n} \hat{\mu}_t^2[(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t]}{\sum_{t=1}^{n} \hat{\mu}_t^4},$$

is derived by using ordinary least squares (OLS) regression of $(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t$ on its approximate expected value of $\sigma^2\hat{\mu}_t^2$. This same idea extends to other nonzero lags of the ACVF by noting that

$$E[(Y_t - \hat{\mu}_t)(Y_{t-\tau} - \hat{\mu}_{t-\tau})] \approx \rho_\epsilon(\tau)\sigma^2_\epsilon \mu_t \mu_{t-\tau}.$$

Regressing $(Y_t - \hat{\mu}_t)(Y_{t-\tau} - \hat{\mu}_{t-\tau})$ on $\hat{\mu}_t\hat{\mu}_{t-\tau}$ leads to the autocovariance estimates

$$\hat{\gamma}_{\epsilon,BJ}(\tau) = \sum_{t=\tau+1}^{n} \hat{\mu}_t\hat{\mu}_{t-\tau}\{(Y_t - \hat{\mu}_t)(Y_{t-\tau} - \hat{\mu}_{t-\tau})\} / \sum_{t=\tau+1}^{n} \hat{\mu}_t^2\hat{\mu}_{t-\tau}^2$$

and corresponding autocorrelation estimates

$$\hat{\rho}_{\epsilon,BJ}(\tau) = \hat{\gamma}_{\epsilon,BJ}(\tau)/\hat{\sigma}^2_{\epsilon,BJ},$$

where $\hat{\sigma}^2_{\epsilon,BJ}$ is as defined above. The Zeger or Brannas and Johansson estimates of variance and autocovariances are not guaranteed to form a non-negative definite sequence and therefore the autocorrelations are not guaranteed to be less than one in absolute values. As we shall see later, both methods produce estimates of autocovariances with large negative bias, while the autocorrelation estimates have considerably less bias. Most of the bias in the autocovariance estimates is directly attributable to the large bias in the estimate of σ^2_ϵ.

2.4.2 Optimally Weighted Estimates

The above estimates are not necessarily optimal in any sense other than being approximately unbiased. In particular the individual terms in the summations are not adjusted to account for unequal variances. By analogy with the use of weights in forming estimates of the variance of the latent process, some form

of weighting could be useful in forming autocovariance estimates. We consider weighted estimates of autocovariances which are required to be unbiased for any underlying stationary latent process. Calculation of the variance of any such estimates will require estimation and knowledge of the autocovariances. However, in the case where the latent process is a sequence of independent random variables the variance of these weighted estimates is readily computable as we will show. Since the hypothesis of independence is of primary interest, obtaining minimum variance estimates is desirable.

More generally, consider weighted estimates of the form:

$$\hat{\sigma}^2_{\epsilon,W} = (\sum_{t=1}^{n} W_t^{-2})^{-1} \sum_{t=1}^{n} W_t^{-2} E_t.$$

where

$$E_t = Y_t^{-2} - \hat{\mu}_t^{-1} \quad \text{and} \quad Y_t = (Y_t - \hat{\mu}_t)/\hat{\mu}_t.$$

Note that

$$\begin{aligned} E(\hat{\sigma}^2_{\epsilon,W}) &= (\sum_{t=1}^{n} W_t^{-2})^{-1} \sum_{t=1}^{n} W_t^{-2} E(E_t) \\ &\approx (\sum_{t=1}^{n} W_t^{-2})^{-1} \sum_{t=1}^{n} W_t^{-2} \mu_t^{-2} [E(Y_t - \mu_t)^2 - \mu_t] \\ &= (\sum_{t=1}^{n} W_t^{-2})^{-1} \sum_{t=1}^{n} W_t^{-2} \mu_t^{-2} [\mu_t + \mu_t^2 \sigma_\epsilon^2 - \mu_t] \\ &= \sigma_\epsilon^2. \end{aligned}$$

so that the weighted estimate is approximately unbiased. Zeger's estimate corresponds to choosing weights $W_t^{-2} = \mu_t^2$ while that of Brannas and Johansson corresponds to choosing weights $W_t^{-2} = \mu_t^4$. Note also that these weighted estimates are not guaranteed to be positive. However, it is unlikely that a negative estimate will be produced from these methods if the test of Section 2.3 supports the presence of a latent process.

It is straightforward to show that the optimal weights for minimizing the variance of $\hat{\sigma}^2_{\epsilon,W}$ are given by

$$W^2_{t,Opt} = 1/\mathrm{Var}(E_t).$$

Calculation of the variances required for this are complicated since they depend on moments up to order 4 for the latent process when it is a sequence of independent random variables. For latent processes with autocorrelation the calculation is further complicated. In addition these higher moments and autocovariances must be estimated.

Because of these potential complications we will limit the discussion to the case where the latent process is assumed to be a sequence of independent and identically distributed random variables. In order to carry out the calculation of third and higher moments we will also assume that $\epsilon_t = e^{\alpha_t}$ is a log-normal process. Then

$$\begin{aligned}\mathrm{Var}(E_t) &\approx \mu_t^{-4}\mathrm{Var}[(Y_t - \mu_t)^2 - \mu_t] \\ &= \mu_t^{-4}\{E[(Y_t - \mu_t)^4] - (\mu_t + \sigma_\epsilon^2\mu_t^2)^2\}.\end{aligned}$$

and using properties of the log-normal distribution.

$$\begin{aligned}E[(Y_t - \mu_t)^4] &= \{\mu_t + \mu_t^2[7e^{\sigma_\alpha^2} - 4] + \mu_t^3[6e^{3\sigma_\alpha^2} - 12e^{\sigma_\alpha^2} + 6] \\ &\quad + \mu_t^4[e^{6\sigma_\alpha^2} - 4e^{3\sigma_\alpha^2} + 6e^{\sigma_\alpha^2} - 3]\}.\end{aligned}$$

so that the optimal weights are given by

$$W^2_{t,Opt} = 1/\mathrm{Var}(E_t) = \mu_t^4/B_t.$$

and the optimal weighted estimator for variance is given by

$$\hat{\sigma}^2_{\epsilon,Opt} = \frac{\sum_{t=1}^n \frac{\hat{\mu}_t^2}{B_t}[(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t]}{\sum_{t=1}^n \frac{\hat{\mu}_t^4}{B_t}}. \qquad (2.17)$$

where

$$\begin{aligned} B_t &= E[(Y_t - \mu_t)^4] - (\mu_t + \sigma_\epsilon^2 \mu_t^2)^2 \\ &= \mu_t + \mu_t^2[7e^{\sigma_a^2} - 5] + \mu_t^3[6e^{3\sigma_a^2} - 14e^{\sigma_a^2} + 8] \\ &\quad + \mu_t^4[e^{6\sigma_a^2} - 4e^{3\sigma_a^2} - e^{2\sigma_a^2} + 8e^{\sigma_a^2} - 4]. \end{aligned} \tag{2.18}$$

Calculation of these optimal weights would require an iterated approach starting with an initial estimate of σ_a^2.

Note that the variance of the optimally weighted estimator is

$$\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Opt}) \approx \frac{1}{\sum_{t=1}^n (\mu_t^4 / B_t)}.$$

and that of Zeger's estimator is

$$\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Z}) \approx \frac{\sum_{t=1}^n B_t}{(\sum_{t=1}^n \mu_t^2)^2}.$$

For the polio data, using the GLM fit to obtain $\hat{\mu}_t$ and using the value of $\sigma_a^2 = 0.57$, the values of these variances are approximately $\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Opt}) \approx 0.46^2$ and $\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Z}) \approx 0.53^2$. These indicate a modest improvement in estimation of variance using the optimal weighting based on the (incorrect in this case) assumption that the latent process is independent.

As noted above, in the non-independent case, the calculation of optimal weights will be complicated by their dependence on unknown covariances. To implement the above optimal weighting scheme, an initial estimate of the variance σ_a^2 is required. One possibility is to use the weights based on the assumption that the latent process has zero variance and then use the resulting estimate to obtain the optimal weights.

Now turn to the estimates of covariances of the latent process. Consider the weighted estimates

$$\hat{\gamma}_{\epsilon,W}(h) = \frac{1}{\sum_{t=1}^{n-h} W_t W_{t+h}} \sum_{t=1}^{n-h} W_t W_{t+h} \hat{V}_t \hat{V}_{t+h}. \tag{2.19}$$

They are approximately unbiased for the true covariance since the individual terms satisfy

$$\begin{aligned} E(\hat{V}_t\hat{V}_{t+h}) &\approx (\mu_t\mu_{t+h})^{-1}\mathrm{Cov}(Y_t, Y_{t+h}) \\ &= \gamma_\epsilon(h). \end{aligned}$$

The approximate variance of these estimates, under the assumption that the latent process is white noise, is:

$$\mathrm{Var}(\hat{\gamma}_{\epsilon,W}(h)) \approx \frac{1}{(\sum_{t=1}^{n-h} W_t W_{t+h})^2} \sum_{t=1}^{n-h} W_t^2 W_{t+h}^2 (\sigma_\epsilon^2 + 1/\mu_t)(\sigma_\epsilon^2 + 1/\mu_{t+h}). \quad (2.20)$$

The following optimal weights can be derived directly from the finite sample approximation as

$$W_t^{Opt} = (\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)^{-1}$$

which gives the optimal estimates for autocovariance

$$\hat{\gamma}_{\epsilon,W^{Opt}}(h) = \frac{\sum_{t=1}^{n-h}(\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)^{-1}(\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_{t+h})^{-1} \frac{(Y_t-\hat{\mu}_t)(Y_{t+h}-\hat{\mu}_{t+h})}{\hat{\mu}_t\hat{\mu}_{t+h}}}{\sum_{t=1}^{n-h}(\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)^{-1}(\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_{t+h})^{-1}} \quad (2.21)$$

and corresponding optimal autocorrelation estimates

$$\hat{\rho}_{\epsilon,Opt}(h) = \hat{\gamma}_{\epsilon,W^{Opt}}(h)/\hat{\sigma}^2_{\epsilon,Opt}. \quad (2.22)$$

The optimal weight leads to an approximate asymptotic variance of the optimal estimates of autocovariances given by

$$\mathrm{Var}(\hat{\gamma}_{\epsilon,W^{Opt}}(h)) \approx \left[\sum_{t=1}^{n-h}(\sigma_\epsilon^2 + 1/\mu_t)^{-1}(\sigma_\epsilon^2 + 1/\mu_{t+h})^{-1}\right]^{-1}.$$

Zeger's (1988) estimates use weights $W_t^Z = \hat{\mu}_t$ which leads to an approximate variance

$$\mathrm{Var}(\hat{\gamma}_{\epsilon,W^Z}(h)) \approx \frac{1}{(\sum_{t=1}^{n-h}\mu_t\mu_{t+h})^2} \sum_{t=1}^{n-h} \mu_t^2\mu_{t+h}^2(\sigma_\epsilon^2 + 1/\mu_t)(\sigma_\epsilon^2 + 1/\mu_{t+h}).$$

Brannas and Johansson's estimator corresponds to weights $W_t^{BJ} = \hat{\mu}_t^2$ with corresponding approximate variance

$$\mathrm{Var}(\hat{\gamma}_{\epsilon,W^{BJ}}(h)) \approx \frac{1}{(\sum_{t=1}^{n-h} \mu_t^2 \mu_{t+h}^2)^2} \sum_{t=1}^{n-h} \mu_t^4 \mu_{t+h}^4 (\sigma_\epsilon^2 + 1/\mu_t)(\sigma_\epsilon^2 + 1/\mu_{t+h}).$$

We compare the approximate asymptotic variances of ACVF estimates for some sample mean functions. Consider the case where $\mu_t = g(t/n) = \exp(\boldsymbol{f}(t/n)^T \boldsymbol{\beta})$. Then the approximate variances for Zeger and the optimal estimates can be further approximated as follows. Under the assumption of an IID latent process $\{\epsilon_t\}$, for Zeger's estimates we have

$$n\mathrm{Var}(\hat{\gamma}_{\epsilon,W^Z}(h)) \approx I_Z := \frac{\int_0^1 g^2(x)(\sigma_\epsilon^2 g(x) + 1)^2 dx}{\left(\int_0^1 g^2(x)dx\right)^2}. \tag{2.23}$$

For the optimal estimates, we obtain

$$n\mathrm{Var}(\hat{\gamma}_{\epsilon,W^{Opt}}(h)) \approx I_{Opt} := \frac{1}{\int_0^1 g^2(x)(\sigma_\epsilon^2 g(x) + 1)^{-2} dx}. \tag{2.24}$$

Note that these estimates have the same variance when $g(x) = c$, a constant, with common value $(c\sigma_\epsilon^2 + 1)^2/c^2$. When $g(x)$ is not constant and $\sigma_\epsilon^2 = 0$, they also have a common variance of $1/\left(\int_0^1 g^2(x)dx\right)$.

It is easy to show that $I_Z \geq I_{Opt}$ using the Cauchy-Schwartz inequality. In general the larger σ_ϵ^2 and the more variation in the mean term μ_t, the larger the difference between the asymptotic variance for Zeger's proposal and that of the optimal estimates. The integrals can be evaluated readily using standard numerical methods.

2.4.3 Bias Adjustments for Estimates of Autocovariances

In this section we consider bias-correction strategies for the estimators of autocovariance. Exploratory simulations show that for the Zeger estimate of σ_ϵ^2, the numerator seriously underestimates $\sum_{t=1}^{n}[(Y_t - \mu_t)^2 - \mu_t]$ while the denominator

overestimates $\sum_{t=1}^{n}\mu_t^2$. Clearly, both directions of bias in these terms contribute to the bias in the ratio.

To correct for the bias in the estimation of $\hat{\mu}_t$, we use the asymptotics from Theorem 2.1. First note that we may write

$$\hat{\mu}_t = \mu_t \exp(\boldsymbol{x}_{nt}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)).$$

Since $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is approximately normally distributed with mean $\mathbf{0}$ and covariance matrix $G_n = \Omega_{I,n}^{-1} + \Omega_{I,n}^{-1}\Omega_{II,n}\Omega_{I,n}^{-1}$, $\hat{\mu}_t$ has an approximate log-normal distribution with mean

$$E(\hat{\mu}_t) = \mu_t E[\exp(\boldsymbol{x}_{nt}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))] = \mu_t \exp(\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}/2) \tag{2.25}$$

and second moment

$$E(\hat{\mu}_t^2) = \mu_t^2 E[\exp(2\boldsymbol{x}_{nt}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))] = \mu_t^2 \exp(2\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}). \tag{2.26}$$

In other words $\hat{\mu}_t$ and $\hat{\mu}_t^2$ have positive bias. Now consider the term

$$\frac{1}{n}\sum_{t=1}^{n}(Y_t - \mu_t)^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2 + \frac{1}{n}\sum_{t=1}^{n}(\hat{\mu}_t - \mu_t)^2 + \frac{2}{n}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)(\hat{\mu}_t - \mu_t).$$

In analogy with standard regression theory, the last term is negligible. Using (2.25) and (2.26), the second term has approximate expectation given by

$$E\left[\frac{1}{n}\sum_{t=1}^{n}(\hat{\mu}_t - \mu_t)^2\right] \approx \frac{1}{n}\sum_{t=1}^{n}\mu_t^2\left(e^{2\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}} - 2e^{\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}/2} + 1\right). \tag{2.27}$$

A bias-corrected estimate of $\frac{1}{n}\sum_{t=1}^{n}(\hat{\mu}_t - \mu_t)^2$ is then

$$\frac{1}{n}\sum_{t=1}^{n}\hat{\mu}_t^2 e^{-2\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}}\left(e^{2\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}} - 2e^{\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt}/2} + 1\right).$$

Turning to the denominator of $\hat{\sigma}_{r,Z}^2$, an approximate unbiased estimator of μ_t^2 is $\hat{\mu}_t^2 \exp(-2\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt})$ so that

$$E\left(\sum_{t=1}^{n}\hat{\mu}_t^2 \exp(-2\boldsymbol{x}_{nt}^T G_n \boldsymbol{x}_{nt})\right) \approx \sum_{t=1}^{n}\mu_t^2.$$

Based on these approximations to adjust for the biases in the numerator and denominator of Zeger's estimate $\sigma^2_{\epsilon,Z}$, we suggest the estimate of variance given by

$$\hat{\sigma}^2_{Z,UB} = \frac{\sum_{t=1}^{n}\left[(Y_t-\hat{\mu}_t)^2 + \hat{\mu}_t^2\epsilon^{-2d_t}(\epsilon^{2d_t} - 2\epsilon^{d_t/2}+1) - \hat{\mu}_t\epsilon^{-d_t/2}\right]}{\sum_{t=1}^{n}\hat{\mu}_t^2\epsilon^{-2d_t}}.$$

where $d_t = \boldsymbol{x}_{nt}^T\hat{G}_n\boldsymbol{x}_{nt}$, and in which the estimate of the asymptotic covariance matrix is given by

$$\hat{G}_n = \hat{\Omega}_{I,n}^{-1} + \hat{\Omega}_{I,n}^{-1}\hat{\Omega}_{II,n}\hat{\Omega}_{I,n}^{-1}. \tag{2.28}$$

where

$$\hat{\Omega}_{I,n} = \sum_{t=1}^{n}\boldsymbol{x}_{nt}\boldsymbol{x}_{nt}^T\hat{\mu}_t$$

and

$$\hat{\Omega}_{II,n} = \sum_{h=-L}^{L}\sum_{t=\max(1-h,1)}^{\min(n-h,n)}\boldsymbol{x}_{nt}\boldsymbol{x}_{n,t+h}^T\hat{\mu}_t\hat{\mu}_{t+h}\hat{\gamma}_{\epsilon,Z}(h)$$

for some maximum lag L specified somewhat arbitrarily. Although the estimates $\hat{\gamma}_{\epsilon,Z}(h)$ are biased we do not recommend iterated estimates obtained by replacing $\hat{\gamma}_{\epsilon,Z}(h)$ with updated $\hat{\gamma}_{Z,UB}(h)$. Such a replacement if iterated would lead to divergent estimates because the adjustment term in the denominator would get progressively closer to zero. An alternative procedure is to model the $\hat{\gamma}_{\epsilon,Z}(h)$ by use of an autoregressive process or other suitable model. This does not appear to lead to better estimates than those based on the non-parametric forms proposed above. In fact the estimates appear to be worse as measured by bias and variance.

Employing similar approximations we suggest the following bias-corrected estimates for the autocovariances.

$$\hat{\gamma}_{Z,UB}(h) = \frac{\sum_{t=1}^{n-h}\left[(Y_t-\hat{\mu}_t)(Y_{t+h}-\hat{\mu}_{t+h}) + \hat{\mu}_t\hat{\mu}_{t+h}\epsilon^{-c_{t,h}/2}(\epsilon^{c_{t,h}/2} - \epsilon^{d_t/2} - \epsilon^{d_{t+h}/2}+1)\right]}{\sum_{t=1}^{n-h}\hat{\mu}_t\hat{\mu}_{t+h}\epsilon^{-c_{t,h}/2}}.$$

where $c_{t,h} = (\boldsymbol{x}_{nt}+\boldsymbol{x}_{n,t+h})^T\hat{G}_n(\boldsymbol{x}_{nt}+\boldsymbol{x}_{n,t+h})$, and $d_t = \boldsymbol{x}_{nt}^T\hat{G}_n\boldsymbol{x}_{nt}$.

The corresponding autocorrelation estimates are given by

$$\hat{\rho}_{Z,UB}(h) = \hat{\gamma}_{Z,UB}(h)/\hat{\sigma}^2_{Z,UB}. \tag{2.29}$$

Following the same idea, we get the following bias-adjusted optimal estimates for variance:

$$\hat{\sigma}^2_{OPT,UB} = \frac{\sum_{t=1}^{n} \frac{\hat{W}^2_{t,Opt}}{\hat{\mu}_t^2}\left[(Y_t - \hat{\mu}_t)^2 + \hat{\mu}_t^2 e^{-2d_t}(e^{2d_t} - 2e^{d_t/2} + 1) - \hat{\mu}_t e^{-d_t/2}\right]}{\sum_{t=1}^{n} \hat{W}^2_{t,Opt}}.$$

where

$$\hat{W}^2_{t,Opt} = \hat{\mu}_t^4 / \hat{B}_t,$$

and $\hat{B}_t$ is the estimated version of B_t given in (2.18). Similarly, the bias-corrected optimal estimate of the covariance function takes the form

$$\hat{\gamma}_{OPT,UB}(h) = \frac{\sum_{t=1}^{n-h} \frac{\hat{W}_t \hat{W}_{t+h}}{\hat{\mu}_t \hat{\mu}_{t+h}}\left[(Y_t - \hat{\mu}_t)(Y_{t+h} - \hat{\mu}_{t+h}) + \hat{\mu}_t \hat{\mu}_{t+h} e^{-c_{t,h}/2}(e^{c_{t,h}/2} - e^{d_t/2} - e^{d_{t+h}/2} + 1)\right]}{\sum_{t=1}^{n-h} \hat{W}_t \hat{W}_{t+h}}.$$

where

$$\hat{W}_t = (\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)^{-1}.$$

The corresponding autocorrelation estimates are

$$\hat{\rho}_{OPT,UB}(h) = \hat{\gamma}_{OPT,UB}(h)/\hat{\sigma}^2_{OPT,UB}. \tag{2.30}$$

It is straightforward to show that these adjusted estimates are consistent. In particular $\hat{G}_n$ will converge to zero as $n \to \infty$. This implies that the adjustment to the denominator will tend to unity while that of the numerator will tend to zero as is required for the adjustments to be asymptotically negligible.

2.4.4 Comparison of the Estimates

We now describe our simulation results to compare the performance of the Zeger and bias-adjusted Zeger estimates along with the optimal and bias-corrected optimal estimates for autocovariances and autocorrelations of the latent process. In Tables 2.8 - 2.11 we report sample means and sample standard deviations (SD) of these estimators over 5000 replications for two Poisson regression models with different parameters and sample sizes. For each replicate of generated count data

series of length n, $\hat{\mu}_t$ was obtained from a GLM fit. The latent process used in the simulation is either IID or correlated. Since the optimal estimates were derived under the assumption of IID latent process, we would expect optimal estimates to perform better than Zeger estimates in the case of an IID latent process.

In the computation of the bias-adjustment factors, we used $L = 10$ in (2.28) for all realizations. The autocorrelation results $\hat{\rho}_r(h)$ are conditional on $S_a > 1.645$. This was done to eliminate values of autocorrelations badly affected by zero or near zero variance estimates.

We summarize the results for sample size $n = 100$ first (Tables 2.8 and 2.9). Under the IID latent process ($\alpha = 0$) and for both linear and cosine regression cases, the optimal covariance estimates have smaller variances than Zeger estimates although they have similar bias. This holds true for autocorrelated $\{\epsilon_t\}$ and cosine regression case. When $\{\epsilon_t\}$ is autocorrelated and the regression is linear, optimal covariance estimates are slightly less biased than Zeger estimates but have larger variances. The bias-corrected estimates reduce the bias but at the expense of higher variance. Both the bias-adjusted Zeger and bias-adjusted optimal estimates still have same magnitude of bias, the latter has smaller variance.

With the increase of the variance in the latent process (σ_α^2 is from $\log 2$ to $\log 5$), the bias increases in all estimates. When there is substantial autocorrelation present ($\alpha = 0.9, \sigma_\alpha^2 = \log 5$), there is also substantial bias in the estimates of the autocovariances $\gamma_r(h)$ - this occurs in all estimates to a similar degree. This would impact the estimation of the correct asymptotic variance in the GLM estimates of $\boldsymbol{\beta}$ of Theorem 2.1. Positive serial correlation in the latent process would tend to lead to underestimating the correct standard errors of the GLM estimates.

For the estimation of autocorrelations $\hat{\rho}_r(h)$ the bias is not as severe. The bias-adjusted estimates of autocorrelations have better bias properties, although the unadjusted estimates perform reasonably well. The bias-adjusted estimates

of autocorrelations are slightly better in standard deviations than the unadjusted estimates. For some purposes such as construction of a suitable correlation model for use in an efficient estimation procedure this reduction in bias is a good property. This means that for correlation estimation and model identification purposes, the bias-adjusted estimates are preferable to the unadjusted versions. However for purposes in which an unbiased estimate of scale is required even the bias-adjusted estimates of $\hat{\sigma}_\epsilon^2$ are biased towards 0. This large bias impacts the magnitude of the bias correction.

All biases in $n = 100$ simulations depend on the form of the regressor with worse bias for the linear trend regression than the cosine regression. This indicates that any bias adjustment procedure should account for the form of the regression function.

When sample size n increases to 1000 (Tables 2.10 and 2.11), biases in all estimates become smaller. The unadjusted estimates have the most significant decrease in bias. For latent process with modest variance $\sigma_\alpha^2 = \log 2$, the bias has been nearly eradicated. With the increase of the variance in the latent process, the biases increase. The optimal estimates have similar biases as Zeger estimates but consistently smaller variances in all cases. Bias-adjusted estimates have smaller biases but larger variances. Again, all versions of the autocorrelation estimates work well.

Overall, the optimal estimates outperform Zeger estimates. The bias-adjusted optimal estimates outperform the bias-adjusted Zeger estimates. The performance advantages decrease as the sample size increases.

Table 2.8 Autocovariance and autocorrelation estimates of a log-normal AR(1) latent process $\{\epsilon_t\}$ in model (2.1) with linear regression function $\boldsymbol{x}_t^T\boldsymbol{\beta} = 1 + 1t/n$. Sample size $n = 100$.

			Means (SD)			
		True	Z	Z.UB	OPT	OPT.UB
$\phi = 0$	$\hat{\sigma}_\epsilon^2$	1.00	.904 (.405)	.955 (.449)	.894 (.368)	.942 (.406)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	0	-.027 (.117)	-.011 (.125)	-.024 (.107)	-.008 (.115)
	$\hat{\gamma}_\epsilon(2)$	0	-.025 (.113)	-.009 (.122)	-.021 (.106)	-.006 (.113)
	$\hat{\rho}_\epsilon(1)$	0	-.030 (.129)	-.013 (.130)	-.028 (.121)	-.011 (.123)
0†	$\hat{\rho}_\epsilon(2)$	0	-.027 (.128)	-.011 (.129)	-.024 (.120)	-.008 (.121)
$\phi = 0$	$\hat{\sigma}_\epsilon^2$	4.00	2.86 (2.01)	3.32 (3.22)	2.75 (1.79)	3.16 (2.80)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	0	-.071 (.273)	-.028 (.320)	-.051 (.250)	-.008 (.295)
	$\hat{\gamma}_\epsilon(2)$	0	-.069 (.280)	-.025 (.333)	-.052 (.250)	-.009 (.298)
	$\hat{\rho}_\epsilon(1)$	0	-.024 (.104)	-.009 (.104)	-.020 (.097)	-.006 (.098)
0†	$\hat{\rho}_\epsilon(2)$	0	-.025 (.105)	-.010 (.105)	-.021 (.097)	-.007 (.098)
$\phi = 0.9$	$\hat{\sigma}_\epsilon^2$	1.00	.487 (.281)	.714 (.602)	.508 (.297)	.723 (.577)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	.866	.390 (.253)	.585 (.540)	.411 (.264)	.596 (.515)
	$\hat{\gamma}_\epsilon(2)$	.753	.310 (.225)	.487 (.484)	.330 (.234)	.497 (.462)
	$\hat{\gamma}_\epsilon(3)$	.657	.242 (.200)	.402 (.436)	.262 (.207)	.413 (.417)
	$\hat{\gamma}_\epsilon(4)$	.576	.186 (.180)	.332 (.394)	.205 (.190)	.344 (.381)
	$\hat{\gamma}_\epsilon(5)$	.506	.139 (.162)	.272 (.352)	.157 (.169)	.284 (.341)
	$\hat{\gamma}_\epsilon(6)$	.445	.098 (.148)	.220 (.319)	.117 (.156)	.234 (.310)
	$\hat{\rho}_\epsilon(1)$	.866	.788 (.182)	.797 (.158)	.803 (.186)	.807 (.162)
	$\hat{\rho}_\epsilon(2)$	.753	.611 (.208)	.640 (.186)	.629 (.206)	.652 (.184)
	$\hat{\rho}_\epsilon(3)$	.657	.464 (.225)	.510 (.207)	.485 (.216)	.524 (.200)
	$\hat{\rho}_\epsilon(4)$	.576	.347 (.235)	.406 (.220)	.370 (.226)	.421 (.212)
	$\hat{\rho}_\epsilon(5)$	.506	.252 (.242)	.322 (.231)	.276 (.233)	.338 (.221)
10†	$\hat{\rho}_\epsilon(6)$	.445	.172 (.245)	.251 (.236)	.197 (.237)	.268 (.228)
$\phi = 0.9$	$\hat{\sigma}_\epsilon^2$	4.00	1.24 (.831)	2.37 (3.17)	1.38 (1.19)	2.45 (3.10)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	3.26	.919 (.674)	1.82 (2.55)	1.08 (1.71)	1.95 (3.37)
	$\hat{\gamma}_\epsilon(2)$	2.68	.684 (.542)	1.43 (2.03)	.827 (1.66)	1.55 (3.06)
	$\hat{\gamma}_\epsilon(3)$	2.23	.504 (.447)	1.13 (1.67)	.632 (1.55)	1.25 (2.79)
	$\hat{\gamma}_\epsilon(4)$	1.88	.366 (.378)	.901 (1.40)	.482 (1.50)	1.01 (2.64)
	$\hat{\gamma}_\epsilon(5)$	1.59	.259 (.336)	.722 (1.22)	.370 (1.48)	.835 (2.53)
	$\hat{\gamma}_\epsilon(6)$	1.35	.169 (.306)	.569 (1.05)	.266 (1.13)	.670 (1.96)
	$\hat{\rho}_\epsilon(1)$	.814	.727 (.157)	.739 (.140)	.754 (.225)	.756 (.223)
	$\hat{\rho}_\epsilon(2)$	.671	.533 (.184)	.564 (.169)	.563 (.253)	.584 (.255)
	$\hat{\rho}_\epsilon(3)$	.558	.386 (.202)	.431 (.189)	.420 (.264)	.455 (.268)
	$\hat{\rho}_\epsilon(4)$	.469	.275 (.208)	.331 (.198)	.310 (.251)	.355 (.251)
	$\hat{\rho}_\epsilon(5)$	.397	.188 (.209)	.253 (.202)	.227 (.261)	.281 (.265)
7†	$\hat{\rho}_\epsilon(6)$	.338	.118 (.213)	.189 (.207)	.158 (.245)	.219 (.248)

† Number of S_a which was less than 1.645 out of 5000 replications.

Table 2.9 Autocovariance and autocorrelation estimates of a log-normal AR(1) latent process $\{\epsilon_t\}$ in model (2.1) with cosine regression function $\boldsymbol{x}_t^T\boldsymbol{\beta} = 1 + 1\cos(2\pi t/12)$. Sample size $n = 100$.

			Means (SD)			
		True	Z	Z.UB	OPT	OPT.UB
$\phi = 0$	$\hat{\sigma}_\epsilon^2$	1.00	.877 (.476)	.961 (.574)	.875 (.414)	.960 (.480)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	0	-.028 (.137)	-.005 (.152)	-.025 (.120)	-.002 (.131)
	$\hat{\gamma}_\epsilon(2)$	0	-.016 (.148)	.003 (.161)	-.016 (.127)	.002 (.137)
	$\hat{\rho}_\epsilon(1)$	0	-.032 (.162)	-.007 (.162)	-.029 (.146)	-.004 (.144)
0†	$\hat{\rho}_\epsilon(2)$	0	-.019 (.172)	-.000 (.169)	-.019 (.154)	.001 (.150)
$\phi = 0$	$\hat{\sigma}_\epsilon^2$	4.00	2.66 (2.04)	3.40 (4.50)	2.80 (2.09)	3.54 (5.21)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	0	-.066 (.313)	.001 (.412)	-.047 (.277)	.022 (.402)
	$\hat{\gamma}_\epsilon(2)$	0	-.048 (.339)	.004 (.432)	-.037 (.310)	.019 (.601)
	$\hat{\rho}_\epsilon(1)$	0	-.026 (.126)	-.004 (.126)	-.019 (.108)	.001 (.106)
0†	$\hat{\rho}_\epsilon(2)$	0	-.019 (.135)	-.002 (.131)	-.016 (.112)	.000 (.108)
$\phi = 0.9$	$\hat{\sigma}_\epsilon^2$	1.00	.614 (.401)	.831 (.746)	.605 (.366)	.825 (.659)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	.866	.506 (.364)	.694 (.679)	.509 (.336)	.693 (.597)
	$\hat{\gamma}_\epsilon(2)$	.753	.427 (.326)	.594 (.598)	.426 (.300)	.591 (.531)
	$\hat{\gamma}_\epsilon(3)$	.657	.361 (.297)	.509 (.527)	.359 (.274)	.505 (.478)
	$\hat{\gamma}_\epsilon(4)$	.576	.306 (.273)	.439 (.472)	.302 (.258)	.434 (.444)
	$\hat{\gamma}_\epsilon(5)$	.506	.259 (.242)	.378 (.416)	.256 (.236)	.376 (.406)
	$\hat{\gamma}_\epsilon(6)$	.445	.213 (.218)	.323 (.373)	.212 (.217)	.322 (.372)
	$\hat{\rho}_\epsilon(1)$	.866	.816 (.195)	.818 (.174)	.836 (.209)	.828 (.182)
	$\hat{\rho}_\epsilon(2)$	.753	.682 (.223)	.692 (.204)	.692 (.222)	.694 (.199)
	$\hat{\rho}_\epsilon(3)$	.657	.573 (.259)	.587 (.243)	.573 (.247)	.582 (.228)
	$\hat{\rho}_\epsilon(4)$	.576	.481 (.286)	.500 (.277)	.474 (.266)	.488 (.253)
	$\hat{\rho}_\epsilon(5)$	.506	.409 (.303)	.433 (.300)	.399 (.280)	.419 (.272)
34†	$\hat{\rho}_\epsilon(6)$	.445	.339 (.301)	.371 (.303)	.328 (.280)	.356 (.276)
$\phi = 0.9$	$\hat{\sigma}_\epsilon^2$	4.00	1.52 (1.03)	2.61 (2.92)	1.52 (.968)	2.58 (2.62)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	3.26	1.17 (.833)	2.03 (2.35)	1.19 (.794)	2.02 (2.12)
	$\hat{\gamma}_\epsilon(2)$	2.68	.941 (.720)	1.65 (1.96)	.944 (.666)	1.63 (1.75)
	$\hat{\gamma}_\epsilon(3)$	2.23	.774 (.648)	1.37 (1.69)	.764 (.589)	1.34 (1.52)
	$\hat{\gamma}_\epsilon(4)$	1.88	.643 (.587)	1.15 (1.49)	.630 (.536)	1.12 (1.35)
	$\hat{\gamma}_\epsilon(5)$	1.59	.526 (.513)	.960 (1.28)	.518 (.491)	.946 (1.21)
	$\hat{\gamma}_\epsilon(6)$	1.35	.414 (.437)	.786 (1.05)	.412 (.436)	.786 (1.05)
	$\hat{\rho}_\epsilon(1)$	.814	.763 (.168)	.766 (.152)	.783 (.188)	.771 (.161)
	$\hat{\rho}_\epsilon(2)$	.671	.610 (.206)	.618 (.191)	.614 (.204)	.613 (.182)
	$\hat{\rho}_\epsilon(3)$	.558	.500 (.231)	.510 (.222)	.492 (.214)	.497 (.199)
	$\hat{\rho}_\epsilon(4)$	.469	.417 (.251)	.431 (.248)	.404 (.228)	.412 (.220)
	$\hat{\rho}_\epsilon(5)$	.397	.342 (.258)	.362 (.263)	.328 (.236)	.341 (.232)
13†	$\hat{\rho}_\epsilon(6)$	.338	.274 (.255)	.302 (.264)	.261 (.235)	.283 (.234)

† Number of S_a which was less than 1.645 out of 5000 replications.

Table 2.10 Autocovariance and autocorrelation estimates of a log-normal AR(1) latent process $\{\epsilon_t\}$ in model (2.1) with linear regression function $\boldsymbol{x}_t^T\boldsymbol{\beta} = 1 + 1t/n$. Sample size $n = 1000$.

			Means (SD)			
		True	Z	Z.UB	OPT	OPT.UB
$o = 0$	$\hat{\sigma}_\epsilon^2$	1.00	.989 (.211)	.997 (.215)	.986 (.169)	.993 (.172)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	0	-.003 (.042)	-.001 (.042)	-.003 (.038)	-.001 (.039)
	$\hat{\gamma}_\epsilon(2)$	0	-.002 (.041)	.000 (.042)	-.002 (.038)	.000 (.038)
	$\hat{\rho}_\epsilon(1)$	0	-.004 (.042)	-.001 (.042)	-.003 (.039)	-.001 (.039)
0†	$\hat{\rho}_\epsilon(2)$	0	-.002 (.042)	.000 (.042)	-.002 (.038)	.000 (.038)
$o = 0$	$\hat{\sigma}_\epsilon^2$	4.00	3.73 (1.94)	3.82 (2.13)	3.70 (1.76)	3.78 (1.92)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	0	-.009 (.137)	-.001 (.142)	-.007 (.121)	.001 (.125)
	$\hat{\gamma}_\epsilon(2)$	0	-.011 (.128)	-.002 (.132)	-.007 (.116)	.000 (.119)
	$\hat{\rho}_\epsilon(1)$	0	-.003 (.035)	-.000 (.035)	-.002 (.032)	.000 (.032)
0†	$\hat{\rho}_\epsilon(2)$	0	-.003 (.035)	-.000 (.035)	-.002 (.032)	.000 (.032)
$o = 0.9$	$\hat{\sigma}_\epsilon^2$	1.00	.903 (.325)	.975 (.410)	.901 (.282)	.967 (.336)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	.866	.776 (.299)	.840 (.377)	.775 (.261)	.833 (.311)
	$\hat{\gamma}_\epsilon(2)$	.753	.670 (.274)	.728 (.347)	.669 (.239)	.722 (.285)
	$\hat{\gamma}_\epsilon(3)$	.657	.579 (.249)	.633 (.315)	.578 (.217)	.627 (.260)
	$\hat{\gamma}_\epsilon(4)$	.576	.502 (.222)	.552 (.280)	.501 (.197)	.547 (.235)
	$\hat{\gamma}_\epsilon(5)$	.506	.435 (.201)	.482 (.253)	.435 (.179)	.478 (.214)
	$\hat{\gamma}_\epsilon(6)$	.445	.378 (.182)	.422 (.229)	.378 (.164)	.418 (.195)
	$\hat{\rho}_\epsilon(1)$	.866	.855 (.036)	.856 (.035)	.856 (.039)	.857 (.038)
	$\hat{\rho}_\epsilon(2)$	.753	.734 (.051)	.738 (.051)	.736 (.051)	.739 (.050)
	$\hat{\rho}_\epsilon(3)$	.657	.632 (.064)	.639 (.063)	.634 (.061)	.640 (.061)
	$\hat{\rho}_\epsilon(4)$	.576	.547 (.073)	.556 (.072)	.549 (.069)	.557 (.068)
	$\hat{\rho}_\epsilon(5)$	.506	.473 (.080)	.484 (.079)	.475 (.075)	.485 (.074)
0†	$\hat{\rho}_\epsilon(6)$	.445	.410 (.085)	.422 (.084)	.412 (.079)	.424 (.079)
$o = 0.9$	$\hat{\sigma}_\epsilon^2$	4.00	3.04 (1.82)	3.69 (3.14)	3.01 (1.76)	3.59 (2.98)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	3.26	2.45 (1.50)	2.98 (2.60)	2.42 (1.44)	2.89 (2.39)
	$\hat{\gamma}_\epsilon(2)$	2.68	1.99 (1.24)	2.45 (2.16)	1.97 (1.19)	2.37 (1.99)
	$\hat{\gamma}_\epsilon(3)$	2.23	1.64 (1.03)	2.03 (1.80)	1.63 (1.00)	1.97 (1.69)
	$\hat{\gamma}_\epsilon(4)$	1.88	1.36 (.864)	1.70 (1.51)	1.35 (.861)	1.65 (1.44)
	$\hat{\gamma}_\epsilon(5)$	1.59	1.14 (.739)	1.43 (1.29)	1.13 (.745)	1.39 (1.25)
	$\hat{\gamma}_\epsilon(6)$	1.35	.955 (.628)	1.21 (1.09)	.953 (.647)	1.18 (1.09)
	$\hat{\rho}_\epsilon(1)$	.814	.800 (.047)	.802 (.047)	.800 (.048)	.801 (.048)
	$\hat{\rho}_\epsilon(2)$	.671	.651 (.068)	.656 (.068)	.651 (.066)	.655 (.065)
	$\hat{\rho}_\epsilon(3)$	.558	.537 (.081)	.544 (.081)	.537 (.077)	.543 (.077)
	$\hat{\rho}_\epsilon(4)$	.469	.447 (.089)	.457 (.089)	.448 (.084)	.456 (.084)
	$\hat{\rho}_\epsilon(5)$	.397	.374 (.093)	.385 (.092)	.376 (.087)	.385 (.086)
0†	$\hat{\rho}_\epsilon(6)$	.338	.316 (.094)	.327 (.093)	.317 (.087)	.328 (.087)

† Number of S_a which was less than 1.645 out of 5000 replications.

Table 2.11 Autocovariance and autocorrelation estimates of a log-normal AR(1) latent process $\{\epsilon_t\}$ in model (2.1) with cosine regression function $\boldsymbol{x}_t^T\boldsymbol{\beta} = 1 + 1\cos(2\pi t/12)$. Sample size $n = 1000$.

			Means (SD)			
		True	Z	Z.UB	OPT	OPT.UB
$\varphi = 0$	$\hat{\sigma}_\epsilon^2$	1.00	.990 (.267)	.999 (.273)	.986 (.180)	.996 (.184)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	0	-.003 (.051)	-.001 (.052)	-.003 (.043)	-.001 (.043)
	$\hat{\gamma}_\epsilon(2)$	0	-.002 (.053)	.000 (.053)	-.001 (.044)	.001 (.045)
	$\hat{\rho}_\epsilon(1)$	0	-.003 (.052)	-.001 (.052)	-.003 (.044)	-.001 (.044)
0†	$\hat{\rho}_\epsilon(2)$	0	-.002 (.053)	.000 (.053)	-.001 (.045)	.001 (.045)
$\varphi = 0$	$\hat{\sigma}_\epsilon^2$	4.00	3.65 (2.10)	3.76 (2.34)	3.73 (1.82)	3.82 (1.96)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	0	-.009 (.168)	.001 (.175)	-.007 (.128)	.002 (.132)
	$\hat{\gamma}_\epsilon(2)$	0	-.009 (.158)	-.002 (.163)	-.007 (.123)	-.001 (.126)
	$\hat{\rho}_\epsilon(1)$	0	-.003 (.044)	-.000 (.044)	-.002 (.034)	.000 (.034)
0†	$\hat{\rho}_\epsilon(2)$	0	-.003 (.043)	-.001 (.043)	-.002 (.034)	-.000 (.034)
$\varphi = 0.9$	$\hat{\sigma}_\epsilon^2$	1.00	.936 (.348)	.972 (.379)	.927 (.299)	.965 (.328)
$\sigma_a^2 = \log 2$	$\hat{\gamma}_\epsilon(1)$	.866	.807 (.317)	.839 (.344)	.804 (.280)	.836 (.304)
	$\hat{\gamma}_\epsilon(2)$	.753	.700 (.285)	.729 (.311)	.698 (.258)	.726 (.280)
	$\hat{\gamma}_\epsilon(3)$	.657	.610 (.255)	.636 (.278)	.607 (.237)	.634 (.258)
	$\hat{\gamma}_\epsilon(4)$	.576	.532 (.228)	.556 (.249)	.531 (.218)	.555 (.238)
	$\hat{\gamma}_\epsilon(5)$	.506	.465 (.203)	.487 (.222)	.464 (.200)	.486 (.218)
	$\hat{\gamma}_\epsilon(6)$	.445	.407 (.185)	.428 (.202)	.407 (.185)	.428 (.202)
	$\hat{\rho}_\epsilon(1)$	.866	.858 (.043)	.859 (.042)	.863 (.052)	.862 (.051)
	$\hat{\rho}_\epsilon(2)$	.753	.743 (.063)	.745 (.063)	.746 (.060)	.746 (.059)
	$\hat{\rho}_\epsilon(3)$	.657	.648 (.081)	.650 (.081)	.647 (.069)	.649 (.069)
	$\hat{\rho}_\epsilon(4)$	.576	.567 (.095)	.570 (.095)	.564 (.076)	.567 (.076)
	$\hat{\rho}_\epsilon(5)$	.506	.497 (.103)	.501 (.104)	.492 (.083)	.496 (.083)
0†	$\hat{\rho}_\epsilon(6)$	.445	.435 (.106)	.440 (.107)	.431 (.087)	.435 (.088)
$\varphi = 0.9$	$\hat{\sigma}_\epsilon^2$	4.00	3.12 (1.96)	3.46 (2.73)	3.14 (1.77)	3.46 (2.42)
$\sigma_a^2 = \log 5$	$\hat{\gamma}_\epsilon(1)$	3.26	2.53 (1.61)	2.81 (2.26)	2.55 (1.49)	2.82 (2.05)
	$\hat{\gamma}_\epsilon(2)$	2.68	2.10 (1.38)	2.34 (1.94)	2.11 (1.26)	2.33 (1.75)
	$\hat{\gamma}_\epsilon(3)$	2.23	1.77 (1.18)	1.96 (1.66)	1.76 (1.10)	1.95 (1.55)
	$\hat{\gamma}_\epsilon(3)$	1.88	1.49 (1.01)	1.66 (1.44)	1.49 (.960)	1.65 (1.37)
	$\hat{\gamma}_\epsilon(3)$	1.59	1.27 (.859)	1.41 (1.23)	1.26 (.841)	1.41 (1.20)
	$\hat{\gamma}_\epsilon(3)$	1.35	1.08 (.747)	1.20 (1.07)	1.08 (.747)	1.20 (1.07)
	$\hat{\rho}_\epsilon(1)$	.814	.808 (.056)	.808 (.056)	.809 (.055)	.809 (.055)
	$\hat{\rho}_\epsilon(2)$	.671	.671 (.085)	.672 (.085)	.667 (.071)	.667 (.071)
	$\hat{\rho}_\epsilon(3)$	.558	.568 (.103)	.569 (.104)	.557 (.081)	.558 (.081)
	$\hat{\rho}_\epsilon(2)$	.469	.484 (.114)	.486 (.116)	.470 (.087)	.471 (.088)
	$\hat{\rho}_\epsilon(2)$	.397	.415 (.119)	.417 (.122)	.399 (.091)	.401 (.092)
0†	$\hat{\rho}_\epsilon(2)$	.338	.356 (.121)	.359 (.124)	.340 (.093)	.343 (.094)

† Number of S_a which was less than 1.645 out of 5000 replications.

For an IID latent process $\{\epsilon_t\}$ with $\sigma_\epsilon^2 = 1$ (i.e., $\sigma_\alpha^2 = \log 2$), the asymptotic variances in (2.23) and (2.24) give the ratio of asymptotic variances $I_Z/I_{Opt} = 1.814/1.506 = 1.205$ in linear regression case and $I_Z/I_{Opt} = 2.877/1.896 = 1.517$ in the cosine case. The corresponding simulated ratios of variances $\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Z})/\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Opt})$, when $n = 1000$, are $(.211/.169)^2 = 1.559$ and $(.267/.180)^2 = 2.200$ respectively. They are somewhat larger than the theoretical large sample values. When $\phi = 0.9$ and $\sigma_\epsilon^2 = 1$, the corresponding ratios are $(.325/.282)^2 = 1.328$ for linear case and $(.348/.299)^2 = 1.355$ for cosine case, these values are close to the theoretical asymptotic variances. The asymptotic formulas for $\mathrm{Var}(\hat{\gamma}_{\epsilon,W}(h))$ give unbiased estimates of the simulated variances when there is no serial dependence, i.e., in the situation that they are derived under. This means that this formula will be useful in calculating an overall test of autocovariance (see Section 2.5 below). Further the asymptotic formulas provide reasonable estimates of standard deviations even in the cases where the null hypothesis of no serial dependence is not true.

2.5 Tests for Zero Autocorrelation in the Latent Process

Typically, tests of zero correlation in a stationary process are based on functions of the sample ACF. Under the assumption that the process is white noise, the sample ACF at distinct lags are approximately independent $N(0, n^{-1})$ distributed from which asymptotic cutoff values of test statistics can be computed. In the Poisson model setting, if the latent process is white noise with positive variance, then it can be shown that the $\hat{\rho}_{\epsilon,Z}(h)$ are asymptotically distributed as independent normal random variables with standard deviations $\mathrm{s.e.}(\hat{\rho}_{\epsilon,Z}(h))$ that may be different than $n^{-1/2}$.

The Box-Pierce (BP) and Ljung-Box (LB) portmanteau statistics, which are weighted sums of the sample ACF at a fixed number of lags, are often the basis of tests for zero correlation in a time series. For our Poisson model, the main

difficulty related to the use of the BP test is due to the nonconstant variance of the sample ACF of the residuals. Since the variance and covariances of the process have different forms of dependence on the mean function μ_t, there is no single normalization of the residuals that will simultaneously eliminate this dependence from the variance and the covariance. To overcome heteroscedasticity in the variance, we approximate the variance of the sample ACF in the BP statistic under the null hypothesis of a white noise latent process. This approach is similar in spirit to the methods proposed by Lo and MacKinlay (1989) for testing the hypothesis that the increments in a random walk process are uncorrelated when the increments are weakly dependent with a possibly heteroscedastic variance. Also see Lobato, Nankervis and Savin (1998) for modification of the BP statistic in other applications.

Brannas and Johansson (1994) studied the performance of the BP and LB statistics based on three sets of residuals: the Pearson residuals

$$e_t = (Y_t - \hat{\mu}_t)/\hat{\mu}_t^{1/2},$$

the Anscombe residuals

$$e_{tA} = 3(Y_t^{2/3} - \hat{\mu}_t^{2/3})/(2\hat{\mu}_t^{1/6}),$$

and their own residuals defined as

$$e_{t,BJ} = [1 + (\hat{\sigma}_\epsilon^2 \hat{\mu}_t)^{-1}]^{1/2}(Y_t - \hat{\mu}_t).$$

In the definition of the latter, the exact form of $\hat{\sigma}_\epsilon^2$ is not specified and presumably could be any of the variance estimates considered in Sections 2.4.1 - 2.4.3. Unfortunately, and contrary to Brannas and Johansson's claim, none of these residuals approximate the correlation structure of the latent process. They observed that for the the BP and LB statistics constructed from any of these residuals "the

sizes are significantly too high". The large size may be due in part to the bias of the estimates and a nonconstant variance of the individual estimates of autocorrelations. Even if these test statistics based on the sample ACF of residuals could be adjusted to achieve the nominal type I errors, such tests will have little power against some alternative models with large correlations but small variance. To illustrate this point, consider the sample ACF based on the Pearson residuals defined by

$$\hat{\rho}_P(h) = \frac{\sum_{t=1}^{n-h} \epsilon_t \epsilon_{t+h}}{\sum_{t=1}^{n} \epsilon_t^2}.$$

Mean correction of the ϵ_t is not utilized since their sample mean is near 0. Approximating the mean of the ratio by the ratio of the means and using results from Section 2.1, we obtain

$$\begin{aligned} E(\hat{\rho}_P(h)) &\approx \frac{n^{-1}\sum_{t=1}^{n-h} E(\epsilon_t \epsilon_{t+h})}{n^{-1}\sum_{t=1}^{n} E(\epsilon_t^2)} \\ &= \frac{n^{-1}\sum_{t=1}^{n-h} \mu_t^{1/2}\mu_{t+h}^{1/2}\gamma_r(h)}{n^{-1}\sum_{t=1}^{n}(1+\sigma_r^2\mu_t)} \\ &= \frac{n^{-1}\sum_{t=1}^{n-h} \mu_t^{1/2}\mu_{t+h}^{1/2}}{n^{-1}\sum_{t=1}^{n}(\sigma_r^{-2}+\mu_t)}\rho_r(h). \end{aligned} \tag{2.31}$$

Under the assumption that $\boldsymbol{x}_{nt} = \boldsymbol{f}(t/n)$, the last line can be approximated by

$$\frac{\int_0^1 e^{\boldsymbol{f}(x)\boldsymbol{\beta}} dx}{\sigma_r^{-2} + \int_0^1 e^{\boldsymbol{f}(x)\boldsymbol{\beta}} dx}\rho_r(h).$$

As $\sigma_r^2 \to 0$, we see that the mean value of $\hat{\rho}_P(h)$ becomes arbitrarily small. Thus for alternatives consisting of a highly correlated latent process for which σ_r^2 is small, the sample ACF of the Pearson residuals may not provide any evidence of correlation and correlation tests based on Pearson residuals will often fail to reject the null hypothesis of white noise.

Under the null hypothesis of a white noise latent process for our Poisson model, the process $\{\epsilon_t\}$ is heteroscedastic and nonstationary. Following the same lines of reasoning as in Lo and MacKinlay (1989) and Lobato, Nankervis and

Savin (1998), the variance of the correlations under the white noise latent process assumption, can be approximated by

$$\begin{aligned}
\mathrm{Var}(\hat{\rho}_P(h)) &\approx [\frac{1}{n}\sum_{t=1}^{n}\mu_t^{-1}E(Y_t-\mu_t)^2]^{-2}[\frac{1}{n^2}\sum_{t=1}^{n-h}\mu_t^{-1}\mu_{t+h}^{-1}E(Y_t-\mu_t)^2E(Y_{t+h}-\mu_{t+h})^2] \\
&= [\frac{1}{n}\sum_{t=1}^{n}\mu_t^{-1}(\mu_t+\mu_t^2\sigma_\epsilon^2)]^{-2}[\frac{1}{n^2}\sum_{t=1}^{n-h}\mu_t^{-1}\mu_{t+h}^{-1}(\mu_t+\mu_t^2\sigma_\epsilon^2)(\mu_{t+h}+\mu_{t+h}^2\sigma_\epsilon^2)] \\
&= [\frac{1}{n}\sum_{t=1}^{n}(1+\mu_t\sigma_\epsilon^2)]^{-2}[\frac{1}{n^2}\sum_{t=1}^{n-h}(1+\mu_t\sigma_\epsilon^2)(1+\mu_{t+h}\sigma_\epsilon^2)] \\
&:= V_h/n.
\end{aligned}$$

We calculate the variance of sample ACF, V_h/n, for linear trend and cosine regression examples used in Section 2.4.4 and compare them with the value of $(n-h)/[n(n+2)]$ used in the LB statistic. For a given maximum lag L, the standard Ljung-Box statistic, based on Pearson residuals, is given by

$$H_{LB}^2 = n(n+2)\sum_{h=1}^{L}[\hat{\rho}_P^2(h)/(n-h)]. \tag{2.32}$$

Results are listed in Table 2.12 for $n = 100$.

Table 2.12 Variances of the sample autocorrelations for linear and cosine regression models and which used in LB statistic.

Lag h	1	2	3	4	5	6	7	8	9	10	11	12	13
V_h lin	1.04	1.03	1.02	1.00	.99	.98	.97	.95	.94	.93	.92	.90	.89
V_h cos	1.19	1.08	.95	.83	.75	.72	.74	.80	.90	1.01	1.08	1.10	1.05
$\frac{n-h}{n+2}$	.97	.96	.95	.94	.93	.92	.91	.90	.89	.88	.87	.86	.85

The differences are greatest for the cosine regression as might be expected. For some of the lags the correct standard errors are considerably different from those used in the LB statistic, e.g., for the cosine regression case, $V_h/n = 1.19/n$ and $(n-h)/[n(n+2)] = 0.97/n$ for $h = 1$, and $V_h/n = 0.72/n$ and $(n-h)/[n(n+2)] = 0.92/n$ for $h = 6$. This implies that care should be taken when testing a particular lag autocorrelation using the standard estimation technique based on the Pearson residuals.

A natural statistic (analogous to the Box-Jenkins's portmanteau statistic) is proposed for testing for serial dependence in the mean of the observed count time series. For a given maximum lag L and a given set of general estimators $\hat{\rho}_\epsilon(h)$ that are asymptotically normal, define

$$H^2 = \sum_{h=1}^{L} [\hat{\rho}_\epsilon(h)/s.e.(\hat{\rho}_\epsilon(h))]^2. \tag{2.33}$$

Under the hypothesis of independence of the latent process, H^2 will have an approximate χ^2 distribution on L d.f. The standard Ljung-Box statistic H^2_{LB} based on Pearson residuals is a special form of (2.33) with ACF $\hat{\rho}_P(h)$ and variance $(n-h)/[n(n+2)]$. The modified LB statistic, which takes the standard errors into account, is defined by

$$H^2_{LB,M} = n \sum_{h=1}^{L} \hat{\rho}^2_P(h)/\hat{V}_h. \tag{2.34}$$

where $\hat{V}_h$ is an estimate of V_h.

Simulation results comparing the relative performance of six test statistics are summarized in Tables 2.13-2.16. The statistics H^2_Z, $H^2_{Z,LB}$, H^2_{Opt} and $H^2_{Opt,LB}$ in the tables refer to the test statistic in (2.33) using Zeger ACF estimate (2.16), bias-adjusted Zeger ACF estimate (2.29), optimal ACF estimate (2.22) and bias-adjusted optimal ACF estimate (2.30) respectively. The standard error $s.e.(\hat{\rho}_\epsilon(h))$ is determined by equation (2.20) with appropriate weights. The sample size for all realizations is 100 or 1000 and the summary statistics are based on 5000 replications. Both linear and cosine regression functions are used in the simulation.

Table 2.13 lists type I errors of the six test statistics for the linear regression case. With variance $\sigma^2_a = \log 2$ (which is equivalent to $\sigma^2_\epsilon = 1$) and sample size n increases from 100 to 1000, type I errors of H^2_Z, H^2_{Opt}, H^2_{LB} and $H^2_{LB,M}$ increase dramatically. When the variance σ^2_a increases to $\log 5$ (i.e., $\sigma^2_\epsilon = 4$), type I errors are almost doubled over the increase of n except for $H^2_{Z,LB}$ and $H^2_{Opt,LB}$. In all

cases type I errors of $H^2_{Z,UB}$ and $H^2_{Opt,UB}$ only have little changes. They are greater than those of H^2_Z and H^2_{Opt} respectively.

In Table 2.13, the differences of type I errors among all test statistics are more obvious for $n = 100$ than those for $n = 1000$. When sample size is 100, with the exception of the $H^2_{Z,UB}$ and $H^2_{Opt,UB}$ tests, all produce less frequent significant results than would be the case if the test statistics were distributed exactly as chi square. For the case of $n = 1000$ and $\sigma^2_\alpha = \log 2$, the obtained sizes of H^2_Z, H^2_{Opt} and $H^2_{LB,M}$ are closer to the nominal size than other test statistics.

Table 2.13 Type I errors for tests of zero autocorrelations in a lognormal latent process $\{c_t = e^{\alpha_t}\}$ for Poisson model (2.1). The regression function is $\boldsymbol{x}^T_{nt}\boldsymbol{\beta} = 1 + 1t/n$.

	α	0.100	0.050	0.025	0.010	5% Critical
$n = 100$	$P(H^2_Z > \chi^2_{10,(1-\alpha)})$	0.056	0.028	0.015	0.007	16.42
$\sigma^2_\alpha = \log 2$	$P(H^2_{Z,UB} > \chi^2_{10,(1-\alpha)})$	0.086	0.053	0.030	0.018	18.46
	$P(H^2_{Opt} > \chi^2_{10,(1-\alpha)})$	0.063	0.035	0.020	0.009	16.79
	$P(H^2_{Opt,UB} > \chi^2_{10,(1-\alpha)})$	0.102	0.059	0.038	0.020	19.06
	$P(H^2_{LB} > \chi^2_{10,(1-\alpha)})$	0.085	0.047	0.023	0.010	17.92
	$P(H^2_{LB,M} > \chi^2_{10,(1-\alpha)})$	0.062	0.033	0.016	0.007	16.72
$n = 1000$	$P(H^2_Z > \chi^2_{10,(1-\alpha)})$	0.092	0.049	0.029	0.015	18.25
$\sigma^2_\alpha = \log 2$	$P(H^2_{Z,UB} > \chi^2_{10,(1-\alpha)})$	0.100	0.058	0.034	0.017	18.81
	$P(H^2_{Opt} > \chi^2_{10,(1-\alpha)})$	0.095	0.052	0.027	0.014	18.45
	$P(H^2_{Opt,UB} > \chi^2_{10,(1-\alpha)})$	0.103	0.057	0.033	0.016	18.86
	$P(H^2_{LB} > \chi^2_{10,(1-\alpha)})$	0.115	0.068	0.037	0.018	19.57
	$P(H^2_{LB,M} > \chi^2_{10,(1-\alpha)})$	0.094	0.051	0.028	0.013	18.39
$n = 100$	$P(H^2_Z > \chi^2_{10,(1-\alpha)})$	0.038	0.020	0.011	0.004	14.82
$\sigma^2_\alpha = \log 5$	$P(H^2_{Z,UB} > \chi^2_{10,(1-\alpha)})$	0.091	0.063	0.043	0.026	19.40
	$P(H^2_{Opt} > \chi^2_{10,(1-\alpha)})$	0.050	0.030	0.016	0.007	16.01
	$P(H^2_{Opt,UB} > \chi^2_{10,(1-\alpha)})$	0.109	0.076	0.058	0.038	21.34
	$P(H^2_{LB} > \chi^2_{10,(1-\alpha)})$	0.062	0.038	0.020	0.010	16.92
	$P(H^2_{LB,M} > \chi^2_{10,(1-\alpha)})$	0.043	0.024	0.015	0.006	15.45
$n = 1000$	$P(H^2_Z > \chi^2_{10,(1-\alpha)})$	0.079	0.055	0.041	0.028	19.04
$\sigma^2_\alpha = \log 5$	$P(H^2_{Z,UB} > \chi^2_{10,(1-\alpha)})$	0.092	0.065	0.047	0.034	20.11
	$P(H^2_{Opt} > \chi^2_{10,(1-\alpha)})$	0.094	0.059	0.041	0.029	19.37
	$P(H^2_{Opt,UB} > \chi^2_{10,(1-\alpha)})$	0.106	0.070	0.050	0.034	20.40
	$P(H^2_{LB} > \chi^2_{10,(1-\alpha)})$	0.111	0.072	0.053	0.037	20.70
	$P(H^2_{LB,M} > \chi^2_{10,(1-\alpha)})$	0.091	0.058	0.042	0.028	19.24

Table 2.14 Powers of six tests for a lognormal AR(1) latent process $\{\epsilon_t = e^{a_t}\}$ with parameter o. The regression function in Poisson model (2.1) is $\boldsymbol{x}_{nt}^T\boldsymbol{\beta} = 1 + 1t/n$. The power is estimated by the fraction of times that the test statistic values are greater than the 5% critical values listed in Table 2.13.

		$o = 0.2$	$o = 0.4$	$o = 0.6$	$o = 0.9$
$n = 100$	H_Z^2	0.098	0.307	0.662	0.932
$\sigma_a^2 = \log 2$	$H_{Z,LB}^2$	0.118	0.354	0.731	0.959
	H_{Opt}^2	0.113	0.358	0.742	0.952
	$H_{Opt,LB}^2$	0.128	0.392	0.796	0.968
	H_{LB}^2	0.104	0.343	0.733	0.953
	$H_{LB,M}^2$	0.107	0.345	0.731	0.956
$n = 1000$	H_Z^2	0.597	0.999	1.000	1.000
$\sigma_a^2 = \log 2$	$H_{Z,LB}^2$	0.604	0.999	1.000	1.000
	H_{Opt}^2	0.685	1.000	1.000	1.000
	$H_{Opt,LB}^2$	0.701	1.000	1.000	1.000
	H_{LB}^2	0.670	1.000	1.000	1.000
	$H_{LB,M}^2$	0.676	1.000	1.000	1.000
$n = 100$	H_Z^2	0.086	0.258	0.593	0.939
$\sigma_a^2 = \log 5$	$H_{Z,LB}^2$	0.101	0.298	0.657	0.958
	H_{Opt}^2	0.090	0.290	0.678	0.968
	$H_{Opt,LB}^2$	0.102	0.313	0.710	0.971
	H_{LB}^2	0.088	0.285	0.670	0.976
	$H_{LB,M}^2$	0.090	0.287	0.657	0.973
$n = 1000$	H_Z^2	0.315	0.931	0.998	1.000
$\sigma_a^2 = \log 5$	$H_{Z,LB}^2$	0.323	0.938	0.998	1.000
	H_{Opt}^2	0.390	0.967	0.999	1.000
	$H_{Opt,LB}^2$	0.400	0.970	0.999	1.000
	H_{LB}^2	0.372	0.959	0.999	1.000
	$H_{LB,M}^2$	0.371	0.958	0.999	1.000

Table 2.14 gives the powers of six tests for the linear regression case. The powers of all statistics are quite close. With same variance in the latent process, the powers increase as the sample size n increases. For fixed sample size, as variance σ_a^2 increases, the powers decrease when there are moderate autocorrelations present ($o = 0.2$ and $o = 0.4$). Throughout linear regression cases, test $H_{Opt,LB}^2$ is uniformly more powerful than the rest of tests after the sizes of all tests have been calibrated to the nominal level.

Table 2.15 Type I errors for tests of zero autocorrelations in a lognormal latent process $\{\epsilon_t = e^{a_t}\}$ for Poisson model (2.1). The regression function is $\boldsymbol{x}_{nt}^T\boldsymbol{\beta} = 1 + 1\cos(2\pi t/12)$.

	α	0.100	0.050	0.025	0.010	5% Critical
$n = 100$	$P(H_Z^2 > \chi^2_{10,(1-\alpha)})$	0.103	0.058	0.041	0.022	19.10
$\sigma_a^2 = \log 2$	$P(H_{Z,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.154	0.097	0.064	0.042	22.05
	$P(H_{Opt}^2 > \chi^2_{10,(1-\alpha)})$	0.087	0.048	0.030	0.017	18.13
	$P(H_{Opt,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.135	0.091	0.057	0.034	21.34
	$P(H_{LB}^2 > \chi^2_{10,(1-\alpha)})$	0.077	0.042	0.024	0.014	17.70
	$P(H_{LB,M}^2 > \chi^2_{10,(1-\alpha)})$	0.089	0.048	0.029	0.016	18.19
$n = 1000$	$P(H_Z^2 > \chi^2_{10,(1-\alpha)})$	0.108	0.063	0.040	0.018	19.23
$\sigma_a^2 = \log 2$	$P(H_{Z,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.116	0.067	0.041	0.020	19.53
	$P(H_{Opt}^2 > \chi^2_{10,(1-\alpha)})$	0.106	0.058	0.032	0.014	18.88
	$P(H_{Opt,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.114	0.062	0.034	0.016	19.07
	$P(H_{LB}^2 > \chi^2_{10,(1-\alpha)})$	0.085	0.046	0.025	0.012	17.92
	$P(H_{LB,M}^2 > \chi^2_{10,(1-\alpha)})$	0.108	0.056	0.031	0.015	18.58
$n = 100$	$P(H_Z^2 > \chi^2_{10,(1-\alpha)})$	0.109	0.072	0.051	0.028	20.56
$\sigma_a^2 = \log 5$	$P(H_{Z,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.181	0.130	0.100	0.072	26.27
	$P(H_{Opt}^2 > \chi^2_{10,(1-\alpha)})$	0.069	0.042	0.023	0.012	17.56
	$P(H_{Opt,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.149	0.104	0.073	0.049	23.09
	$P(H_{LB}^2 > \chi^2_{10,(1-\alpha)})$	0.063	0.035	0.020	0.010	16.97
	$P(H_{LB,M}^2 > \chi^2_{10,(1-\alpha)})$	0.080	0.050	0.029	0.016	18.34
$n = 1000$	$P(H_Z^2 > \chi^2_{10,(1-\alpha)})$	0.125	0.090	0.068	0.050	23.28
$\sigma_a^2 = \log 5$	$P(H_{Z,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.133	0.097	0.075	0.055	24.11
	$P(H_{Opt}^2 > \chi^2_{10,(1-\alpha)})$	0.103	0.070	0.049	0.034	20.39
	$P(H_{Opt,UB}^2 > \chi^2_{10,(1-\alpha)})$	0.110	0.079	0.057	0.039	21.27
	$P(H_{LB}^2 > \chi^2_{10,(1-\alpha)})$	0.092	0.064	0.046	0.033	19.83
	$P(H_{LB,M}^2 > \chi^2_{10,(1-\alpha)})$	0.108	0.077	0.056	0.038	21.24

Type I errors of the six tests for cosine regression case are summarized in Table 2.15. When variance $\sigma_a^2 = \log 2$ and sample size n increases from 100 to 1000, type I errors of H_Z^2, H_{Opt}^2, H_{LB}^2 and $H_{LB,M}^2$ increase slightly. Magnitude of the increase is bigger when variance $\sigma_a^2 = \log 5$. Unlike linear cases, type I errors of $H_{Z,UB}^2$ and $H_{Opt,UB}^2$ decrease as sample size increases. The bias-adjusted versions have bigger sizes which are too high than their unadjusted counterparts, these differences in sizes are more significant when $n = 100$. Statistics H_{Opt}^2, H_{LB}^2 and $H_{LB,M}^2$ have smaller sizes than the nominal ones when sample size is 100.

For $n = 1000$, the sizes of H^2_{LB} are closer to the nominal sizes than other test statistics.

Table 2.16 Powers of six tests for a lognormal AR(1) latent process $\{\epsilon_t = e^{\alpha_t}\}$ with parameter ϕ. The regression function in Poisson model (2.1) is $\boldsymbol{x}_{nt}^T\boldsymbol{\beta} = 1 + 1\cos(2\pi t/12)$. The power is estimated by the fraction of times that the test statistic values are greater than the 5% critical values listed in Table 2.15.

		$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$	$\phi = 0.9$
$n = 100$	H^2_Z	0.073	0.201	0.472	0.914
$\sigma^2_\alpha = \log 2$	$H^2_{Z,UB}$	0.098	0.261	0.572	0.939
	H^2_{Opt}	0.083	0.266	0.607	0.939
	$H^2_{Opt,UB}$	0.108	0.330	0.678	0.955
	H^2_{LB}	0.093	0.294	0.653	0.950
	$H^2_{LB,M}$	0.083	0.252	0.598	0.941
$n = 1000$	H^2_Z	0.377	0.979	1.000	1.000
$\sigma^2_\alpha = \log 2$	$H^2_{Z,UB}$	0.395	0.982	1.000	1.000
	H^2_{Opt}	0.550	0.999	1.000	1.000
	$H^2_{Opt,UB}$	0.578	0.999	1.000	1.000
	H^2_{LB}	0.611	0.999	1.000	1.000
	$H^2_{LB,M}$	0.527	0.998	1.000	1.000
$n = 100$	H^2_Z	0.071	0.141	0.384	0.929
$\sigma^2_\alpha = \log 5$	$H^2_{Z,UB}$	0.099	0.220	0.527	0.951
	H^2_{Opt}	0.083	0.241	0.603	0.965
	$H^2_{Opt,UB}$	0.119	0.316	0.691	0.973
	H^2_{LB}	0.098	0.267	0.635	0.971
	$H^2_{LB,M}$	0.081	0.210	0.547	0.962
$n = 1000$	H^2_Z	0.180	0.758	0.990	1.000
$\sigma^2_\alpha = \log 5$	$H^2_{Z,UB}$	0.189	0.775	0.992	1.000
	H^2_{Opt}	0.341	0.952	0.999	1.000
	$H^2_{Opt,UB}$	0.353	0.956	0.999	1.000
	H^2_{LB}	0.354	0.946	0.999	1.000
	$H^2_{LB,M}$	0.271	0.904	0.999	1.000

Table 2.16 lists the powers of all six test statistics for cosine case. Throughout the powers are much higher for all test statistics in the $n = 1000$ case. As the variance σ^2_α increases, the powers decrease for all the tests in n=1000 case when there are moderate autocorrelations present. For $n = 100$, the powers of $H^2_{Opt,UB}$ are bigger than the others.

From these tables we see that the relative performance of these test statistics is mixed and highly dependent on the form of the regression functions. Throughout the simulation, H^2_{Opt} provides approximately similar performance to the modified Ljung-Box statistic $H^2_{LB,M}$. In the linear case, $H^2_{Opt,UB}$ and $H^2_{Z,UB}$ are superior than all other statistics. In the cosine case, when $n = 1000$, H^2_{LB} is more powerful than any other statistics; when $n = 100$, $H^2_{Opt,UB}$ is more powerful than others; and H^2_{Opt} and $H^2_{LB,M}$ statistics give better type I error rates.

2.6 Example

We illustrate the above methods with an application to counts of daily admissions for asthma to Campbelltown Hospital in the Sydney, Australia metropolitan area. Figure 2.4 shows the daily number of asthma presentations from January 1, 1990 - December 31, 1993. An analysis of temporally related effects identified the following: (i) no upward or downward trend in counts at this location; (ii) a triple peaked annual cycle modeled by pairs of the form $\cos(2\pi kt/365).\sin(2\pi kt/365)$, where t is the day number and $k = 1, 2, 3, 4, 5, 8$; and (iii) a day of the week effect best characterized by separate indicator variables for Sundays and Mondays to model the increased level of admittance for these days compared to that for Tuesday to Saturday.

A detailed preliminary analysis of the possible effects of meteorological variables, daily maximum and minimum temperatures and humidity, and pollution variables ozone, NO and NO_2, identified humidity at lags of approximately 12 to 18 days as the only variable that appears to have an association with asthma presentations. A humidity variable, H_t, was constructed as

$$H_t = \frac{1}{7}\sum_{i=0}^{6} h_{t-12-i},$$

where h_t is the residual from an annual-cycle harmonic model fitted to the daily average value of humidity at 0900 and 1500 hours.

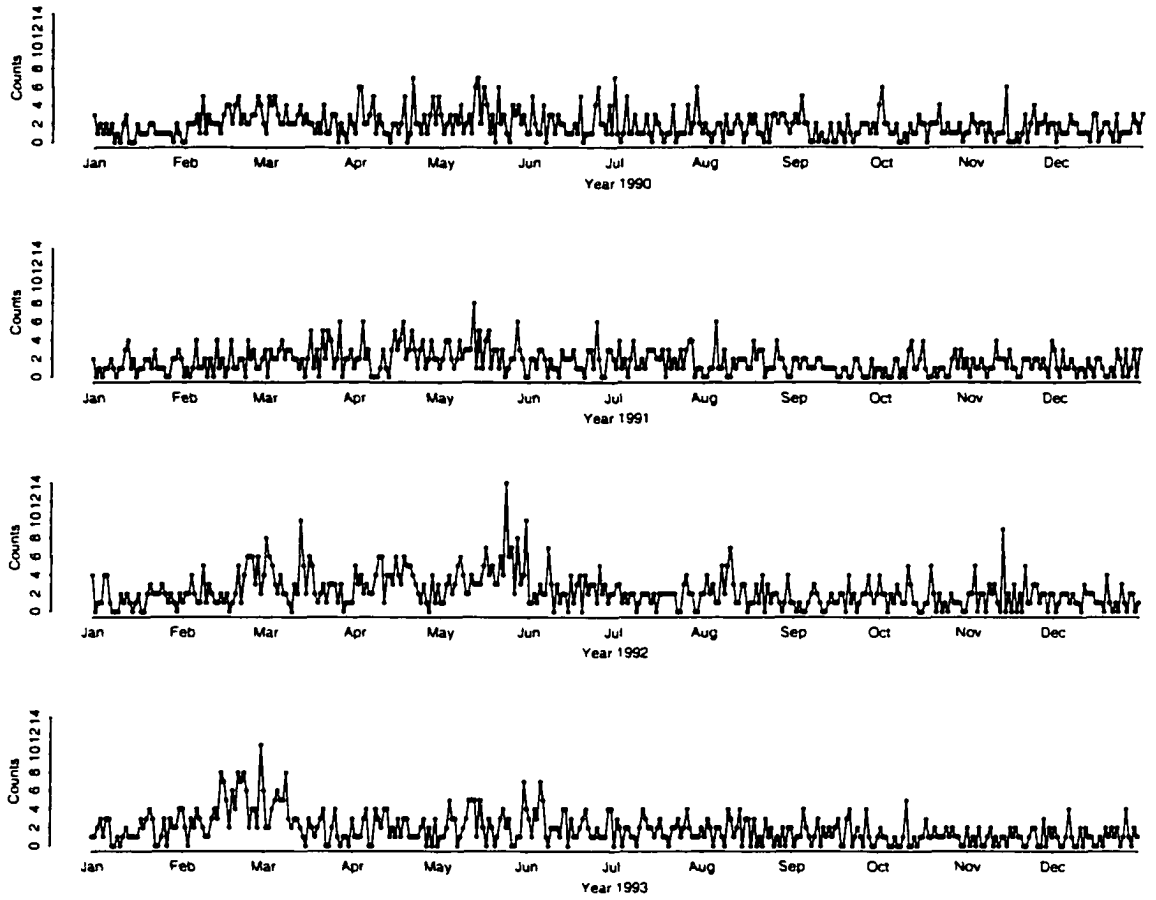

Figure 2.4: Asthma presentations at Campbelltown hospital during January 1, 1990 - December 31, 1993.

In the following tables, the results for the coefficients of the harmonic components are not given because they are of no interest to the analysis. Results of the ordinary generalized linear model fit as well as various standard errors and diagnostic procedures for the presence of a latent process are provided in Tables 2.17 and 2.18.

Table 2.17 Generalized linear model regression estimators for Campbelltown asthma count series. The two standard errors are computed using a standard generalized linear model analysis without a latent process and using Theorem 2.1, respectively.

Effect	$\hat{\beta}$	GLM s.e.$\{\hat{\beta}\}$	$\hat{G}$ s.e.$\{\hat{\beta}\}$
Sunday	0.230	0.051	0.055
Monday	0.236	0.051	0.055
H_t	0.210	0.048	0.066

GLM s.e.$\{\hat{\beta}\}$, standard error of $\hat{\beta}$ based on a generalized linear model fit assuming independence; $\hat{G}$ s.e.$\{\hat{\beta}\}$, standard error of $\hat{\beta}$ based on the asymptotic covariance matrix given in (2.28).

The t-ratios for the coefficient of humidity for the two estimators of standard error in Table 2.17 are 4.41 and 3.19, respectively. It is also clear from Table 2.17 that the effect of lagged seven-day average humidity is highly significant when we use the proper standard error based on the asymptotic covariance results of Theorem 2.1.

The statistic S_a has an observed value of 3.30, which is highly significant and clearly indicates the presence of a latent process. Table 2.18 reports values of test statistics for autocorrelation in the latent process. The modified Ljung-Box statistic based on Pearson residuals indicates significant autocorrelation. The portmanteau test based on bias-adjusted autocorrelations, $\hat{\rho}_{Z,UB}(h)$, provides stronger evidence of serial correlation in the latent process than does the modified Ljung-Box test based on Pearson residuals.

Table 2.18 Tests of correlation in the latent process for the asthma data with p-values in parentheses.

	Degrees of freedom		
Test statistic	5	10	15
$H^2_{Z,UB}$	44.63 (1.72×10^{-8})	74.86 (5.08×10^{-12})	81.32 (4.00×10^{-11})
$H^2_{LB,M}$	10.78 (0.056)	25.60 (0.004)	26.83 (0.030)

Table 2.19 Autocovariance and autocorrelation estimates for the asthma data.

lag h	$\hat{\gamma}_Z(h)$	s.e.$\{\hat{\gamma}_Z(h)\}$	$\hat{\gamma}_{Z,UB}(h)$	$\hat{\rho}_Z(h)$	$\hat{\rho}_{Z,UB}(h)$	s.e.$\{\hat{\rho}_{Z,UB}(h)\}$	$\hat{\rho}_P(h)$
0	0.054	0.024	0.067	1.0	1.0		1.0
1	0.041	0.014	0.053	0.76	0.79	0.209	0.047
2	0.030	0.015	0.041	0.56	0.62	0.224	0.021
3	0.038	0.015	0.050	0.71	0.74	0.224	0.055
4	0.023	0.015	0.033	0.42	0.50	0.224	0.033
5	0.025	0.015	0.036	0.47	0.54	0.224	0.026
6	0.020	0.015	0.030	0.37	0.45	0.224	0.025
7	0.046	0.014	0.057	0.85	0.85	0.209	0.072
8	0.024	0.015	0.033	0.44	0.50	0.224	0.035

Table 2.19 provides details of various autocovariance and autocorrelation estimates. Based on the estimates $\hat{\rho}_Z$, we would conclude that the autocorrelations are significant at lags 1, 2, 3 and 7 days. The autocorrelation estimates using either $\hat{\rho}_Z$ or $\hat{\rho}_{Z,UB}$ demonstrate the need for an autoregressive latent process with nonzero coefficients at lags 1, 2, 3 and 7. The autocorrelation function based on the Pearson residuals is completely misleading. If we use the calculation in (2.31), the expected value of $\hat{\rho}_P(1)$ is

$$E\{\hat{\rho}_P(1)\} \simeq \frac{1.934}{(0.054)^{-1} + 1.939}(0.76) = 0.0718,$$

which explains the small observed values of the $\hat{\rho}_P$ in Table 2.19.

3. ESTIMATION FOR PARAMETER DRIVEN MODELS

In this chapter, we compare several estimation methods for the parameter driven models. Suppose $\{\epsilon_t\}$ is the latent process in a parameter driven model, and

$$Y_t|\epsilon_t, \boldsymbol{x}_t \sim \text{Poisson}(\epsilon_t \exp\{\boldsymbol{x}_t^T\boldsymbol{\beta}\}) \text{ independently.} \tag{3.1}$$

where $\boldsymbol{x}_t$ is a $r \times 1$ regressor, $\boldsymbol{\beta}$ is a r-dimensional regression parameter, $\epsilon_t = \exp(\alpha_t)$ and $\{\alpha_t\}$ is an autoregressive process of order p denoted by AR(p). That is, $\{\alpha_t\}$ satisfies the recursions

$$\alpha_t = \phi_1\alpha_{t-1} + \cdots + \phi_p\alpha_{t-p} + z_t, \quad z_t \sim \text{IID}N(0, \sigma^2). \tag{3.2}$$

The likelihood of the complete data $(\boldsymbol{y}, \boldsymbol{\alpha}) = (y_1, \ldots, y_n; \alpha_1, \ldots, \alpha_n)$ is given by

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\alpha}) &= f(\boldsymbol{y}|\boldsymbol{\alpha})f(\boldsymbol{\alpha}) \\ &= [\Pi_{i=1}^n f(y_i|\alpha_i)]f(\boldsymbol{\alpha}) \\ &= [\Pi_{i=1}^n \frac{\exp\{-\exp(\alpha_i+\boldsymbol{x}_i^T\boldsymbol{\beta})\}\exp\{(\alpha_i+\boldsymbol{x}_i^T\boldsymbol{\beta})y_i\}}{y_i!}]\frac{\exp\{-\frac{1}{2}\boldsymbol{\alpha}^T V\boldsymbol{\alpha}\}}{(2\pi)^{n/2}|V^{-1}|^{1/2}} \\ &= \frac{|V|^{1/2}}{C_1}\exp\{-\textstyle\sum_{i=1}^n \exp(\alpha_i + \boldsymbol{x}_i^T\boldsymbol{\beta}) + \sum_{i=1}^n(\alpha_i + \boldsymbol{x}_i^T\boldsymbol{\beta})y_i - \frac{1}{2}\boldsymbol{\alpha}^T V\boldsymbol{\alpha}\} \\ &= \frac{|V|^{1/2}}{C_1}\exp\{-(e^{\boldsymbol{\alpha}})^T e^{X\boldsymbol{\beta}} + \boldsymbol{y}^T\boldsymbol{\alpha} + \boldsymbol{y}^T X\boldsymbol{\beta} - \tfrac{1}{2}\boldsymbol{\alpha}^T V\boldsymbol{\alpha}\}. \end{aligned} \tag{3.3}$$

where $C_1 = (2\pi)^{n/2}(\Pi_{i=1}^n y_i!)$, $e^{\boldsymbol{\alpha}} = (e^{\alpha_1}, \cdots, e^{\alpha_n})^T$, $X = (\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_n^T)^T$, $e^{X\boldsymbol{\beta}} = (e^{\boldsymbol{x}_1^T\boldsymbol{\beta}}, \cdots, e^{\boldsymbol{x}_n^T\boldsymbol{\beta}})^T$ and V^{-1} is the covariance matrix of $\boldsymbol{\alpha}$.

The objective is to estimate the model parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \sigma^2)^T$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_r)^T$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$. Since the likelihood of the observed data $f(\boldsymbol{y})$ is an n-fold integral which does not have a simple closed form, maximum likelihood estimation based on $f(\boldsymbol{y})$ is intractable. Likelihood-based estimation methods may require computationally intensive methods such as Markov chain Monte Carlo (Chan and Ledolter, 1995). We review the existing estimation methods in the following section, then propose a new method using an approximation

to the likelihood. A comparison of estimation methods are then examined based on fitting a real data set and simulation results.

3.1 Review of Existing Methods

Several methods have been developed for the estimation of the parameter driven models. Chan and Ledolter (1995) proposed Monte Carlo EM (MCEM) algorithm. The objective function is calculated by Monte Carlo. In the E-step, the conditional expectation of the log-likelihood of the complete data $(\boldsymbol{y}, \boldsymbol{\alpha})$ given the observed data $\boldsymbol{y}$ is estimated by averaging the conditional log-likelihood of simulated sets of complete data, i.e., they estimate

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(i)}) = E_{\boldsymbol{\psi}^{(i)}}[\log f(\boldsymbol{Y}, \boldsymbol{\alpha})|\boldsymbol{Y} = \boldsymbol{y}]$$

by

$$Q^{(m)}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(i)}) = \frac{1}{m}\sum_{j=1}^{m}\left[\log f(\boldsymbol{y}, \boldsymbol{\alpha}^{(j)})\right].$$

where $\boldsymbol{\psi}$ is the model parameters to be estimated, and $\boldsymbol{\alpha}^{(j)}, j = 1, \cdots, m$ are the Monte Carlo samples drawn from the conditional distribution of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ using the Gibbs Sampler. In the M-step, the conditional expectation $Q^{(m)}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(i)})$ is maximized with respect to $\boldsymbol{\psi}$ to obtain the updated estimate $\boldsymbol{\psi}^{(i+1)}$. The algorithm is then iterated until convergence. Each M-step usually requires iterations because when $\boldsymbol{\psi}$ changes, so does the distribution of $\boldsymbol{\alpha}|\boldsymbol{y}$. The Monte Carlo samples $\boldsymbol{\alpha}^{(j)}, j = 1, \cdots, m$ must be updated at each iteration of the maximization. This is a "many samples" method (Geyer, 1996) because it requires one sample $\boldsymbol{\alpha}^{(j)}, j = 1, \cdots, m$ per evaluation of the objective function $Q^{(m)}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(i)})$. The EM algorithm converges at a rather slow linear rate. A time consuming method like Gibbs sampling used at each iteration makes the convergence even slower.

As a viable alternative to the MCEM algorithm, the Monte Carlo implementation of the Newton-Raphson (MCNR) algorithm was suggested by Kuk and

Cheng (1997). It is computationally more efficient than the MCEM algorithm as it converges at a faster quadratic rate. Both algorithms require simulation from $\boldsymbol{\alpha}|\boldsymbol{y}$ with the aid of methods like Gibbs sampling and rejective sampling.

Let $l(\cdot)$ denote the log-likelihood, $l'(\cdot)$ and $l''(\cdot)$ denote the first and second derivatives of $l(\cdot)$ with respect to $\boldsymbol{\psi}$ respectively. The standard errors of the estimated parameters are calculated by inverting $-l''(\boldsymbol{y})$. The calculation of l' and l'' can be obtained from the following expressions (Louis, 1982):

$$\begin{aligned} l'(\boldsymbol{y}) &= E_{\boldsymbol{\psi}}\left[l'(\boldsymbol{Y},\boldsymbol{\alpha})|\boldsymbol{Y}=\boldsymbol{y}\right] \\ l''(\boldsymbol{y}) &= E_{\boldsymbol{\psi}}\left[l''(\boldsymbol{Y},\boldsymbol{\alpha})|\boldsymbol{Y}=\boldsymbol{y}\right] + E_{\boldsymbol{\psi}}\left[l'(\boldsymbol{Y},\boldsymbol{\alpha})(l'(\boldsymbol{Y},\boldsymbol{\alpha}))^T|\boldsymbol{Y}=\boldsymbol{y}\right] \\ &\quad -l'(\boldsymbol{y},\boldsymbol{\psi})(l'(\boldsymbol{y}))^T. \end{aligned}$$

For the case in that the above conditional expectations cannot be performed analytically, the Monte Carlo approximations of $l'(\boldsymbol{y})$ and $l''(\boldsymbol{y})$ are obtained.

3.1.1 Durbin and Koopman's Method

Durbin and Koopman (1997) estimated the model parameters based on a Monte Carlo estimation of the log-likelihood of the observed data $\boldsymbol{y}$. For a given set of model parameters $\boldsymbol{\psi}$, a normal density is used to approximate the non-Gaussian density of $\boldsymbol{y}$ given $\boldsymbol{\alpha}$, then Monte Carlo simulation and importance sampling are used to approximate the exact likelihood of $\boldsymbol{y}$, $f(\boldsymbol{y},\boldsymbol{\psi})$, and $f(\boldsymbol{y},\boldsymbol{\psi})$ is then maximized with respect to $\boldsymbol{\psi}$ numerically.

More specifically, let $g(\boldsymbol{\alpha}|\boldsymbol{y})$ be the approximate Gaussian density of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ for a given set of model parameters $\boldsymbol{\psi}$ and $g(\boldsymbol{y}|\boldsymbol{\alpha})$ is the approximating density to the Poisson density $f(\boldsymbol{y}|\boldsymbol{\alpha})$. The likelihood of $\boldsymbol{y}$ is given by

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{\psi}) &= \int f(\boldsymbol{y},\boldsymbol{\alpha})d\boldsymbol{\alpha} = \int f(\boldsymbol{y}|\boldsymbol{\alpha})f(\boldsymbol{\alpha})d\boldsymbol{\alpha} \\ &= \int \frac{g(\boldsymbol{y}|\boldsymbol{\alpha})f(\boldsymbol{\alpha})}{g(\boldsymbol{\alpha}|\boldsymbol{y})}\frac{f(\boldsymbol{y}|\boldsymbol{\alpha})}{g(\boldsymbol{y}|\boldsymbol{\alpha})}g(\boldsymbol{\alpha}|\boldsymbol{y})d\boldsymbol{\alpha} \\ &= L_g(\boldsymbol{\psi})\int \frac{f(\boldsymbol{y}|\boldsymbol{\alpha})}{g(\boldsymbol{y}|\boldsymbol{\alpha})}g(\boldsymbol{\alpha}|\boldsymbol{y})d\boldsymbol{\alpha} \\ &= L_g(\boldsymbol{\psi})E_g\left[\frac{f(\boldsymbol{y}|\boldsymbol{\alpha})}{g(\boldsymbol{y}|\boldsymbol{\alpha})}\right]. \end{aligned} \tag{3.4}$$

where E_g is the expectation with respect to Gaussian density $g(\boldsymbol{\alpha}|\boldsymbol{y})$, which is used as the importance density, and $L_g(\boldsymbol{\psi}) = \frac{g(\boldsymbol{y}|\boldsymbol{\alpha})f(\boldsymbol{\alpha})}{g(\boldsymbol{\alpha}|\boldsymbol{y})} = g(\boldsymbol{y})$ which is independent of the latent process $\boldsymbol{\alpha}$. The quantity $L_g(\boldsymbol{\psi})$ is the likelihood of $\boldsymbol{y}$ as if $\boldsymbol{y}$ is generated from the Gaussian state space model

$$\begin{aligned} y_t &= \alpha_t + \delta_t, \quad \delta_t \sim N(\mu_t, H_t) \\ \alpha_t &= \phi_1\alpha_{t-1} + \cdots + \phi_p\alpha_{t-p} + z_t, \quad z_t \sim N(0, \sigma^2). \end{aligned} \tag{3.5}$$

where

$$\begin{aligned} \mu_t &= y_t - \alpha_t^* - e^{-(\alpha_t^* + \boldsymbol{x}_t^T\boldsymbol{\beta})}(y_t - e^{(\alpha_t^* + \boldsymbol{x}_t^T\boldsymbol{\beta})}), \\ H_t &= \exp(-(\alpha_t^* + \boldsymbol{x}_t^T\boldsymbol{\beta})). \end{aligned}$$

and $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_n^*)^T$ is the mean of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ based on the above Gaussian state space model. The Kalman filter can be used to obtain $g(\boldsymbol{y})$. The normal density $g(\boldsymbol{y}|\boldsymbol{\alpha})$ is given by

$$g(\boldsymbol{y}|\boldsymbol{\alpha}) = \prod_{t=1}^{n} g(y_t|\alpha_t) \quad \text{with} \quad y_t|\alpha_t \sim N(\alpha_t + \mu_t, H_t). \tag{3.6}$$

The simulation smoother (De Jong and Shephard, 1995) is used to draw samples $\boldsymbol{\alpha}^{(j)}, j = 1, \cdots, N$ from $g(\boldsymbol{\alpha}|\boldsymbol{y})$. Then $f(\boldsymbol{y}, \boldsymbol{\psi})$ is approximated by

$$f(\boldsymbol{y}, \boldsymbol{\psi}) \approx \hat{f}(\boldsymbol{y}, \boldsymbol{\psi}) = L_g(\boldsymbol{\psi}) \left[\frac{1}{N} \sum_{j=1}^{N} \frac{f(\boldsymbol{y}|\boldsymbol{\alpha}^{(j)})}{g(\boldsymbol{y}|\boldsymbol{\alpha}^{(j)})} \right]. \tag{3.7}$$

Function $\hat{f}(\boldsymbol{y}, \boldsymbol{\psi})$ is maximized with respect to $\boldsymbol{\psi}$. This is also a "many samples" method because $g(\boldsymbol{\alpha}|\boldsymbol{y})$ changes as the model parameters $\boldsymbol{\psi}$ change. We have to simulate a fresh set of $\boldsymbol{\alpha}^{(j)}$ from new $g(\boldsymbol{\alpha}|\boldsymbol{y})$ at each iteration of the maximization of $\hat{f}(\boldsymbol{y}, \boldsymbol{\psi})$. This could make it difficult to find the optimum value to the locally rough character of the estimated likelihood. Moreover, this method is complicated and not easy to implement.

Let $\hat{\boldsymbol{\psi}}$ be the value that maximizes $\log \hat{f}(\boldsymbol{y}, \boldsymbol{\psi})$, and $\tilde{\boldsymbol{\psi}}$ be the value that would have been obtained by maximizing the true $\log f(\boldsymbol{y}, \boldsymbol{\psi})$ if this had been known. To compute standard errors due solely to simulation, Durbin and Koopman (1997) expanded $\partial \log \hat{f}(\boldsymbol{y}, \boldsymbol{\phi})/\partial\boldsymbol{\psi}$ about $\tilde{\boldsymbol{\psi}}$, put $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}$, and obtained

$$\hat{\boldsymbol{\psi}} - \tilde{\boldsymbol{\psi}} \approx -\left[\frac{\partial^2 \log \hat{f}(\boldsymbol{y}, \tilde{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} \right]^{-1} \frac{\partial \log \hat{f}(\boldsymbol{y}, \tilde{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}}.$$

So the mean square error matrix due to simulation is given by

$$E_g(\hat{\boldsymbol{\psi}} - \tilde{\boldsymbol{\psi}})^2 =: \mathrm{MSE}_g(\hat{\boldsymbol{\psi}}) \approx \hat{\Sigma} E_g \left[\frac{\partial \log \hat{f}(\boldsymbol{y}, \tilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} \frac{\partial \log \hat{f}(\boldsymbol{y}, \tilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}^T} \right] \hat{\Sigma},$$

where

$$\hat{\Sigma} = - \left[\frac{\partial^2 \log \hat{f}(\boldsymbol{y}, \hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \right]^{-1}.$$

The matrix $\hat{\Sigma}$ is an asymptotic estimate of the variance matrix of $\tilde{\boldsymbol{\psi}}$ under sampling variation of $\boldsymbol{y}$ ignoring simulation. It can be approximated numerically from neighboring values of $\boldsymbol{\psi}$. Define $\boldsymbol{q}^{(i)} = \partial \left[\frac{f(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)})}{g(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)})} \right] / \partial \boldsymbol{\psi}$ and $\bar{\boldsymbol{q}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{q}^{(i)}$, then $E_g \left[\frac{\partial \log \hat{f}(\boldsymbol{y}, \tilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}} \frac{\partial \log \hat{f}(\boldsymbol{y}, \tilde{\boldsymbol{\psi}})}{\partial \boldsymbol{\psi}^T} \right]$ can be estimated by

$$\frac{\sum_{i=1}^{N} \left[(\boldsymbol{q}^{(i)} - \bar{\boldsymbol{q}})(\boldsymbol{q}^{(i)} - \bar{\boldsymbol{q}})^T \right]}{\left[\sum_{i=1}^{N} \frac{f(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)})}{g(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)})} \right]^2}.$$

3.1.2 Kuk's Method

Let $L(\boldsymbol{\psi})$ be the likelihood of $\boldsymbol{y}$ and $L_w(\boldsymbol{\psi}_0)$ the likelihood of $\boldsymbol{y}$ under a working model. That is, $L_w(\boldsymbol{\psi}_0) = \int w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0) d\boldsymbol{\alpha}$, where $w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi})$ is the joint density of $(\boldsymbol{y}, \boldsymbol{\alpha})$ under the working model. Kuk (1997) derived a relative likelihood formula by using Geyer's (1994) results:

$$\begin{aligned} \frac{L(\boldsymbol{\psi})}{L_w(\boldsymbol{\psi}_0)} &= \frac{\int f(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi}) f(\boldsymbol{\alpha};\boldsymbol{\psi}) d\boldsymbol{\alpha}}{\int w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0) d\boldsymbol{\alpha}} \\ &= \int \frac{f(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi}) f(\boldsymbol{\alpha};\boldsymbol{\psi})}{w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0)} \left[\frac{w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0)}{\int w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0) d\boldsymbol{\alpha}} \right] d\boldsymbol{\alpha} \\ &= \int \frac{f(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi}) f(\boldsymbol{\alpha};\boldsymbol{\psi})}{w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0)} w(\boldsymbol{\alpha}|\boldsymbol{y}; \boldsymbol{\psi}_0) d\boldsymbol{\alpha} \\ &= E_w \left[\frac{f(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi}) f(\boldsymbol{\alpha};\boldsymbol{\psi})}{w(\boldsymbol{y}, \boldsymbol{\alpha}; \boldsymbol{\psi}_0)} | \boldsymbol{y}; \boldsymbol{\psi}_0 \right]. \end{aligned} \tag{3.8}$$

where the expectation is with respect to the conditional density $w(\boldsymbol{\alpha}|\boldsymbol{y}; \boldsymbol{\psi}_0)$ of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ under the working model at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, and $\boldsymbol{\psi}_0$ is a reference point of the relative likelihood. In order to approximate the relative likelihood $L(\boldsymbol{\psi})/L_w(\boldsymbol{\psi}_0)$, one takes a sample $\boldsymbol{\alpha}$ from distribution $w(\boldsymbol{\alpha}|\boldsymbol{y}; \boldsymbol{\psi}_0)$ (note that here $\boldsymbol{\psi}_0$ is fixed)

then evaluate the expectation E_w defined in (3.8). This is a "single sample" method because it uses the same sample $\boldsymbol{\alpha}$ from the distribution $w(\boldsymbol{\alpha}|\boldsymbol{y};\boldsymbol{\psi}_0)$ for all evaluations of the objective function $L(\boldsymbol{\psi})/L_w(\boldsymbol{\psi}_0)$ in the process of optimization.

In particular, if we use Durbin and Koopman's (1997) approximation - a normal density $g(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi})$ to approximate Poisson density $f(\boldsymbol{y}|\boldsymbol{\alpha},\boldsymbol{\psi})$, then we have working model $w(\boldsymbol{y},\boldsymbol{\alpha};\boldsymbol{\psi}) = g(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi})f(\boldsymbol{\alpha};\boldsymbol{\psi})$, where $g(\boldsymbol{y}|\boldsymbol{\alpha};\boldsymbol{\psi})$ is given by (3.6) and $f(\boldsymbol{\alpha};\boldsymbol{\psi})$ is an n-dimensional normal density. By drawing samples $\boldsymbol{\alpha}^{(i)}, i = 1,\cdots,N$ from the density $g(\boldsymbol{\alpha}|\boldsymbol{y};\boldsymbol{\psi}_0)$ for fixed $\boldsymbol{\psi}_0$, we will be able to approximate the relative likelihood as

$$\frac{L(\boldsymbol{\psi})}{L_w(\boldsymbol{\psi}_0)} \approx \frac{1}{N}\sum_{i=1}^{N}\left[\frac{f(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)};\boldsymbol{\psi})f(\boldsymbol{\alpha}^{(i)};\boldsymbol{\psi})}{g(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)};\boldsymbol{\psi}_0)f(\boldsymbol{\alpha}^{(i)};\boldsymbol{\psi}_0)}\right]. \tag{3.9}$$

Geyer (1996) suggested updating $\boldsymbol{\psi}_0$ to the maximizer of the relative likelihood and repeating the Monte Carlo approximation and maximization using the new $\boldsymbol{\psi}_0$. By updating $\boldsymbol{\psi}_0$ a few times, one should get better approximations of the relative likelihood function near the true maximum likelihood estimate.

Geyer (1996) provided the asymptotic distribution of $\hat{\boldsymbol{\psi}}$, the maximizer of the relative likelihood $\log\frac{L(\boldsymbol{\psi})}{L_w(\boldsymbol{\psi}_0)} := l_N(\boldsymbol{\psi})$. The variance of $\hat{\boldsymbol{\psi}}$ is approximately $B^{-1}(\hat{\boldsymbol{\psi}})A(\hat{\boldsymbol{\psi}})B^{-1}(\hat{\boldsymbol{\psi}})$, where a consistent estimator of $B(\hat{\boldsymbol{\psi}})$ is $-\frac{\partial^2 l_N(\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}$, and $A(\hat{\boldsymbol{\psi}}) = \mathrm{Var}[\partial l_N(\hat{\boldsymbol{\psi}})/\partial\boldsymbol{\psi}]$. $A(\hat{\boldsymbol{\psi}})$ can be estimated by

$$\frac{\sum_{i=1}^{N}\boldsymbol{u}_i\boldsymbol{u}_i^T - N\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{u}_i\right)^T}{\left(\sum_{i=1}^{N}\frac{f(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})f(\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})}{g(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi}_0)f(\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi}_0)}\right)^2}.$$

where $\boldsymbol{u}_i = \left[\frac{\partial[f(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})f(\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi})]}{\partial\boldsymbol{\psi}}\right]/[g(\boldsymbol{y}|\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi}_0)f(\boldsymbol{\alpha}^{(i)},\boldsymbol{\psi}_0)]$. Here the quantity $B^{-1}(\hat{\boldsymbol{\psi}})A(\hat{\boldsymbol{\psi}})B^{-1}(\hat{\boldsymbol{\psi}})$ is the estimated variance of the difference between the Monte Carlo approximations to the MLE $\hat{\boldsymbol{\psi}}$ and the exact MLE which we are unable to calculate. It is also called simulation error. The estimate of difference between the exact MLE and the true parameter value of $\boldsymbol{\psi}$ is an entirely different problem.

The variance matrix of this difference is the inverse Fisher information which is estimated by $B^{-1}(\hat{\boldsymbol{\psi}})$. It is the error due to the variation of the observed data $\boldsymbol{y}$. The variance formulas provided by Geyer (1996) are the same as those given in Durbin and Koopman (1997).

Both Durbin and Koopman's, and Kuk's estimation methods are based on Monte Carlo approximation and importance sampling. The implementation of the algorithms are complicated. In the following section, we take a different approach to the approximation of the likelihood of observed data.

3.2 Approximation to the Likelihood of Observed Data $\boldsymbol{y}$

We approximate the likelihood of the complete data $(\boldsymbol{y}, \boldsymbol{\alpha})$ so that $\boldsymbol{\alpha}$ can easily be integrated. We then compute the marginal distribution of $\boldsymbol{y}$ and maximize it to get the estimates of the model parameters.

3.2.1 Approximate Likelihood

A Taylor expansion at $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_n^*)^T$ on the term $(e^{\boldsymbol{\alpha}})^T e^{X\boldsymbol{\beta}}$ in equation (3.3) gives

$$(e^{\boldsymbol{\alpha}})^T e^{X\boldsymbol{\beta}} = \boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}} + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T K \boldsymbol{b}^* + \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T BK(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \quad (3.10)$$

where $\boldsymbol{b}^* = (e^{\alpha_1^*}, \ldots, e^{\alpha_n^*})^T$, B is the diagonal matrix $\mathrm{diag}(e^{\alpha_1^*}, \ldots, e^{\alpha_n^*})$, and K is the diagonal matrix $\mathrm{diag}(e^{\boldsymbol{x}_1^T\boldsymbol{\beta}}, \ldots, e^{\boldsymbol{x}_n^T\boldsymbol{\beta}})$. Let

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - K\boldsymbol{b}^* + BK\boldsymbol{\alpha}^*. \quad (3.11)$$

and use equation (3.3), the approximate likelihood of the complete data $(\boldsymbol{y}, \boldsymbol{\alpha})$ is then given by

$$\begin{aligned} f_a(\boldsymbol{y}, \boldsymbol{\alpha}) &= \frac{|V|^{1/2}}{C_1} \exp\{\boldsymbol{y}^T\boldsymbol{\alpha} + \boldsymbol{y}^T X\boldsymbol{\beta} - \tfrac{1}{2}\boldsymbol{\alpha}^T V\boldsymbol{\alpha} \\ &\quad -[\boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}} + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T K\boldsymbol{b}^* + \tfrac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T BK(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)]\} \\ &= \frac{|V|^{1/2}}{C_1} \exp\{(\boldsymbol{y} - K\boldsymbol{b}^* + BK\boldsymbol{\alpha}^*)^T\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^T(BK + V)\boldsymbol{\alpha} \\ &\quad -\tfrac{1}{2}\boldsymbol{\alpha}^{*T} BK\boldsymbol{\alpha}^* + \boldsymbol{\alpha}^{*T} K\boldsymbol{b}^* + \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}}\} \\ &= \frac{|V|^{1/2}}{C_1} \exp\{\tilde{\boldsymbol{y}}^T\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^T(BK + V)\boldsymbol{\alpha} \\ &\quad -\tfrac{1}{2}\boldsymbol{\alpha}^{*T} BK\boldsymbol{\alpha}^* + \boldsymbol{\alpha}^{*T} K\boldsymbol{b}^* + \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}}\} \\ &= \frac{|V|^{1/2}}{C_1} \exp\{-\tfrac{1}{2}[\boldsymbol{\alpha} - (BK + V)^{-1}\tilde{\boldsymbol{y}}]^T (BK + V)[\boldsymbol{\alpha} - (BK + V)^{-1}\tilde{\boldsymbol{y}}] \\ &\quad +\tfrac{1}{2}\tilde{\boldsymbol{y}}^T(BK + V)^{-1}\tilde{\boldsymbol{y}} - \tfrac{1}{2}\boldsymbol{\alpha}^{*T} BK\boldsymbol{\alpha}^* + \boldsymbol{\alpha}^{*T} K\boldsymbol{b}^* + \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}}\}. \end{aligned}$$

So the conditional distribution of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ is

$$\boldsymbol{\alpha}|\boldsymbol{y} \sim N\left((BK + V)^{-1}\tilde{\boldsymbol{y}}, (BK + V)^{-1}\right). \tag{3.12}$$

and the approximate distribution of $\boldsymbol{y}$ is

$$\begin{aligned} f_a(\boldsymbol{y}) &= \frac{|V|^{1/2}}{|BK + V|^{1/2}(\prod_{i=1}^n y_i!)} \exp\{\tfrac{1}{2}\tilde{\boldsymbol{y}}^T(BK + V)^{-1}\tilde{\boldsymbol{y}} \\ &\quad -\tfrac{1}{2}\boldsymbol{\alpha}^{*T} BK\,\boldsymbol{\alpha}^* + \boldsymbol{\alpha}^{*T} K\boldsymbol{b}^* + \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}}\}. \end{aligned} \tag{3.13}$$

The calculation of $|BK + V|$ and $\tilde{\boldsymbol{y}}^T(BK + V)^{-1}\tilde{\boldsymbol{y}}$ can be done by using the innovations algorithm (Brockwell and Davis, 1991) to avoid inverting the $n \times n$ matrix $(BK + V)$ directly. Some calculation details are given in section 3.2.2. Since $E(\boldsymbol{\alpha}|\boldsymbol{y}) = (BK + V)^{-1}\tilde{\boldsymbol{y}}$, $\boldsymbol{\alpha}^*$ can be calculated as the converged value of $\boldsymbol{\alpha}^{(j)} = (\alpha_1^{(j)}, \cdots, \alpha_n^{(j)})^T$ which is obtained recursively from the equation

$$\boldsymbol{\alpha}^{(j+1)} = (B^{(j)}K + V)^{-1}(\boldsymbol{y} - K\boldsymbol{b}^{(j)} + B^{(j)}K\boldsymbol{\alpha}^{(j)}). \tag{3.14}$$

where $B^{(j)} = \text{diag}(e^{\alpha_1^{(j)}}, \cdots, e^{\alpha_n^{(j)}})$, and $\boldsymbol{b}^{(j)} = (e^{\alpha_1^{(j)}}, \cdots, e^{\alpha_n^{(j)}})^T$. The convergence of $\boldsymbol{\alpha}^{(j)}$ is usually reached within $5 \sim 7$ iterations for a given set of parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \sigma^2)^T$.

Based on approximate distribution of $\boldsymbol{y}$ given in equation (3.13), parameters of the model are then estimated as follows:

1. Select initial values of $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\phi} = \boldsymbol{\phi}^{(0)}$, $\sigma^2 = \sigma^{2(0)}$ and set $j = 0$;

2. For fixed $\boldsymbol{\alpha}^{(j)}, \boldsymbol{\phi}^{(j)}$ and $\sigma^{2(j)}$, maximize $(\boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{b}^{*T} e^{X\boldsymbol{\beta}})$ with respect to $\boldsymbol{\beta}$ to get $\boldsymbol{\beta}^{(j+1)}$; this is comparable to Poisson regression;

3. For fixed $\boldsymbol{\alpha}^{(j)}$ and $\boldsymbol{\beta}^{(j+1)}$, maximize $[\log \frac{|V|}{|BK+V|} + \tilde{\boldsymbol{y}}^T (BK+V)^{-1}\tilde{\boldsymbol{y}}]$ to find $\boldsymbol{\phi}^{(j+1)}$ and $\sigma^{2(j+1)}$, the estimates of $\boldsymbol{\phi}$ and σ^2 respectively;

4. For fixed $\boldsymbol{\beta}^{(j+1)}, \boldsymbol{\phi}^{(j+1)}$ and $\sigma^{2(j+1)}$, use equation (3.14) iteratively and take the converged value as $\boldsymbol{\alpha}^*$, and set $\boldsymbol{\alpha}^{(j+1)} = \boldsymbol{\alpha}^*$;

5. Increment j, go to step 2 and continue to convergence.

3.2.2 Some Calculation Details

To carry out the above calculations for parameter estimation, we need to compute $\tilde{\boldsymbol{y}}^T (BK+V)^{-1}\tilde{\boldsymbol{y}}$, $|BK+V|$ and $|V|$, where matrix V^{-1} is the covariance matrix of $(a_1, \cdots, a_n)$ from the latent process $\{a_t\}$, and $\{a_t\}$ is assumed to be the AR(p) process given in (3.2). Assuming $n - p > p$, we can write

$$\boldsymbol{Z} = A\boldsymbol{\alpha}. \tag{3.15}$$

where $\boldsymbol{Z} = (a_1, \cdots, a_p, z_{p+1}, \cdots, z_n)^T$, $\boldsymbol{\alpha} = (a_1, \cdots, a_p, a_{p+1}, \cdots, a_n)^T$, and A is a lower-triangular matrix with all diagonal elements 1, i.e.,

$$A = \begin{pmatrix} I_{p\times p} & 0_{p\times(n-p)} \\ \\ A_1 & A_2 \end{pmatrix}, \tag{3.16}$$

where

$$A_1 = \begin{pmatrix} -\phi_p & -\phi_{p-1} & \cdots & -\phi_1 \\ 0 & -\phi_p & \cdots & -\phi_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\phi_p \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}, \text{ and}$$

$$A_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ -\phi_1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \cdots & \vdots & \vdots \\ -\phi_p & -\phi_{p-1} & \ddots & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi_p & \ddots & -\phi_1 & 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & -\phi_p & -\phi_{p-1} & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & -\phi_p & \cdots & -\phi_1 & 1 \end{pmatrix}.$$

Using (3.15), we have

$$E(\boldsymbol{Z}\boldsymbol{Z}^T) = E(A\boldsymbol{a}\boldsymbol{a}^T A^T) = AE(\boldsymbol{a}\boldsymbol{a}^T)A^T = AV^{-1}A^T. \tag{3.17}$$

Since

$$E(\boldsymbol{Z}\boldsymbol{Z}^T) = E\left[(a_1, \cdots, a_p, z_{p+1}, \cdots, z_n)^T (a_1, \cdots, a_p, z_{p+1}, \cdots, z_n)\right]$$

$$= \begin{pmatrix} \Gamma_{p\times p} & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & \sigma^2 I_{n-p} \end{pmatrix} =: G, \tag{3.18}$$

where Γ is the covariance matrix of $(a_1, \cdots, a_p)$ that can be written as

$$\Gamma = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \cdots & \cdots & \cdots & \cdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{pmatrix},$$

with $\gamma(h) = \mathrm{Cov}(a_t, a_{t+h})$. From (3.17) and (3.18), we have

$$\begin{aligned} V &= A^T G^{-1} A = A^T \begin{pmatrix} \Gamma^{-1} & 0 \\ 0 & \sigma^{-2} I_{n-p} \end{pmatrix} A \\ &= \begin{pmatrix} I & A_1^T \\ 0 & A_2^T \end{pmatrix} \begin{pmatrix} \Gamma^{-1} & 0 \\ 0 & \sigma^{-2} I_{n-p} \end{pmatrix} \begin{pmatrix} I & 0 \\ A_1 & A_2 \end{pmatrix} \\ &= \begin{pmatrix} \Gamma^{-1} + \frac{1}{\sigma^2} A_1^T A_1 & \frac{1}{\sigma^2} A_1^T A_2 \\ \frac{1}{\sigma^2} A_2^T A_1 & \frac{1}{\sigma^2} A_2^T A_2 \end{pmatrix}. \end{aligned} \tag{3.19}$$

From the AR(p) recursion $a_t - \phi_1 a_{t-1} - \cdots - \phi_p a_{t-p} = z_t$, one can use the Yule-Walker equations to compute $\gamma(0), \gamma(1), \cdots, \gamma(p)$.

Now we are ready to calculate $\tilde{\boldsymbol{y}}^T(BK + V)^{-1}\tilde{\boldsymbol{y}}$. Suppose that $BK + V$ is the covariance matrix of $\tilde{\boldsymbol{y}}$. The best linear predictors of the components of $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \cdots, \tilde{y}_n)^T$ are given by

$$\hat{\tilde{y}}_{j+1} = \theta_{j1}(\tilde{y}_j - \hat{\tilde{y}}_j) + \theta_{j2}(\tilde{y}_{j-1} - \hat{\tilde{y}}_{j-1}) + \ldots + \theta_{jj}(\tilde{y}_1 - \hat{\tilde{y}}_1).$$

with $\hat{\tilde{y}}_1 = 0$. Thus, writing $\hat{\tilde{\boldsymbol{y}}} = (\hat{\tilde{y}}_1, \cdots, \hat{\tilde{y}}_n)^T$, we have

$$\begin{aligned} \tilde{\boldsymbol{y}} &= \tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}} + \hat{\tilde{\boldsymbol{y}}} \\ &= \tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}} + (C - I)(\tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}}) \\ &= C(\tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}}). \end{aligned}$$

where I is the $n \times n$ identity matrix, C is the $n \times n$ matrix given by

$$C = \begin{pmatrix} \theta_{00} & 0 & 0 & \cdots & 0 \\ \theta_{11} & \theta_{10} & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & \theta_{20} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & \theta_{n-1,0} \end{pmatrix}. \tag{3.20}$$

where $\theta_{i0} = 1$ for $i = 0, 1, \ldots, n-1$. Taking covariances, we have

$$\begin{aligned} BK + V &= E(\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}^T) \\ &= CDC^T. \end{aligned}$$

where $D = \mathrm{diag}(v_0, v_1, \ldots, v_{n-1})$, and v_i is the mean squared error of the one-step predictor of $\tilde{y}_i$. The determinant of $BK + V$ is given by

$$|BK + V| = |C||D||C^T| = \Pi_{i=0}^{n-1} v_i.$$

and

$$\begin{aligned} \tilde{\boldsymbol{y}}^T(BK + V)^{-1}\tilde{\boldsymbol{y}} &= \left[C(\tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}})\right]^T (CDC^T)^{-1} \left[C(\tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}})\right] \\ &= (\tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}})^T D^{-1} (\tilde{\boldsymbol{y}} - \hat{\tilde{\boldsymbol{y}}}) \\ &= \textstyle\sum_{i=1}^{n} \frac{(\tilde{y}_i - \hat{\tilde{y}}_i)^2}{v_{i-1}}. \end{aligned}$$

So the quantities $\tilde{\boldsymbol{y}}^T(BK + V)^{-1}\tilde{\boldsymbol{y}}$ and $|BK + V|$ can be obtained directly from the innovations algorithm.

The lower triangular matrix A defined in (3.16) has diagonal elements 1, so that

$$|V| = |A^T| \begin{vmatrix} \Gamma^{-1} & 0 \\ 0 & \Sigma^{-1} \end{vmatrix} |A| = |\Gamma^{-1}|\,|\Sigma^{-1}| = |\Gamma^{-1}|\sigma^{-2(n-p)}.$$

In the special case of $p = 1$, i.e., $\{\alpha_t\} \sim \mathrm{AR}(1)$, $|V| = \frac{1-\phi^2}{\sigma^{2n}}$, and

$$V = \sigma^{-2} \begin{pmatrix} 1 & -\phi & 0 & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & 0 & \cdots & 0 \\ 0 & -\phi & 1+\phi^2 & -\phi & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{pmatrix}.$$

3.2.3 Connections between the Approximate Likelihood and Durbin and Koopman's Importance Density

An approximate density of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ proposed in (3.12) is based on the Taylor expansion of a term in the density of $\boldsymbol{y}$ given $\boldsymbol{\alpha}$ which is a Poisson density in our model. Durbin and Koopman (1997) have a different approach to find their importance density $g(\boldsymbol{\alpha}|\boldsymbol{y})$, though they did not write down this conditional density explicitly. In fact, these two densities are exactly the same as we show below.

Durbin and Koopman (1997) approximated the density $f(y_t|\alpha_t)$ by a normal density $g(y_t|\alpha_t)$ and obtained an approximating model.

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t, \quad \epsilon_t \sim N(\mu_t, H_t) \text{ independently.} \\ \alpha_t &= \phi_1\alpha_{t-1} + \cdots + \phi_p\alpha_{t-p} + Z_t, \quad Z_t \sim IIDN(0, \sigma^2). \end{aligned} \tag{3.21}$$

In this case, $g(y_t|\alpha_t)$ is the density of a normally distributed random variable with mean $\mu_t + \alpha_t$ and variance H_t. They chose μ_t and H_t so that $\partial l_t(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = 0$ and $\partial^2 l_t(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T = 0$, where $l_t(\boldsymbol{\alpha}) = \log f(y_t|\alpha_t) - \log g(y_t|\alpha_t)$. They obtained

$$\begin{aligned} \mu_t &= y_t - \hat{\alpha}_t - y_t e^{-(\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta})} + 1 \text{ and} \\ H_t &= e^{-(\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta})}. \end{aligned} \tag{3.22}$$

Here $\hat{\alpha}_t$ is calculated recursively: start with initial values of μ_t and H_t; use Kalman smoothing to obtain $\hat{\alpha}_t^{(1)}$ in terms of $y_1, y_2, \cdots, y_n$; get a new set of μ_t and H_t; use Kalman smoothing to get an updated $\hat{\alpha}_t^{(2)}$; and continue until convergence of the $\hat{\alpha}_t^{(j)}$. The $\hat{\alpha}_t$ represents the converged value of the sequence $\hat{\alpha}_t^{(1)}, \hat{\alpha}_t^{(2)}, \cdots$.

The approximating density of $\boldsymbol{y}$ given $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} g(\boldsymbol{y}|\boldsymbol{\alpha}) &= \prod_{t=1}^n g(y_t|\alpha_t) = \prod_{t=1}^n \left[\frac{1}{\sqrt{2\pi}H_t^{1/2}} \exp\{-\tfrac{1}{2}(y_t - \mu_t - \alpha_t)^2 H_t^{-1}\}\right] \\ &= (2\pi)^{-\frac{n}{2}} \prod_{t=1}^n \{\exp(\tfrac{1}{2}(\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta})) \exp\left[-\tfrac{1}{2}e^{\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta}}(\hat{\alpha}_t + y_t e^{-(\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta})} - 1 - \alpha_t)^2\right]\} \\ &= (2\pi)^{-\frac{n}{2}} \exp\left[\tfrac{1}{2}\sum_{t=1}^n (\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta}) - \tfrac{1}{2}\sum_{t=1}^n e^{\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta}}(\hat{\alpha}_t + y_t e^{-(\hat{\alpha}_t + \boldsymbol{x}_t^T\boldsymbol{\beta})} - 1 - \alpha_t)^2\right]. \end{aligned}$$

We write it in matrix form as

$$\begin{aligned} g(\boldsymbol{y}|\boldsymbol{\alpha}) = (2\pi)^{-\frac{n}{2}} \exp\Big[&-\tfrac{1}{2}[\hat{\boldsymbol{\alpha}} + (B_3K)^{-1}\boldsymbol{y} - \mathbf{1} - \boldsymbol{\alpha}]^T B_3 K [\hat{\boldsymbol{\alpha}} + (B_3K)^{-1}\boldsymbol{y} - \mathbf{1} - \boldsymbol{\alpha}] \\ &+ \tfrac{1}{2}\mathbf{1}^T(\hat{\boldsymbol{\alpha}} + X\boldsymbol{\beta})\Big]. \end{aligned}$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_n)^T$, $B_{\hat{\alpha}} = \text{diag}(e^{\hat{\alpha}_1}, \cdots, e^{\hat{\alpha}_n})$ and $K = \text{diag}(e^{\boldsymbol{x}_1^T\boldsymbol{\beta}}, \cdots, e^{\boldsymbol{x}_n^T\boldsymbol{\beta}})$.

The approximating joint density of $\boldsymbol{y}$ and $\boldsymbol{\alpha}$ is

$$\begin{aligned}
g(\boldsymbol{y}, \boldsymbol{\alpha}) &= g(\boldsymbol{y}|\boldsymbol{\alpha})g(\boldsymbol{\alpha}) \\
&= (2\pi)^{-\frac{n}{2}} \exp\Big[-\tfrac{1}{2}[\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1} - \boldsymbol{\alpha}]^T B_{\hat{\alpha}} K [\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1} - \boldsymbol{\alpha}] \\
&\quad + \tfrac{1}{2}\mathbf{1}^T(\hat{\boldsymbol{\alpha}} + X\boldsymbol{\beta})\Big] (2\pi)^{-\frac{n}{2}} |V|^{\frac{1}{2}} \exp(-\tfrac{1}{2}\boldsymbol{\alpha}^T V \boldsymbol{\alpha}) \\
&= (2\pi)^{-n} |V|^{\frac{1}{2}} \exp\Big[-\tfrac{1}{2}[\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1} - \boldsymbol{\alpha}]^T B_{\hat{\alpha}} K [\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1} - \boldsymbol{\alpha}] \\
&\quad + \tfrac{1}{2}\mathbf{1}^T(\hat{\boldsymbol{\alpha}} + X\boldsymbol{\beta}) - \tfrac{1}{2}\boldsymbol{\alpha}^T V \boldsymbol{\alpha}\Big] \\
&= (2\pi)^{-n} |V|^{\frac{1}{2}} \exp\{-\tfrac{1}{2}\boldsymbol{\alpha}^T(B_{\hat{\alpha}}K + V)\boldsymbol{\alpha} + (B_{\hat{\alpha}}K\hat{\boldsymbol{\alpha}} + \boldsymbol{y} - K\boldsymbol{b}_{\hat{\alpha}})^T\boldsymbol{\alpha} \\
&\quad - \tfrac{1}{2}[\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1}]^T B_{\hat{\alpha}} K [\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1}] + \tfrac{1}{2}\mathbf{1}^T(\hat{\boldsymbol{\alpha}} + X\boldsymbol{\beta})\} \\
&= (2\pi)^{-n} |V|^{\frac{1}{2}} \exp\{-\tfrac{1}{2}[\boldsymbol{\alpha} - (B_{\hat{\alpha}}K + V)^{-1}\tilde{\boldsymbol{y}}_{\hat{\alpha}}]^T (B_{\hat{\alpha}}K + V)[\boldsymbol{\alpha} - (B_{\hat{\alpha}}K + V)^{-1}\tilde{\boldsymbol{y}}_{\hat{\alpha}}] \\
&\quad + \tfrac{1}{2}\tilde{\boldsymbol{y}}_{\hat{\alpha}}^T(B_{\hat{\alpha}}K + V)^{-1}\tilde{\boldsymbol{y}}_{\hat{\alpha}} - \tfrac{1}{2}[\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1}]^T B_{\hat{\alpha}} K [\hat{\boldsymbol{\alpha}} + (B_{\hat{\alpha}}K)^{-1}\boldsymbol{y} - \mathbf{1}] \\
&\quad + \tfrac{1}{2}\mathbf{1}^T(\hat{\boldsymbol{\alpha}} + X\boldsymbol{\beta})\},
\end{aligned}$$

where V^{-1} is the covariance matrix of $\alpha_1, \cdots, \alpha_n$, $\boldsymbol{b}_{\hat{\alpha}} = (e^{\hat{\alpha}_1}, \cdots, e^{\hat{\alpha}_n})^T$ and

$$\tilde{\boldsymbol{y}}_{\hat{\alpha}} = \boldsymbol{y} - K\boldsymbol{b}_{\hat{\alpha}} + B_{\hat{\alpha}}K\hat{\boldsymbol{\alpha}}. \tag{3.23}$$

So the approximating conditional density of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ is

$$\boldsymbol{\alpha}|\boldsymbol{y} \sim N((B_{\hat{\alpha}}K + V)^{-1}\tilde{\boldsymbol{y}}_{\hat{\alpha}}, (B_{\hat{\alpha}}K + V)^{-1}). \tag{3.24}$$

From (3.12), (3.14) and (3.24), we know that $\boldsymbol{\alpha}^*$ and $\hat{\boldsymbol{\alpha}}$ are identical. They are the mean of Gaussian density of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$. So the approximate density of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ obtained in Section 3.2.1 is the same as the importance density derived from Durbin and Koopman's method.

The key step in Durbin and Koopman's approach is the construction of the quasi-observation $(\boldsymbol{y} - \boldsymbol{\mu})$ as they try to make $g(\boldsymbol{y}|\boldsymbol{\alpha})$ close to $f(\boldsymbol{y}|\boldsymbol{\alpha})$ as a function of $\boldsymbol{\alpha}$ with $\boldsymbol{y}$ fixed at the observed value. Schall (1991) proposed an algorithm for

estimating parameters of the generalized linear models with random effects. The link function $g(.)$ applied to the observed data $\boldsymbol{y}$ is linearized, giving to the first order

$$g(\boldsymbol{y}) = g(\boldsymbol{\mu}) + (\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) = \boldsymbol{z},$$

where $\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{e})$, and $\boldsymbol{e}$ is a vector of random errors. A linear random effects model is then derived for $\boldsymbol{z}$. The original observations $\boldsymbol{y}$ are replaced by "adjusted dependent variable" $\boldsymbol{z}$ and parameter estimation is based on the model for $\boldsymbol{z}$ instead of $\boldsymbol{y}$. Actually, as Kuk (1997) pointed out, Durbin and Koopman's quasi-observation $(\boldsymbol{y} - \boldsymbol{\mu})$ which can be derived from (3.22) is the same as Schall's adjusted dependent variable $\boldsymbol{z}$ evaluated at $\boldsymbol{\mu} = e^{X\boldsymbol{\beta}+\boldsymbol{\alpha}}$.

Let $\boldsymbol{\beta}$ represent the fixed effects, and $\boldsymbol{b}$ the random effects. Schall's (1991) objective function for his model is the posterior likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ under a normal prior of $\boldsymbol{b}$, which is proportional to $f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})g(\boldsymbol{b})$. $f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{b})$ is, by assumption, in the exponential family. The posterior log likelihood is equivalent to $\log g(\boldsymbol{z}|\boldsymbol{\beta},\boldsymbol{b}) + \log g(\boldsymbol{b})$, where g(.) is a normal density. Schall's model deals with random effects that are observable. The situation with our model is different. In our model, $(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha})$ does not have a normal density. Our latent process $\{\alpha_t\}$ which is comparable to the random effects $\boldsymbol{b}$ in Schall's model is unobservable and correlated. That makes our model parameter estimation much more complicated to implement.

3.3 Comparison of the Estimation Methods

From the introduction in Sections 3.1 and 3.2, we know that Durbin and Koopman's (DK's) method is computationally intensive requiring Monte Carlo approximation and iteration in order to maximize the objective function (log likelihood). Kuk's one-sample method is conceptually simpler, but has an important issue: how to find a "good" ψ_0 to start the maximization procedure? With the

setup of our parameter driven model, this initial reference point ψ_0 will definitely have impact on the final estimates of the model parameters. Our approximate likelihood method is free of simulation and relatively easy to implement, while the objective function is not an exact likelihood function. Moreover, there is no closed form formula to calculate the variance of the estimates.

In this section, we compare the relative performance of the above three estimation methods. A real data set (the polio data) and simulation results under different models are used. All the source code was written in Fortran 90. User times of computation are obtained by using Unix command "timex" on an IBM RS/6000 computer. It is the time spent in the system (CPU time) when the computation command is complete.

3.3.1 Polio Data Results

The following model is used to fit the polio data.

$$Y_t|\alpha_t, \boldsymbol{x}_t \sim \text{Poisson}(\exp\{\boldsymbol{x}_t^T\boldsymbol{\beta} + \alpha_t\}) \text{ independently,}$$

where $\{\alpha_t\}$ is an AR(1) process satisfying the recursions

$$\alpha_t = \phi_1\alpha_{t-1} + z_t, \quad z_t \sim \text{IID}N(0, \sigma^2),$$

and $\boldsymbol{x}_t^T = (1, \frac{t}{1000}, \cos(\frac{2\pi t}{12}), \sin(\frac{2\pi t}{12}), \cos(\frac{2\pi t}{6}), \cos(\frac{2\pi t}{6}))$.

Table 3.1 gives the results of the fitted model using Durbin and Koopman's (DK), Kuk's, and our approximation methods. The initial values of $\boldsymbol{\beta}$ are obtained by fitting a Poisson regression model to the data, assuming no temporal dependence in the data. The initial values for ϕ and σ^2 are 0 and 1 respectively. Results from Chan and Ledolter's MCEM, and Kuk and Cheng's MCNR algorithms are also listed for reference. The Monte Carlo sample size is $N = 1000$ for both DK's and Kuk's methods.

Table 3.1. Parameter estimates for fitting the parameter driven model to the polio data using different estimation methods. The figures within parentheses are standard errors, and numbers within curly brackets are standard errors due to Monte Carlo simulation.

Estimation Method	DK's	Kuk's	Our method	MCEM	MCNR
Intercept (log rate in Jan. 1976)	-.032 (.167) {.106}	.043 (.096) {.013}	.387 (.267)	.211 (.125)	.243 (.278)
trend $\times 10^{-3}$	-3.78 (2.86) {.179}	-5.51 (1.53) {.175}	-4.01 (2.79)	-4.62 (1.38)	-3.81 (2.83)
$\cos(2\pi t/12)$	-.101 (.149) {.013}	-.106 (.111) {.016}	.152 (.138)	.149 (.090)	.161 (.145)
$\sin(2\pi t/12)$	-.497 (.162) {.036}	-.502 (.123) {.010}	-.465 (.150)	-.495 (.116)	-.481 (.165)
$\cos(2\pi t/6)$	.198 (.128) {.019}	.173 (.112) {.014}	.403 (.118)	.439 (.102)	.413 (.127)
$\sin(2\pi t/6)$	-.363 (.126) {.021}	-.390 (.110) {.006}	-.008 (.119)	-.042 (.099)	-.011 (.125)
o	.662 (.118) {.021}	.711 (.057) {.010}	.663 (.205)	.894 (.036)	.661 (.218)
σ^2	.269 (.196) {.072}	.274 (.055) {.017}	.244 (.087)	.082	.272 (.627)
User time (seconds)	2034	3182†	1595††		

† The user time is based on iterating reference point ψ_0 6 times.

†† The user time includes both estimation time and time that computes the standard errors.

In our approximate likelihood method, the objective function is an approximation of the exact likelihood of $\boldsymbol{y}$. We are unable to calculate the Hessian matrix. The method we used to calculate the standard errors of our fitted parameters is as follows. We fit the model using our method to the polio data first, then use the fitted parameter values to generate 1000 sets of data of length 168. By fitting the

model to these 1000 sets of data, we are able to calculate the standard deviations of the estimated parameters. These standard deviations are listed in Table 3.1 as standard errors for our method.

All five estimation methods give comparable estimated parameter values. Especially, our method generates very similar estimates as MCNR does; DK and KUK's estimated parameters are very close except for the trend term with Kuk's estimates having smaller standard errors. MCEM algorithm has smallest standard errors.

The simulation standard errors are relatively small compared to the usual standard errors for either DK's or Kuk's method. For example, in the case of DK's method, the simulation standard error is 6.3% of the estimated standard error for the trend term, which implies that the simulation variance is only 0.4% of the usual variance; while Kuk's method gives variance ratio 1.3%. For the latent process parameters ϕ and σ^2, the ratios of the simulation variance to the usual variance are 3.2% and 13.5% respectively for DK's method; and 3.1% and 9.6% respectively for Kuk's method.

We did not run MCEM and MCNR methods on our computer so we cannot compare the computing time of these two methods. According to Kuk and Cheng (1997), MCNR procedure on the polio data is eight times faster than a modified version of the MCEM algorithm. Our approximate likelihood method is the fastest among the three methods. This is what we expected because there is no Monte Carlo simulation involved in the procedure. Somehow to our surprise, Kuk's method is not faster than DK's. Kuk's is a "single sample" method in which for all evaluations of the objective function $L(\boldsymbol{\psi})/L_u(\boldsymbol{\psi}_0)$, same sample $\boldsymbol{\alpha}$ from distribution $g(\boldsymbol{\alpha}|\boldsymbol{y};\boldsymbol{\psi}_0)$ is used. One would expect that this method saves some computing time over DK's "many samples" method in which a fresh set of $\boldsymbol{\alpha}$ is generated from new $g(\boldsymbol{\alpha}|\boldsymbol{y})$ at each iteration of the maximization of the likelihood

$f(\boldsymbol{y}, \boldsymbol{\psi})$. For the polio data, it takes 1.89 seconds of user time for Kuk's method to produce one evaluation of the objective function, while DK's method takes 8.40 seconds of user time. It is the iteration of the reference point $\boldsymbol{\psi}_0$ that slows down DK's procedures. We choose the GLM estimate of $\boldsymbol{\beta}$ and $\theta = 0$ and $\sigma^2 = 1$ as initial value of $\boldsymbol{\psi}_0$ in our calculation. Iteration time 6 for $\boldsymbol{\psi}_0$ is an arbitrarily chosen number. Obviously, how to find a good $\boldsymbol{\psi}_0$ and when to stop iterating $\boldsymbol{\psi}_0$ are important issues here. Geyer (1996) stated that it is not necessary to iterate $\boldsymbol{\psi}_0$ to convergence, as soon as new $\hat{\boldsymbol{\psi}}$ is reasonably close to $\boldsymbol{\psi}_0$, it is time to stop the iteration.

3.3.2 Simulation Results

We conducted a simulation study to compare the relative performance of three estimation methods: DK's, Kuk's and our approximate likelihood methods introduced in Sections 3.1 and 3.2. Simulation is based on the model

$$Y_t|\alpha_t \sim \text{Poisson}(e^{\boldsymbol{x}_t^T\boldsymbol{\beta}+\alpha_t}) \quad \text{independently}$$

with one of the following three model schemes. For each model and a set of given parameter values, 1000 sets of data of length n=100 are generated, then a corresponding model is fitted to each set of generated data using each of the three estimation methods. The average of each estimated parameter and the standard deviation are then obtained. The standard errors of parameters are also calculated using DK's and Kuk's methods. The initial values of the parameters in the estimation are the true values. In the cases of DK's and Kuk's methods, the Monte Carlo sampling size is $N = 1000$.

3.3.2.1 Model 1

Model 1 has a constant regression function and an AR(1) latent process, i.e., $\boldsymbol{x}_t \equiv 1$, and $\alpha_t = \phi\alpha_{t-1} + Z_t$, where $Z_t \sim IID N(0, \sigma^2)$. Constant regression

function assures that the observed data $\{Y_t\}$ is stationary. Figure 3.1 shows a sample path of $\{Y_t\}$ for this model with parameters $\beta = 0.7$, $\phi = 0.5$ and $\sigma^2 = 0.3$.

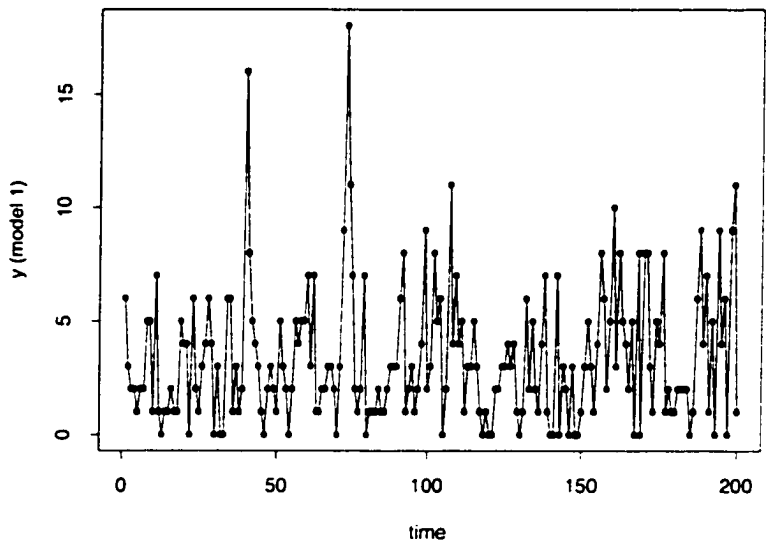

Figure 3.1: A sample path of model 1 with $\beta = 0.7$, $\phi = 0.5$ and $\sigma^2 = 0.3$.

Table 3.2 Parameter estimates of Model 1 with $\boldsymbol{x}_t \equiv 1$ and $\alpha_t = \phi\alpha_{t-1} + Z_t$ as latent process, using three estimation methods.

	DK's (n=100)	Kuk's (n=100)	Our method (n=100)	Our method (n=200)	True value of parameters
ave($\hat{\beta}$)	.707	.672	.818	.823	β=.7
sd($\hat{\beta}$)	.130	.142	.128	.096	
ave($\hat{se}(\hat{\beta})$)	.132	.093			
ave($\hat{\phi}$)	.438	.450	.456	.483	ϕ=.5
sd($\hat{\phi}$)	.193	.195	.203	.129	
ave($\hat{se}(\hat{\phi})$)	.170	.142			
ave($\hat{\sigma}^2$)	.294	.291	.257	.258	σ^2=.3
sd($\hat{\sigma}^2$)	.108	.110	.088	.066	
ave($\hat{se}(\hat{\sigma}^2)$)	.170	.108			
User time (seconds)	469260	363381	429	758	

Table 3.2 gives the parameter estimates of Model 1 using three different estimation methods. The average (ave) and standard deviations (sd) of the estimated parameters are calculated based on fitted models to 1000 sets of simulated data. The average of standard errors (ave(se)) of the parameters are based on 1000 replicates of inverse Hessian matrices.

Figures 3.2 and 3.3 show typical histograms and Q-Q plots of 1000 replicates of estimated parameters, respectively. It is evident that the distributions of $\hat{\beta}$ from all three estimation methods are normal.

From table 3.2 we know that the averages of $\hat{\beta}$ of DK's and Kuk's methods are close to the true value of $\beta = 0.7$. Hypothesis test $H_0 : E(\hat{\beta}) = 0.7$ vs. $H_a : E(\hat{\beta}) \neq 0.7$ can be used to detect bias for $\hat{\beta}$. The results of the test show that DK's $\hat{\beta}$ is not biased, but $\hat{\beta}$ from the other two methods are biased. $\hat{\beta}$ from our method has a bigger bias. For all three methods, $\hat{\phi}$ is estimated to have negative bias of about 10%, and the standard deviations of $\hat{\phi}$ are close among the three methods. Averages of standard errors are smaller than the corresponding standard deviation of $\hat{\phi}$ for both DK's and Kuk's methods.

The estimator of σ^2 from DK's and Kuk's methods are about 2% and 3% biased, respectively; $\hat{\sigma}^2$ in our method is approximately 14% negatively biased with a mean of 0.257 and interquartile range of 0.198 to 0.313 when the true value of σ^2 is 0.3. The average of the standard error of σ^2 from DK's method, $ave(\hat{se}(\hat{\sigma}^2))$, is much bigger than the standard deviation of the estimated σ^2, $sd(\hat{\sigma}^2)$, while these two values from Kuk's method are very close.

Table 3.2 also lists the results for $n = 200$ (length of time series in each replicate) using our approximate likelihood method. All the averages of the estimated parameters are greater than their counterparts in the case of $n = 100$. The bias of $\hat{\beta}$ is bigger, while biases of $\hat{\phi}$ and $\hat{\sigma}^2$ are slightly smaller.

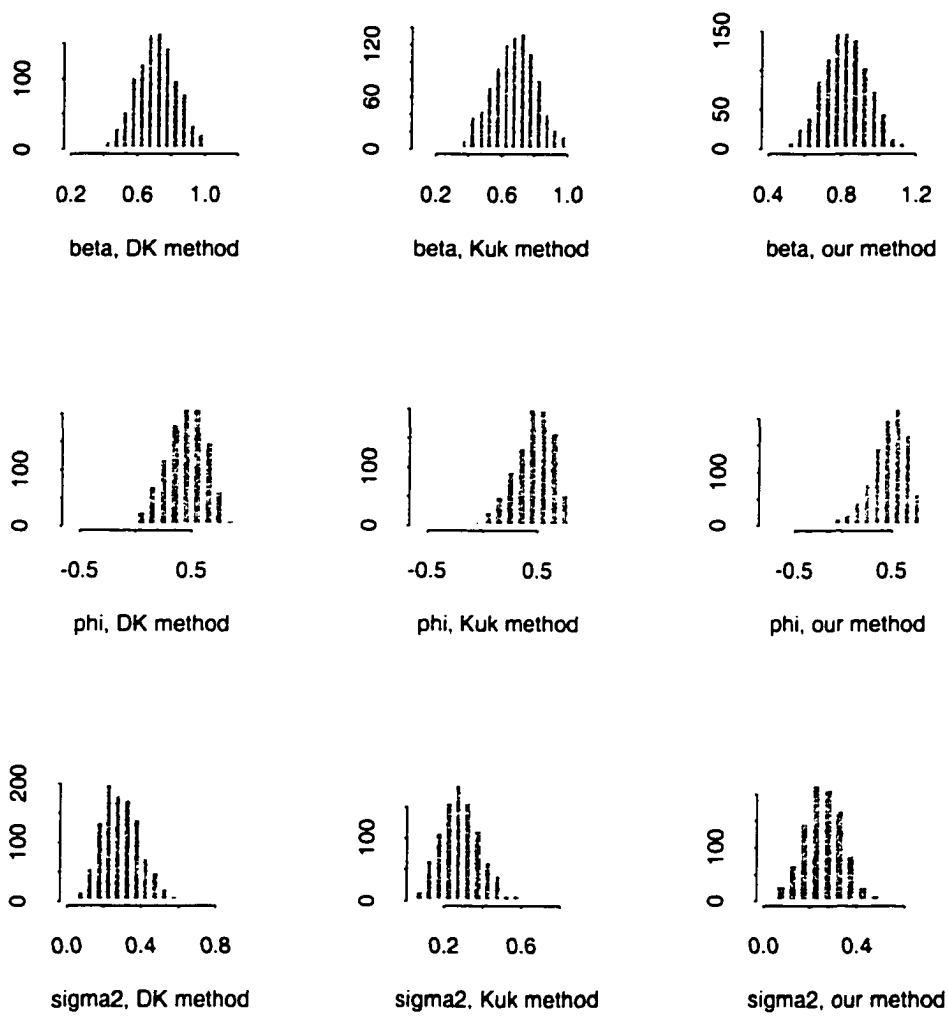

Figure 3.2: Histograms of the estimated parameters $\hat{\beta}$, $\hat{\phi}$, and $\hat{\sigma}^2$.

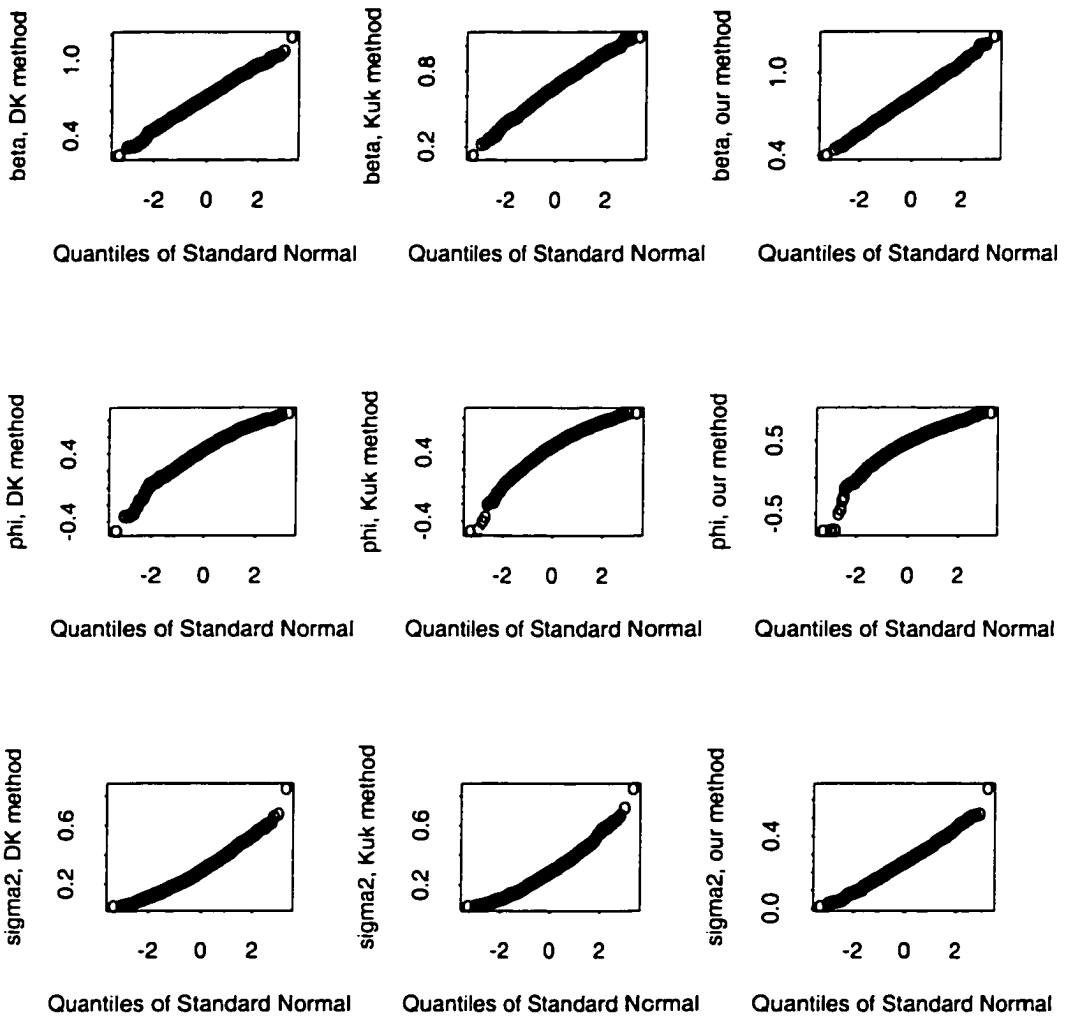

Figure 3.3: Q-Q plots of the estimated parameters $\hat{\beta}$, $\hat{\phi}$, and $\hat{\sigma}^2$.

Compared with DK's estimation method, Kuk's method is about 1.3 times faster, and our method is about 1000 times faster. Kuk's method is not as fast as we expected. The iteration of reference point ψ_0 slows down the computing of relative likelihood (the objective function).

3.3.2.2 Model 2

Second model in this simulation has a linear regression with 6 regressors and an AR(1) latent process, i.e., $\boldsymbol{x}_t = x_{t1}\beta_1 + \cdots + x_{t6}\beta_6$, and $\alpha_t = \phi\alpha_{t-1} + Z_t$, where $Z_t \sim IID\,N(0, \sigma^2)$. The observed $\{Y_t\}$ is not stationary. Figure 3.4 gives a sample path for this model. The parameter values used for generating the sample path are taken from the fitted model to the polio data using our estimation method (see Table 3.1). These values are also used as true parameter values in the simulation for this model and as initial values for the estimation.

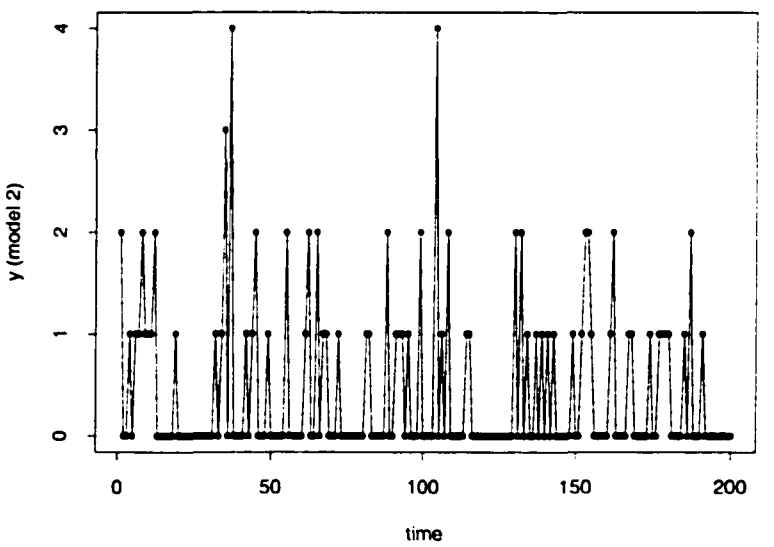

Figure 3.4: A sample path of Model 2 with $\phi = 0.663$, $\sigma^2 = 0.244$ and $\boldsymbol{\beta} = (0.387, -4.01, 0.152, -0.465, 0.403, -0.00758)^T$.

Table 3.3 summarizes the results from three estimation methods. The histograms and Q-Q plot of the estimated parameters reveal similar patterns as those in Model 1: $\hat{\beta}_1$ through $\hat{\beta}_6$ seem to be normally distributed, $\hat{\phi}$ is skewed to the left, and $\hat{\sigma}^2$ is skewed to the right.

The estimated parameters $\hat{\beta}_1$ - $\hat{\beta}_6$ from both DK's and Kuk's methods, and $\hat{\beta}_2$ - $\hat{\beta}_6$ from our method are not biased. For all three methods, $\hat{\phi}$ is estimated to have negative bias of about 19%, with largest standard deviation from our method. The magnitudes of bias of $\hat{\phi}$ are bigger than those of Model 1 (10%). Averages of standard errors of $\hat{\phi}$ are much smaller than standard deviation of $\hat{\phi}$ for both DK's and Kuk's methods.

The estimated parameter σ^2 from DK's, Kuk's, and our methods are about 8%, 15%, and 24% negatively biased, respectively. Again, the magnitudes of bias are bigger than those in Model 1. The average of standard error of σ^2 from DK's method is much bigger than the standard deviation of $\hat{\sigma}^2$, while these two values from Kuk's method are close. These findings are consistent with those from Model 1.

We also give results for $n = 200$ using our approximate likelihood method. Like $n = 100$ case, the estimated parameters $\hat{\beta}_2$ - $\hat{\beta}_6$ are not biased. But bias of $\hat{\beta}_1$ is bigger, and biases of $\hat{\phi}$ and $\hat{\sigma}^2$ are smaller, compared to the $n = 100$ case.

Comparisons of computing time of three estimation methods show that DK's method is about 1.8 times faster than Kuk's method which is contrary to what we found in Model 1. With Model 1, we have three unknown parameters in the estimation; with model 2, we have eight parameters. This increase on the number of unknown parameters has reduced the computing speed for Kuk's method. Our method is still the fastest among the three methods.

Table 3.3 Parameter estimates of Model 2 with $\boldsymbol{x}_t^T = (1, t/1000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))$ and $\alpha_t = \phi\alpha_{t-1} + Z_t$ as latent process, using three estimation methods.

	DK's (n=100)	Kuk's (n=100)	Our method (n=100)	Our method (n=200)	True value of parameters
ave($\hat{\beta}_1$)	.395	.365	.463	.477	β_1=.387
sd($\hat{\beta}_1$)	.327	.341	.346	.246	
ave($\hat{se}(\hat{\beta}_1)$)	.324	.242			
ave($\hat{\beta}_2$)	-4.15	-4.070	-3.852	-3.883	β_2= -4.01
sd($\hat{\beta}_2$)	5.624	5.830	5.802	2.129	
ave($\hat{se}(\hat{\beta}_2)$)	5.772	4.854			
ave($\hat{\beta}_3$)	.142	.138	.139	.147	β_3=.152
sd($\hat{\beta}_3$)	.180	.192	.172	.124	
ave($\hat{se}(\hat{\beta}_3)$)	.164	.156			
ave($\hat{\beta}_4$)	-.470	-.473	-.460	-.454	β_4= -.465
sd($\hat{\beta}_4$)	.193	.190	.191	.138	
ave($\hat{se}(\hat{\beta}_4)$)	.184	.171			
ave($\hat{\beta}_5$)	.404	.413	.395	.397	β_5= .403
sd($\hat{\beta}_5$)	.155	.158	.142	.110	
ave($\hat{se}(\hat{\beta}_5)$)	.151	.142			
ave($\hat{\beta}_6$)	.0072	-.0094	-.0077	-.0064	β_6=-.0076
sd($\hat{\beta}_6$)	.146	.150	.151	.108	
ave($\hat{se}(\hat{\beta}_6)$)	.147	.137			
ave($\hat{\phi}$)	.539	.540	.531	.618	ϕ=.663
sd($\hat{\phi}$)	.256	.274	.294	.175	
ave($\hat{se}(\hat{\phi})$)	.192	.124			
ave($\hat{\sigma}^2$)	.224	.207	.186	.199	σ^2= .244
sd($\hat{\sigma}^2$)	.112	.123	.101	.082	
ave($\hat{se}(\hat{\sigma}^2)$)	.235	.097			
User time (seconds)	1649451	3054986	1802	3157	

3.3.2.3 Model 3

Model 3 has a constant regressor and an AR(5) latent process, i.e., $\boldsymbol{x}_t \equiv 1$, and $\alpha_t = \phi_1\alpha_{t-1} + \cdots + \phi_5\alpha_{t-5} + Z_t$, where $Z_t \sim IIDN(0, \sigma^2)$. In this model, the observed $\{Y_t\}$ is stationary. Figure 3.5 shows a sample path for this model.

Estimation results using DK's and our methods are listed in Table 3.4. Kuk's method is excluded because in the process of maximization of log relative likeli-

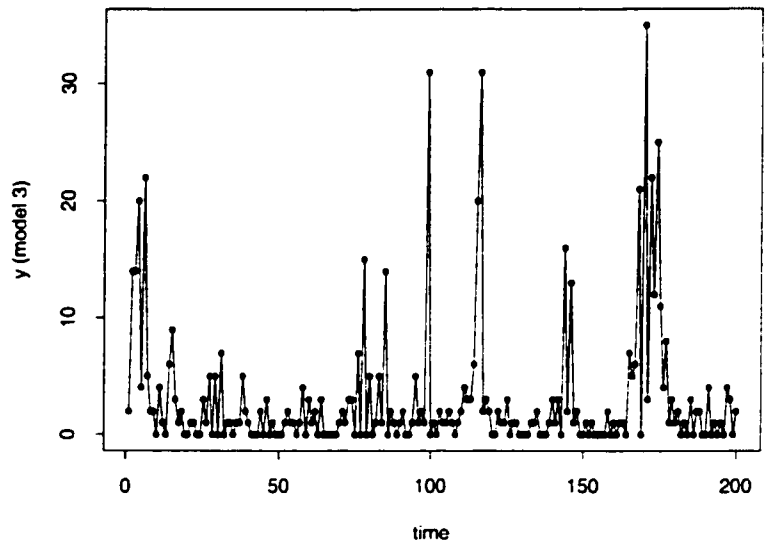

Figure 3.5: A sample path of Model 3 with parameters $\beta = 0.3$, $\phi_1 = 0.1$, $\phi_2 = 0.72$, $\phi_3 = -0.148$, $\phi_4 = -0.0944$, $\phi_5 = 0.0192$, and $\sigma^2 = 1$.

hood, the Hessian matrix produced from optimization algorithm is not positive definite for all the replications.

The Q-Q plots of estimated parameters from both methods are shown in Figures 3.6 and 3.7. For DK's method, the estimated parameters of the AR(5) process $\hat{\phi}_1$ - $\hat{\phi}_5$ seem to be normally distributed. $\hat{\phi}_1$, $\hat{\phi}_3$, $\hat{\phi}_4$ and $\hat{\phi}_5$ are not biased or slightly biased. $\hat{\beta}$ is about 24% positively biased while $\hat{\sigma}^2$ is about 26% negatively biased. The averages of standard errors is smaller than the standard deviation of the estimated parameter except $\hat{\sigma}^2$.

Our estimation method is about 1.5 times faster but produces biased estimates for all model parameters. The means of estimated parameter values are not close to the true parameter values.

Table 3.4 Parameter estimates of Model 3 with $x_t \equiv 1$, and $\alpha_t = \phi_1\alpha_{t-1} + \cdots + \phi_5\alpha_{t-5} + Z_t$ as latent process, using two estimation methods.

	DK's	Kuk's	Our method	True value of parameters
ave($\hat{\beta}$)	.373		1.643	β=.3
sd($\hat{\beta}$)	.338		1.485	
ave($\hat{se}(\hat{\beta})$)	.269			
ave($\hat{\phi_1}$)	.087		.0628	ϕ_1=.1
sd($\hat{\phi_1}$)	.197		.205	
ave($\hat{se}(\hat{\phi_1})$)	.169			
ave($\hat{\phi_2}$)	.651		-.103	ϕ_2=.72
sd($\hat{\phi_2}$)	.226		.200	
ave($\hat{se}(\hat{\phi_2})$)	.180			
ave($\hat{\phi_3}$)	-.126		.144	ϕ_3=-.148
sd($\hat{\phi_3}$)	.296		.239	
ave($\hat{se}(\hat{\phi_3})$)	.233			
ave($\hat{\phi_4}$)	-.123		.071	ϕ_4=-.094
sd($\hat{\phi_4}$)	.205		.228	
ave($\hat{se}(\hat{\phi_4})$)	.181			
ave($\hat{\phi_5}$)	.0052		.103	ϕ_5=.0192
sd($\hat{\phi_5}$)	.228		.212	
ave($\hat{se}(\hat{\phi_5})$)	.185			
ave($\hat{\sigma}^2$)	.738		1.629	σ^2=1.0
sd($\hat{\sigma}^2$)	.217		2.684	
ave($\hat{se}(\hat{\sigma}^2)$)	.250			
User time (seconds)	1100227		733770	

In summary, Durbin and Koopman's method produces relatively less biased estimates for model parameters of parameter driven models. Comparable estimated parameter values can be obtained from Kuk's method when the latent process is an AR(1) process. The computing time of Kuk's method is shorter than DK's when the number of unknown parameters is three. The advantage of computing speed of the "single sample" Kuk's method over the "many sample" DK's method vanished when the number of unknown parameters in the model increased to eight. Our approximate likelihood method produces unbiased estimates for the non-constant regression parameters when the latent process is AR(1). Es-

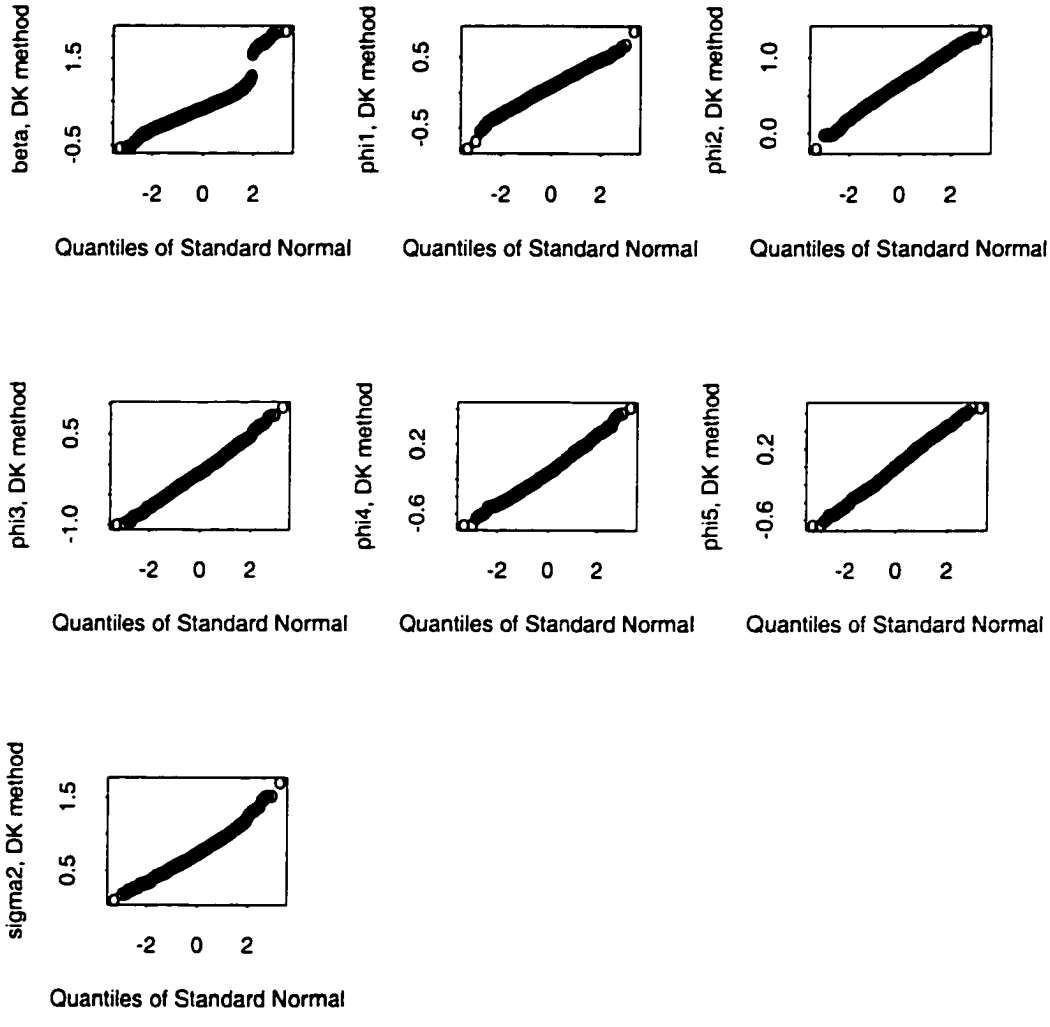

Figure 3.6: Q-Q plots of the estimated parameters of Model 3 using DK's method.

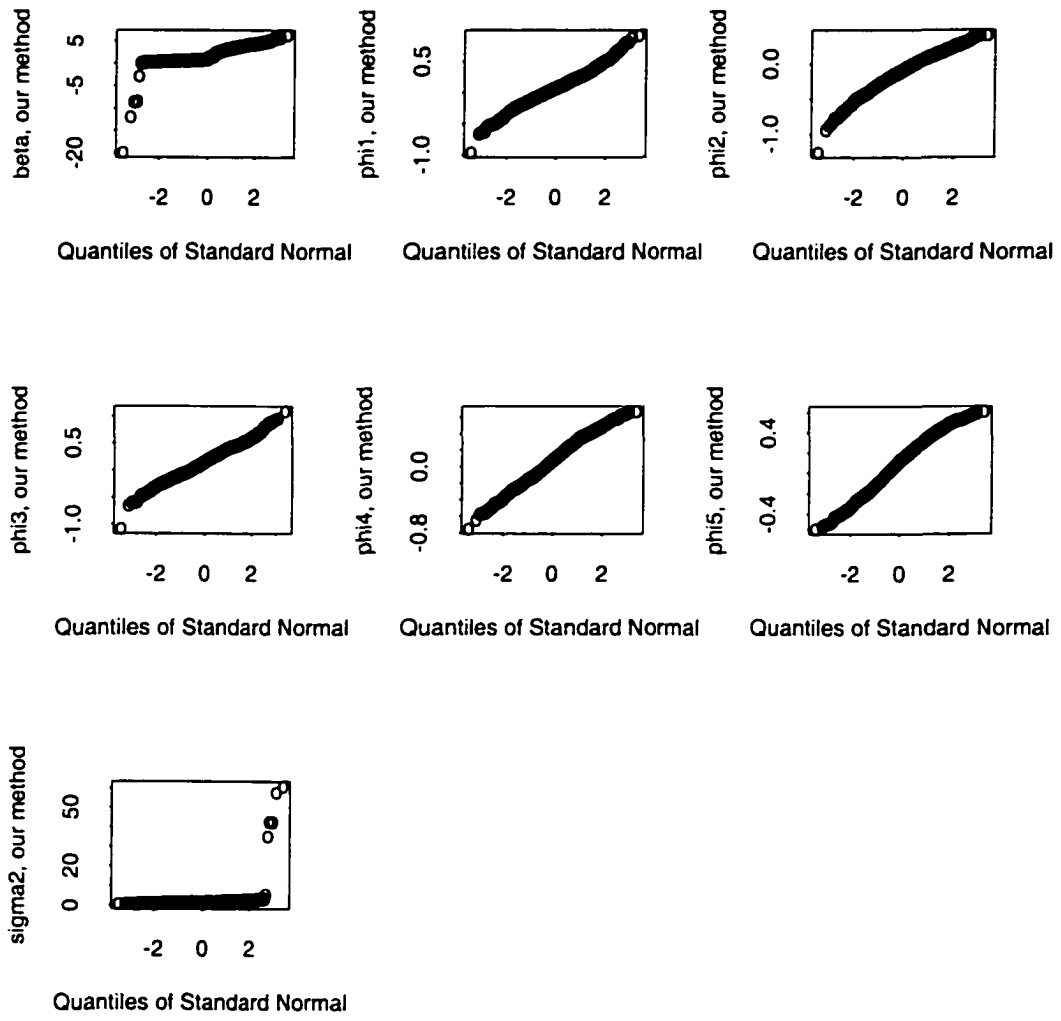

Figure 3.7: Q-Q plots of the estimated parameters of Model 3 using our approximate likelihood method.

timates for constant term in all three models are biased using our method. The computing time for AR(1) latent process model using our method is significantly shorter than the others.

4. SUMMARY AND CONCLUSIONS

In the previous chapters, we considered a class of parameter driven models in which a latent stochastic process is present in the mean of observed time series of counts $\{Y_t\}$. Much research effort has been devoted to fitting such a model. Existing techniques rely on the specification of a suitable model for the correlation structure in the latent process. As in linear regression with correlated errors, there is a need for model diagnostic and identification techniques. In this thesis, we focused on model identification, diagnosing the existence of a latent process, hypothesis testing for any evidence of autocorrelation in the latent process, and parameter estimation. In this chapter, we summarize our methods, approaches and important results we have obtained in this work.

Chapter 1 gave an introduction to generalized linear models and state-space models. Existing research related to the parameter driven models were also reviewed.

In chapter 2, for fitting the observed count data to our parameter driven model, we carried out an analogous model fitting procedure as for a linear model with time series errors. A consistent estimator of regression parameter $\boldsymbol{\beta}$ is needed first. A natural and easy way to compute consistent estimators of $\boldsymbol{\beta}$ is to use a standard generalized linear model analysis. We showed the consistency and asymptotic normality of the Poisson maximum likelihood estimator of $\boldsymbol{\beta}$ in a parameter driven model where the likelihood ignores the presence of serial correlations. A proof when the latent process is strongly mixing was provided. We

also derived simple formulas for the effect of autocovariance of latent process on standard errors of the regression coefficients.

The next step in the modeling process is to test for the existence of a latent process. Based on higher moment properties of Poisson distribution and Pearson residuals, we developed a test statistic Q and its modified version $\tilde{Q}$ and Q^*. The size and size adjusted power of four statistics $Q, \tilde{Q}, Q^*$ and S_a were compared for linear and cosine regression functions via simulation, where S_a is an existing test statistic. Based on our limited simulation, we have found that the test statistic S_a appeared to have similar type I error rates as $\tilde{Q}$ and Q^* but larger power. Further research is needed to demonstrate that S_a is preferred in all circumstances.

After the existence of a latent process has been established, we need to estimate the variance and autocovariances of the latent process. The existing Zeger's method-of-moment estimates of autocovariances, $\hat{\gamma}_{r,Z}$ have large negative bias which is directly attributable to the large bias in the estimate of variance σ_r^2. We proposed an "optimally" weighted estimates for the variance and autocovariances in the sense of minimum variance. We have found via simulation that under an IID latent process specification the optimal covariance estimates have smaller variances than Zeger estimates with approximately the same bias. We then developed bias corrections to both Zeger and the optimal estimates. The bias-adjusted estimates reduce bias but at the expense of higher variance. With increasing variance in the latent process, the bias increases in all estimates. Substantial bias caused by large autocorrelation presented in the latent process would impact the estimation of the correct asymptotic variance of the GLM estimates of $\boldsymbol{\beta}$. Overall, the optimal estimates outperform Zeger estimates and the bias-adjusted optimal estimates outperform the bias-adjusted Zeger estimates.

To test for zero autocorrelation in the latent process, we proposed a test statistic H^2 that is analogous to the Box-Jenkins's portmanteau statistic. We

considered four different test statistics H^2_Z, $H^2_{Z,UB}$, H^2_{Opt}, and $H^2_{Opt,UB}$ derived from H^2 by using Zeger, bias-adjusted Zeger, optimal, and bias-adjusted optimal estimates of autocorrelations, respectively. Under the null hypothesis of a white noise latent process, the Pearson residuals are heteroscedastic and nonstationary. We approximated the variance of the sample autocorrelation function of Pearson residuals and used it to obtain a modified Ljung-Box test statistic $H^2_{LB,M}$. A simulation comparison of the empirical size and power of above five test statistics along with standard Ljung-Box test H^2_{LB} was reported. We found that relative performance of these test statistics was highly dependent on the form of the regression functions. Throughout the simulation, results of H^2_{Opt} were similar to those of $H^2_{LB,M}$. The bias-adjusted statistics $H^2_{Z,UB}$ and $H^2_{Opt,UB}$ were uniformly more powerful than their unadjusted counterparts after the sizes of the tests had been calibrated to the nominal level. In the linear case, $H^2_{Opt,UB}$ and $H^2_{Z,UB}$ were superior than the others. In the cosine case, H^2_{LB} was more powerful than the other when sample size $n = 1000$; when $n = 100$, $H^2_{Opt,UB}$ was more powerful.

We applied our model identification procedures to daily asthma presentations at a hospital in Sydney. The data indicated presence of a latent process. Both tests $H^2_{Z,UB}$ and $H^2_{LB,M}$ rejected the hypothesis of zero autocorrelation in the latent process. The autocorrelations based on the Pearson residuals were very small and misleading. Both Zeger and bias-adjusted Zeger estimates of autocorrelations demonstrated the need for an autoregressive latent process with nonzero coefficients at lags 1,2,3 and 7.

In chapter 3, we focused on parameter estimation of the parameter driven model, in which the latent process is unobservable. The likelihood based on the observed data $\boldsymbol{y}$ is an n-fold integral that does not have a simple closed form. Existing likelihood-based estimation methods such as Durbin and Koopman's (DK) importance sampling and Kuk's Monte Carlo Newton Raphson require intense

Monte Carlo simulation in order to approximate the n-fold integral in the density function of $\boldsymbol{y}$.

We presented a new approach for estimating model parameters which avoids Monte Carlo simulation. A Taylor expansion was applied to a term in the joint density of the observed data $\boldsymbol{y}$ and the latent process $\boldsymbol{\alpha}$. As a result, an approximate likelihood of the complete data $(\boldsymbol{y}, \boldsymbol{\alpha})$ and the approximate conditional density of $\boldsymbol{\alpha}$ given $\boldsymbol{y}$ were obtained. Then the approximate distribution of $\boldsymbol{y}$ was derived and maximized to get estimates of model parameters.

Applications of our estimation method, DK's and Kuk's methods to time series of monthly polio counts in the U.S. were used to illustrate the methods. All three methods produced comparable parameter estimates. Kuk's method gave a significant trend term, but the trend was not significant using DK's and our approximate methods.

A simulation study was conducted to compare the relative performance of these three estimation methods. For a parameter driven model with a constant regression function and an AR(1) latent process, all estimates obtained from the three methods were biased with one exception. The estimate of β under DK's method was nearly unbiased. The magnitudes of bias were dependent on the estimation methods. All three methods had similar standard deviations for the estimated parameters. For a parameter driven model with trend and seasonal terms and an AR(1) latent process, all estimates of regression parameters were unbiased except for the constant term under our method. Estimated parameters of latent process ϕ and σ^2 were biased using all three methods. The magnitude of bias was bigger than that from the first model. Again, all three methods produced comparable standard deviations of the estimated parameters. The third model in our simulation had a constant regressor and an AR(5) latent process. DK's method gave unbiased estimates for the ϕ's. Overall, DK's method produced

relatively less biased estimates for model parameters. When the latent process was AR(1), our method gave reasonable estimates for non-constant regression parameters, and the computing time was considerably shorter than for the other methods.

5. APPENDIX

Appendix A

Poisson Regression Model $\log \mu_t = \beta_0 + \beta_1 t$

Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a vector consisting of independent observations from a Poisson distribution with mean $E(Y_t) = \mu_t$, $t = 1, \cdots, n$. Consider a regression model

$$\log(\mu_t) = \beta_0 + \beta_1 t. \tag{5.1}$$

We will show that if $\beta_1 < 0$, all the observations Y_n are zero beyond some large n, so there are no consistent estimators for β_0 and β_1; while if $\beta_1 > 0$, the weak consistency and asymptotic normality of MLEs for β_0 and β_1 hold.

Case 1. $\beta_1 < 0$

Let $A_n = \{\omega : Y_n(\omega) = 0\}$, then $\{A_n\}$ is an independent sequence of events, and $\{\omega : Y_n(\omega) = 0, \forall n \text{ large}\} = \cup_{k=1}^{\infty} \cap_{n=k}^{\infty} \{\omega : Y_n(\omega) = 0\} = \liminf_n A_n$. Then

$$\begin{aligned} \sum_{n=1}^{\infty} P(A_n^c) &= \sum_{n=1}^{\infty} P\{\omega : Y_n(\omega) > 0\} \\ &= \sum_{n=1}^{\infty} \left[1 - e^{-e^{\beta_0 + \beta_1 n}}\right] < \infty \end{aligned}$$

since $\lim_{n\to\infty} \frac{1-e^{-e^{\beta_0+\beta_1(n+1)}}}{1-e^{-e^{\beta_0+\beta_1 n}}} = e^{\beta_1} < 1$ by the ratio test for series of positive terms. Thus $P(\limsup_n A_n^c) = 0$ by the Borel-Cantelli lemma. This is equivalent to

$$P\{\omega : Y_n(\omega) = 0, \forall n \text{ large}\} = 1.$$

Case 2: $\beta_1 > 0$

Let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, the log-likelihood of $(y_1, \cdots, y_n)$ is

$$\begin{aligned} l(\boldsymbol{\beta}) &= \textstyle\sum_{t=1}^{n}(y_t \log \mu_t - \mu_t - \log(y_t!)) \\ &= \textstyle\sum_{t=1}^{n}[y_t\beta_0 + y_t\beta_1 t - e^{\beta_0+\beta_1 t} - \log(y_t!)] \end{aligned} \tag{5.2}$$

and the Fisher information is given by

$$F_n(\boldsymbol{\beta}) := \mathrm{E}(\frac{\partial l}{\partial \boldsymbol{\beta}}\frac{\partial l}{\partial \boldsymbol{\beta}^T}) = -\mathrm{E}(\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}) = e^{\beta_0}\begin{pmatrix} \sum_{t=1}^{n} e^{\beta_1 t} & \sum_{t=1}^{n} te^{\beta_1 t} \\ \sum_{t=1}^{n} te^{\beta_1 t} & \sum_{t=1}^{n} t^2 e^{\beta_1 t} \end{pmatrix}. \tag{5.3}$$

where

$$\begin{aligned} \sum_{t=1}^{n} e^{\beta_1 t} &= \frac{e^{\beta_1}(1-e^{n\beta_1})}{1-e^{\beta_1}}. \\ \sum_{t=1}^{n} te^{\beta_1 t} &= \frac{e^{\beta_1} - (n+1)e^{(n+1)\beta_1} + ne^{(n+2)\beta_1}}{(1-e^{\beta_1})^2}. \end{aligned}$$

and

$$\begin{aligned} \sum_{t=1}^{n} t^2 e^{\beta_1 t} &= \frac{-n^2 e^{(n+1)\beta_1} + 2\sum_{t=1}^{n} te^{\beta_1 t} - \sum_{t=1}^{n} e^{\beta_1 t}}{1-e^{\beta_1}} \\ &= \frac{e^{\beta_1} + e^{2\beta_1} - (n+1)^2 e^{(n+1)\beta_1} + (2n^2+2n-1)e^{(n+2)\beta_1} - n^2 e^{(n+3)\beta_1}}{(1-e^{\beta_1})^3}. \end{aligned}$$

1) We show that the minimum eigenvalue of the Fisher information diverges: $\lambda_{min} F_n(\boldsymbol{\beta}) \to \infty$.

Let $A = \sum_{t=1}^{n} e^{\beta_1 t}$, $B = \sum_{t=1}^{n} te^{\beta_1 t}$ and $C = \sum_{t=1}^{n} t^2 e^{\beta_1 t}$, then

$$\begin{aligned} AC - B^2 &= \frac{e^{2\beta_1}(e^{n\beta_1}-1)^2 - n^2 e^{(n+2)\beta_1}(e^{\beta_1}-1)^2}{(e^{\beta_1}-1)^4} \\ &= \frac{e^{2\beta_1}}{(e^{\beta_1}-1)^4}(e^{2n\beta_1}) + o(e^{2n\beta_1}). \\ A + C &= \frac{-2e^{\beta_1} + e^{2\beta_1} - e^{3\beta_1} + (n^2+2n+2)e^{(n+1)\beta_1} - (2n^2+2n+1)e^{(n+2)\beta_1} + (n^2+1)e^{(n+3)\beta_1}}{(e^{\beta_1}-1)^3} \\ &= \frac{-2e^{\beta_1} + e^{2\beta_1} - e^{3\beta_1} + e^{(n+1)\beta_1}\left(2 - e^{\beta_1} + e^{2\beta_1}\right) - 2ne^{(n+1)\beta_1}(e^{\beta_1}-1) + n^2 e^{(n+1)\beta_1}(e^{\beta_1}-1)^2}{(e^{\beta_1}-1)^3} \\ &= \frac{1}{(e^{\beta_1}-1)}\left(n^2 e^{(n+1)\beta_1}\right) + o\left(n^2 e^{(n+1)\beta_1}\right). \end{aligned}$$

So we have

$$\frac{AC - B^2}{(A+C)^2} \to 0 \quad \text{and} \quad \frac{AC - B^2}{A+C} \to \infty \quad \text{as} \quad n \to \infty.$$

Let $X = \frac{4(AC-B^2)}{(A+C)^2}$, then

$$\begin{aligned} \lambda_{min} F_n(\boldsymbol{\beta}) &= \frac{e^{\beta_0}}{2}\left[A + C - \sqrt{(A+C)^2 - 4(AC - B^2)}\right] \\ &= \frac{e^{\beta_0}}{2}(A+C)\left(1 - \sqrt{1-X}\right). \end{aligned}$$

A Taylor expansion at 0 gives $\sqrt{1-X} \approx 1 - \frac{1}{2}X$, and hence

$$\begin{aligned}\lambda_{min} F_n(\boldsymbol{\beta}) &\approx \frac{e^{\beta_0}}{2}(A+C)\left[1-(1-\frac{1}{2}X)\right] \\ &= e^{\beta_0}\left(\frac{AC-B^2}{A+C}\right) \to \infty, \quad \text{as } n \to \infty.\end{aligned}$$

2) $\boldsymbol{z}_n' F_n^{-1}(\boldsymbol{\beta})\boldsymbol{z}_n \to 0$ as $n \to \infty$, where $\boldsymbol{z}_n' = (1, n)$.

Since

$$F_n^{-1}(\boldsymbol{\beta}) = \frac{e^{\beta_0}}{C_1}\begin{pmatrix} \sum_{t=1}^n t^2 e^{\beta_1 t} & -\sum_{t=1}^n t e^{\beta_1 t} \\ -\sum_{t=1}^n t e^{\beta_1 t} & \sum_{t=1}^n e^{\beta_1 t} \end{pmatrix}.$$

where

$$\begin{aligned}C_1 &= (\sum_{t=1}^n e^{\beta_0+\beta_1 t})(\sum_{t=1}^n t^2 e^{\beta_0+\beta_1 t}) - (\sum_{t=1}^n t e^{\beta_0+\beta_1 t})^2 \\ &= e^{2\beta_0}\frac{e^{3\beta_1} - n^2 e^{(n+2)\beta_1} + (2n^2-2)e^{(n+3)\beta_1} - n^2 e^{(n+4)\beta_1} + e^{(2n+3)\beta_1}}{(1-e^{\beta_1})^4}.\end{aligned}$$

$$\begin{aligned}&\boldsymbol{z}_n' F_n^{-1}(\boldsymbol{\beta})\boldsymbol{z}_n \\ =\ & \frac{e^{\beta_0}}{C_1}\left(\sum_{t=1}^{\infty} t^2 e^{\beta_1 t} - n\sum_{t=1}^{\infty} t e^{\beta_1 t}, -\sum_{t=1}^{\infty} t e^{\beta_1 t} + n\sum_{t=1}^{\infty} e^{\beta_1 t}\right)\begin{pmatrix} 1 \\ n \end{pmatrix} \\ =\ & \frac{e^{\beta_0}}{C_1}\left[\sum_{t=1}^{\infty} t^2 e^{\beta_1 t} - 2n\sum_{t=1}^{\infty} t e^{\beta_1 t} + n^2\sum_{t=1}^{\infty} e^{\beta_1 t}\right] \\ =\ & \frac{e^{\beta_0}}{C_1}\left[\frac{(n-1)^2 e^{\beta_1} + (-2n^2+2n+1)e^{2\beta_1} + n^2 e^{3\beta_1} - e^{(n+1)\beta_1} - e^{(n+2)\beta_1}}{(1-e^{\beta_1})^3}\right] \\ =\ & \frac{e^{-\beta_0}(1-e^{\beta_1})\left[(n-1)^2 e^{\beta_1} + (-2n^2+2n+1)e^{2\beta_1} + n^2 e^{3\beta_1} - e^{(n+1)\beta_1} - e^{(n+2)\beta_1}\right]}{e^{3\beta_1} - n^2 e^{(n+2)\beta_1} + (2n^2-2)e^{(n+3)\beta_1} - n^2 e^{(n+4)\beta_1} + e^{(2n+3)\beta_1}} \\ \to\ & 0 \quad \text{as} \quad n \to \infty.\end{aligned}$$

By Theorems 1, 3 and Example (i) of Fahrmeir and Kaufmann (1985), the weak consistency and asymptotic normality of MLE hold. That is, there exists a sequence $\{\hat{\boldsymbol{\beta}}_n\}$ of MLE's such that $F_n^{T/2}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I_{2\times 2})$, where

$$F_n^{T/2}(\boldsymbol{\beta}) = e^{\beta_0/2}\begin{pmatrix} \sqrt{\sum_{t=1}^n e^{\beta_1 t}} & \frac{\sum_{t=1}^n t e^{\beta_1 t}}{\sqrt{\sum_{t=1}^n e^{\beta_1 t}}} \\ 0 & \sqrt{\frac{(\sum_{t=1}^n e^{\beta_1 t})(\sum_{t=1}^n t^2 e^{\beta_1 t}) - (\sum_{t=1}^n t e^{\beta_1 t})^2}{\sum_{t=1}^n e^{\beta_1 t}}} \end{pmatrix}.$$

Appendix B

Poisson Regression Model $\log \mu_t = \beta_0 + \beta_1 t/n$

Consider independent observations $\{Y_{nt}, t = 1, \cdots, n\}$ from Poisson distributions with mean $E(Y_{nt}) = \mu_{nt} = e^{\beta_0+\beta_1 t/n}$. Let $L_{n,n}(\boldsymbol{\beta})$ be the log-likelihood based on the observations $y_{n1}, y_{n2}, \cdots, y_{nn}$, and denote $\dot{L}_{n,n}(\boldsymbol{\beta})$ and $\ddot{L}_{n,n}(\boldsymbol{\beta})$ the first and second derivatives of $L_{n,n}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ respectively. Then

$$\begin{aligned} L_{n,n}(\boldsymbol{\beta}) &= -\sum_{t=1}^{n} e^{\beta_0+\beta_1 t/n} + \sum_{t=1}^{n} y_{nt}(\beta_0 + \beta_1 t/n) - \sum_{t=1}^{n} \log(y_{nt}!), \\ \dot{L}_{n,n}(\boldsymbol{\beta}) &= \begin{pmatrix} -\sum_{t=1}^{n} e^{\beta_0+\beta_1 t/n} + \sum_{t=1}^{n} y_{nt} \\ -\sum_{t=1}^{n} e^{\beta_0+\beta_1 t/n}(\frac{t}{n}) + \sum_{t=1}^{n} y_{nt}(\frac{t}{n}) \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n}(y_{nt} - e^{\beta_0+\beta_1 t/n}) \\ \frac{1}{n}\sum_{t=1}^{n}(t y_{nt} - t e^{\beta_0+\beta_1 t/n}) \end{pmatrix}. \end{aligned}$$

and

$$-\ddot{L}_{n,n}(\boldsymbol{\beta}) = e^{\beta_0} \begin{pmatrix} \sum_{t=1}^{n} e^{\beta_1 t/n} & \frac{1}{n}\sum_{t=1}^{n} t e^{\beta_1 t/n} \\ \frac{1}{n}\sum_{t=1}^{n} t e^{\beta_1 t/n} & \frac{1}{n^2}\sum_{t=1}^{n} t^2 e^{\beta_1 t/n} \end{pmatrix}.$$

We show that the weak consistency and asymptotic normality of the MLE of $\boldsymbol{\beta}$ as follows.

1) The score function $\dot{L}_{n,n}(\boldsymbol{\beta})$ is a martingale array.

Let $\mathcal{F}_{n,t-1} = \sigma(y_{kj}, j \le i, k \le n)$, then

$$\mathrm{E}\left(\dot{L}_{n,k}(\boldsymbol{\beta})|\mathcal{F}_{n,k-1}\right) = \begin{pmatrix} \sum_{t=1}^{k-1}(y_{nt} - e^{\beta_0+\beta_1 t/n}) \\ \frac{1}{n}\sum_{t=1}^{k-1}(t y_{nt} - t e^{\beta_0+\beta_1 t/n}) \end{pmatrix} = \dot{L}_{n,k-1}(\boldsymbol{\beta}).$$

And since $\mathrm{E}(y_{nk} - e^{\beta_0+\beta_1 k/n}) = 0$, $\{\dot{L}_{n,n}(\boldsymbol{\beta}), \mathcal{F}_{n,n}, n \ge 1\}$ is a zero-mean, square integrable martingale array. Let $a_{n,k}(\boldsymbol{\beta}) = \dot{L}_{n,k}(\boldsymbol{\beta}) - \dot{L}_{n,k-1}(\boldsymbol{\beta})$ be the difference sequence, then

$$a_{n,k}(\boldsymbol{\beta}) = \begin{pmatrix} y_{nk} - e^{\beta_0+\beta_1 k/n} \\ \frac{1}{n}(k y_{nk} - k e^{\beta_0+\beta_1 k/n}) \end{pmatrix} = (y_{nk} - e^{\beta_0+\beta_1 k/n}) \begin{pmatrix} 1 \\ \frac{k}{n} \end{pmatrix}.$$

2) The conditional variance of $a_{n,k}(\boldsymbol{\beta})$ converges.

$$\begin{aligned} \mathrm{Var}\left[a_{n,t}(\boldsymbol{\beta})|\mathcal{F}_{n,t-1}\right] &= \mathrm{E}\left[a_{n,t}(\boldsymbol{\beta}) a'_{n,t}(\boldsymbol{\beta})|\mathcal{F}_{n,t-1}\right] \\ &= \mathrm{E}\left[(y_{nt} - e^{\beta_0+\beta_1 t/n})^2 \begin{pmatrix} 1 \\ \frac{t}{n} \end{pmatrix} (1, \frac{t}{n})|\mathcal{F}_{n,t-1}\right] \\ &= \begin{pmatrix} e^{\beta_0+\beta_1 t/n} & \frac{t}{n} e^{\beta_0+\beta_1 t/n} \\ \frac{t}{n} e^{\beta_0+\beta_1 t/n} & \frac{t^2}{n^2} e^{\beta_0+\beta_1 t/n} \end{pmatrix}. \end{aligned}$$

Thus

$$G_{n,n}(\boldsymbol{\beta}) := \sum_{t=1}^{n} \mathrm{Var}\,[a_{n,t}(\boldsymbol{\beta})|\mathcal{F}_{n,t-1}] = e^{\beta_0} \begin{pmatrix} \sum_{t=1}^{n} e^{\beta_1 t/n} & \frac{1}{n}\sum_{t=1}^{n} t e^{\beta_1 t/n} \\ \frac{1}{n}\sum_{t=1}^{n} t e^{\beta_1 t/n} & \frac{1}{n^2}\sum_{t=1}^{n} t^2 e^{\beta_1 t/n} \end{pmatrix} = -\ddot{L}_{n,n}(\boldsymbol{\beta}).$$

and

$$\frac{1}{n} G_{n,n}(\boldsymbol{\beta}) \to e^{\beta_0} \begin{pmatrix} \frac{e^{\beta_1}-1}{\beta_1} & \frac{1-e^{\beta_1}+\beta_1 e^{\beta_1}}{\beta_1^2} \\ \frac{1-e^{\beta_1}+\beta_1 e^{\beta_1}}{\beta_1^2} & \frac{-2+2e^{\beta_1}-2\beta_1 e^{\beta_1}+\beta_1^2 e^{\beta_1}}{\beta_1^3} \end{pmatrix} := V \qquad (5.4)$$

which is positive definite.

3) The Lindeberg condition is satisfied.

For any $\epsilon > 0$,

$$\begin{aligned}
& \sum_{t=1}^{n} \mathrm{E}\left[\frac{1}{n} a'_{n,t} a_{n,t} I_{\{\frac{1}{n} a'_{n,t} a_{n,t} \geq \epsilon^2\}}\right] \\
= & \sum_{t=1}^{n} \mathrm{E}\left[\frac{1}{n}(y_{nt} - e^{\beta_0+\beta_1 t/n})^2 (1+\frac{t^2}{n^2}) I_{\{\frac{1}{n}(y_{nt}-e^{\beta_0+\beta_1 t/n})^2(1+\frac{t^2}{n^2}) \geq \epsilon^2\}}\right] \\
\leq & \sum_{t=1}^{n} \mathrm{E}\left[\frac{(1+\frac{t^2}{n^2})^2 (y_{nt} - e^{\beta_0+\beta_1 t/n})^4}{n^2\epsilon^2}\right] \\
\leq & \frac{1}{n^2\epsilon^2} \sum_{t=1}^{n} \mathrm{E}\left[4(y_{nt} - e^{\beta_0+\beta_1 t/n})^4\right] \\
= & \frac{4}{n^2\epsilon^2} \sum_{t=1}^{n} \left[e^{\beta_0+\beta_1 t/n} + 3e^{2\beta_0+2\beta_1 t/n}\right] \\
= & \frac{4}{n^2\epsilon^2} \left[e^{\beta_0} \frac{e^{\beta_1/n}(1-e^{\beta_1})}{1-e^{\beta_1/n}} + 3e^{2\beta_0} \frac{e^{2\beta_1/n}(1-e^{2\beta_1})}{1-e^{2\beta_1/n}}\right] \\
\to & 0
\end{aligned}$$

4) The asymptotic distribution of the score function is normal.

Combining 1), 2) and 3), and using Corollary 3.1 of Hall and Heyde (1980), we obtain that the asymptotic distribution of the score function $\dot{L}_{n,n}(\boldsymbol{\beta})$ is normal, i.e.,

$$\frac{1}{\sqrt{n}} \dot{L}_{n,n}(\boldsymbol{\beta}) \xrightarrow{d} V^{1/2} Z. \qquad (5.5)$$

where $V^{1/2}$ is the left Cholesky square root of the matrix V defined in (5.4), and Z is a standard normal random variable.

5) The Continuity condition holds.

Define a neighborhood of $\boldsymbol{\beta}$ as

$$N_n(\delta) = \{\tilde{\boldsymbol{\beta}}_{n,n} : ||\sqrt{n}(\tilde{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta})|| \leq \delta\} \quad \text{for } \delta > 0.$$

where $||\boldsymbol{v}||$ for vector $\boldsymbol{v}$ is defined as $\max(\boldsymbol{\lambda}'\boldsymbol{v})$ for $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$. Let $\tilde{\boldsymbol{\beta}}_{n,n} = (\tilde{\beta}_0, \tilde{\beta}_1)'$, then

$$\begin{aligned} & \sup_{\tilde{\boldsymbol{\beta}}_{n,n} \in N_n(\delta)} ||\tfrac{1}{n}\left[-\ddot{L}_{n,n}(\tilde{\boldsymbol{\beta}}_{n,n}) - G_{n,n}(\boldsymbol{\beta})\right]|| \\ = & \sup_{\tilde{\boldsymbol{\beta}}_{n,n} \in N_n(\delta)} ||\tfrac{1}{n}\begin{pmatrix} \sum_{t=1}^n (e^{\tilde{\beta}_0+\tilde{\beta}_1 t/n} - e^{\beta_0+\beta_1 t/n}) & \sum_{t=1}^n \frac{t}{n}(e^{\tilde{\beta}_0+\tilde{\beta}_1 t/n} - e^{\beta_0+\beta_1 t/n}) \\ \sum_{t=1}^n \frac{t}{n}(e^{\tilde{\beta}_0+\tilde{\beta}_1 t/n} - e^{\beta_0+\beta_1 t/n}) & \sum_{t=1}^n \frac{t^2}{n^2}(e^{\tilde{\beta}_0+\tilde{\beta}_1 t/n} - e^{\beta_0+\beta_1 t/n}) \end{pmatrix}|| . \end{aligned} \tag{5.6}$$

A Taylor expansion gives

$$e^{\tilde{\beta}_0+\tilde{\beta}_1 t/n} - e^{\beta_0+\beta_1 t/n} = \left[(\tilde{\beta}_0 - \beta_0) + (\tilde{\beta}_1 - \beta_1)(\frac{t}{n})\right] e^{\beta_0^*+\beta_1^* t/n},$$

where $\beta_0^* + \beta_1^* t/n$ lies between $\tilde{\beta}_0 + \tilde{\beta}_1 t/n$ and $\beta_0 + \beta_1 t/n$. Since $\tilde{\boldsymbol{\beta}}_{n,n} \in N_n(\delta)$, thus $|\tilde{\beta}_0 - \beta_0| \leq \delta/\sqrt{n}$, and $|\tilde{\beta}_1 - \beta_1| \leq \delta/\sqrt{n}$. Hence

$$\begin{aligned} & \frac{1}{n}\sum_{t=1}^n (e^{\tilde{\beta}_0+\tilde{\beta}_1 t/n} - e^{\beta_0+\beta_1 t/n}) \\ = & \frac{1}{n}\sum_{t=1}^n \left[(\tilde{\beta}_0 - \beta_0) + (\tilde{\beta}_1 - \beta_1)(\frac{t}{n})\right] e^{\beta_0^*+\beta_1^* t/n} \\ = & \frac{1}{n}(\tilde{\beta}_0 - \beta_0)e^{\beta_0^*}\sum_{t=1}^n e^{\beta_1^* t/n} + \frac{1}{n^2}(\tilde{\beta}_1 - \beta_1)e^{\beta_0^*}\sum_{t=1}^n t e^{\beta_1^* t/n} \\ \leq & \frac{\delta}{\sqrt{n}}e^{\beta_0^*}\left[\frac{1}{n}\sum_{t=1}^n e^{\beta_1^* t/n}\right] + \frac{\delta}{\sqrt{n}}e^{\beta_0^*}\left[\frac{1}{n^2}\sum_{t=1}^n t e^{\beta_1^* t/n}\right] \\ \to & \ 0 \quad \text{as } n \to \infty. \end{aligned}$$

Similarly, we can show all other elements of the matrix in (5.6) converge to 0 as $n \to \infty$. Therefore,

$$\sup_{\tilde{\boldsymbol{\beta}}_{n,n} \in N_n(\delta)} ||\frac{1}{n}\left[-\ddot{L}_{n,n}(\tilde{\boldsymbol{\beta}}_{n,n}) - G_{n,n}(\boldsymbol{\beta})\right]|| \to 0 \quad \text{holds for any } \delta > 0. \tag{5.7}$$

6) The MLE of $\boldsymbol{\beta}$ exists and is unique asymptotically and is weakly consistent.

Since $-\ddot{L}_{n,n}(\boldsymbol{\beta})$ is a covariance matrix, it is positive definite for all n, and hence there is at most one zero of the score function $\dot{L}_{n,n}(\boldsymbol{\beta})$. The solution of

$\dot{L}_{n,n}(\boldsymbol{\beta}) = \mathbf{0}$ gives a local (global) maximum of the likelihood if it exists. For any n, $\delta > 0$, and $N_n(\delta) = \{\tilde{\boldsymbol{\beta}}_{n,n} : ||\sqrt{n}(\tilde{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta})|| \leq \delta\}$, the event

$$\{L_{n,n}(\tilde{\boldsymbol{\beta}}) - L_{n,n}(\boldsymbol{\beta}) < 0 \text{ for all } \tilde{\boldsymbol{\beta}} \in \partial N_n(\delta)\}$$

implies that there is a local maximum inside of $N_n(\delta)$, where $\partial N_n(\delta)$ is the boundary of $N_n(\delta)$. From the above discussion, there is only one local (global) maximum, so this local (global) maximum must be located at the MLE $\hat{\boldsymbol{\beta}}_{n,n}$. Now we need to show that for any given $\eta > 0$, there exist $\delta > 0$ and n_1 such that

$$P\left[L_{n,n}(\tilde{\boldsymbol{\beta}}) - L_{n,n}(\boldsymbol{\beta}) < 0 \text{ for all } \tilde{\boldsymbol{\beta}} \in \partial N_n(\delta)\right] \geq 1 - \eta \tag{5.8}$$

for all $n \geq n_1$. A Taylor expansion gives

$$L_{n,n}(\tilde{\boldsymbol{\beta}}) - L_{n,n}(\boldsymbol{\beta}) = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\dot{L}_{n,n}(\boldsymbol{\beta}) + \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}^*_{n,n}$ is between $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$.

Let $\boldsymbol{\lambda} = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})/\delta$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$, then

$$\begin{aligned}
& P\left[L_{n,n}(\tilde{\boldsymbol{\beta}}) - L_{n,n}(\boldsymbol{\beta}) < 0 \text{ for all } \tilde{\boldsymbol{\beta}} \in \partial N_n(\delta)\right] \\
= \; & P\left[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\dot{L}_{n,n}(\boldsymbol{\beta}) + \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) < 0\right] \\
= \; & P\left[\frac{\delta}{\sqrt{n}}\boldsymbol{\lambda}'\dot{L}_{n,n}(\boldsymbol{\beta}) < -\frac{1}{2}\frac{\delta^2}{n}\boldsymbol{\lambda}'\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n})\boldsymbol{\lambda}\right]
\end{aligned}$$

Using the definition $||\frac{1}{\sqrt{n}}\dot{L}_{n,n}(\boldsymbol{\beta})|| = \max\left(\boldsymbol{\lambda}'\frac{1}{\sqrt{n}}\dot{L}_{n,n}(\boldsymbol{\beta})\right)$ for $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$, it suffices to show

$$P\left[||\frac{1}{\sqrt{n}}\dot{L}_{n,n}(\boldsymbol{\beta})|| < \frac{\delta}{2n}\boldsymbol{\lambda}'(-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}))\boldsymbol{\lambda}\right] \geq 1 - \eta \quad \text{for any given } \eta > 0.$$

Since $-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n})$ is positive definite for all n, and $\frac{1}{n}(-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}))$ converges to a positive definite matrix, so we can find a positive constant c such that $\frac{1}{n}\boldsymbol{\lambda}'(-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}))\boldsymbol{\lambda} \geq c$ for all n. By Markov's inequality,

$$P\left[||V^{-1/2}Z|| < \frac{\delta c}{2}\right] \geq 1 - \frac{2}{\delta c}\mathrm{E}||V^{-1/2}Z|| := 1 - \eta.$$

thus

$$P\left[\|V^{1/2}Z\| < \frac{\delta}{2n}\boldsymbol{\lambda}'(-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}))\boldsymbol{\lambda}\right] \geq 1-\eta.$$

by (5.5) we have

$$P\left[\|\frac{1}{\sqrt{n}}\dot{L}_{n,n}(\boldsymbol{\beta})\| < \frac{\delta}{2n}\boldsymbol{\lambda}'(-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}))\boldsymbol{\lambda}\right] \geq 1-\eta.$$

Given $\eta > 0$, choose δ such that (5.8) holds for $n \geq n_1$. For such δ, we have

$$P\left[\|\sqrt{n}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta})\| \leq \delta\right] > 1-\eta.$$

and hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta}) = O_p(1)$. $\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta} = O_p(\frac{1}{\sqrt{n}}) = o_p(1)$, i.e.

$$\hat{\boldsymbol{\beta}}_{n,n} \xrightarrow{p} \boldsymbol{\beta}. \tag{5.9}$$

7) The MLE of $\boldsymbol{\beta}$ is asymptotically normal.

By a Taylor expansion,

$$\dot{L}_{n,n}(\boldsymbol{\beta}) - \dot{L}_{n,n}(\hat{\boldsymbol{\beta}}_{n,n}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{n,n})'\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}).$$

since $\dot{L}_{n,n}(\hat{\boldsymbol{\beta}}_{n,n}) = \mathbf{0}$,

$$\dot{L}_{n,n}(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{n,n})'\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n}).$$

thus,

$$\frac{1}{\sqrt{n}}\dot{L}_{n,n}(\boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{n,n} - \boldsymbol{\beta})'\left[\frac{1}{n}\left(-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n})\right)\right]. \tag{5.10}$$

The continuity condition (5.7) and (5.4) give

$$\frac{1}{n}\left[-\ddot{L}_{n,n}(\boldsymbol{\beta}^*_{n,n})\right] \xrightarrow{p} V \quad \text{for any } \boldsymbol{\beta}^*_{n,n} \in N_n(\delta). \tag{5.11}$$

From (5.5), (5.10) and (5.11), we have $\sqrt{n}(\hat{\boldsymbol{\beta}}_{nn} - \boldsymbol{\beta})' \xrightarrow{d} V^{1/2}ZV^{-1}$, i.e.

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{nn} - \boldsymbol{\beta}) \xrightarrow{d} V^{-T/2}Z, \tag{5.12}$$

where $V^{-T/2}$ is the right Cholesky square root of V^{-1}.

REFERENCES

1. Blais, M. MacGibbon, B. & Roy, R. (2000). Limit theorems for regression models of time series of counts. *Statistics and Probability Letters*, 46 (2), 161-168.

2. Brannas, K. & Johansson, P. (1994). Time series count data regression. *Commun. Statist.- Theory and Meth.* 23 (10), 2907-2925.

3. Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods* (2nd ed.). New York: Springer-Verlag.

4. Brockwell, P. J. & Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. New York: Springer.

5. Burnett, R. T., Shedde, J. & Krewski, D. (1991). Nonlinear regression models for correlated count data. *Environmetrics* 3 (2), 211-222.

6. Campbell, M. J. (1994). Time series regression for counts: an investigation into the relationship between Sudden Infant Death Syndrome and environmental temperature. *Journal of Royal Statistical Society*, Series A 157, Part 2, 191-208.

7. Chan, K. S. & Ledolter J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association* 90, 242-252.

8. Cox, D. R. (1981). Statistical analysis of time series, some recent developments. *Scandinavia Journal of Statistics* 8, 93-115.

9. Davidson, J. (1992). A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes. *Econometric Theory* 8, 313-329.

10. Davis, R. A., Dunsmuir, W. T. M. & Wang, Y. (1998). Modelling time series of count data. In *Asymptotics, Nonparametrics and Time Series* (ed Subir Ghosh), Marcel Dekker.

11. Davis, R. A., Dunsmuir, W. T. M. & Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika* 87, 491-505.

12. Davydov, Yu A. (1970). The invariance principle for stationary processes. *Theory of Probability and Its Applications* 15, 487-498.

13. Dean, C. B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association* 87, 451-457.

14. Dean, C. & Lawless J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* 84, 467-472.

15. De Jong, P. & Shephard, N. (1995). The simulation smoother for time series models. *Biometrika* 82, 339-350.

16. Diggle, P. J., Liang, K. Y. & Zeger S. L. (1994). *Analysis of longitudinal data.* New York: Oxford University Press.

17. Durbin, J. & Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84, 669-684.

18. 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.

19. Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear model. *Annals of Statistics* 13, 342-368.

20. Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models.* New York: Springer-Verlag.

21. Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of Royal Statistical Society*, Series B 56, 261-274.

22. Geyer, C. J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice*, eds. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J., 241-258. London: Chapman and Hall.

23. Gourieroux, C, Monfort, A. & Trognon, A. (1984a). Pseudo maximum likelihood methods: theory. *Econometrica* 52, 681-700.

24. Gourieroux, C, Monfort, A. & Trognon, A. (1984b). Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica* 52, 701-720.

25. Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and Its Application.* New York: Academic Press.

26. Hannan, E. J. (1970). *Multiple Time Series.* New York: John Wiley & Sons.

27. Hannan, E. J. & Heyde, C. C. (1972). On limit theorems for quadratic functions of discrete time series. *The Annals of Mathematical statistics* 43, 6, 2058-2066.

28. Ibragimov, I. A. & Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*, ed. Kingman, J. F. C., Wolters-Noordhoff Publishing, Groningen, the Netherlands.

29. Jorgensen, B., Lundbye-Christensen, S., Song, X.-K. & Sun, L. (1995). A state space model for multivariate longitudinal count data. *Technique Report* 148, Department of Statistics, University of British Columbia.

30. Jorgensen, B., Lundbye-Christensen, S., Song, X.-K. & Sun, L. (1996). A longitudinal study of emergency room visits and air pollution for Prince George, British Columbia. *Statistics in Medicine* 15, 823-836.

31. Kuk, A. Y. C. (1997). A note on the use of approximating model in Monte Carlo maximum likelihood estimation. Working paper. Department of Statistics, University of New South Wales.

32. Kuk A. Y. C. & Cheng Y. W. (1997). The Monte Carlo Newton-Raphson algorithm. *Journal of statistical computing and simulation* 59, 233-250.

33. Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

34. Lo, A. W. & MacKinlay A. C. (1989). The size and power of the variance ratio test in finite samples: a Monte Carlo investigation. *Journal of Econometrics* 40, 203-238.

35. Lobato, I., Nankervis J. & Savin N. E. (1998). Testing that stock returns are uncorrelated using a modified Box-Pierce Q test. Technical report. Department of Economic, University of Iowa, Iowa City 52242.

36. Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, Series B 44, 226-233.

37. McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics* 11, 59-67.

38. McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

39. Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A 135, 370-384.

40. Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. *Econometric Theory* 7, 186-199.

41. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C: the Art of Scientific Computing* (2nd ed.). Cambridge University Press.

42. Rockafellar R.T. (1970). *Convex Analysis*. Princeton University Press, Princeton, New Jersey.

43. Roussas, G. G., Tran, L. T. & Ioannides, D. A. (1992). Fixed design regression for time series: asymptotic normality. *Journal of Multivariate Analysis* 40, 262-291.

44. Schall, R. (1991). Estimation in generalised linear models with random effects. *Biometrika* 78, 719-727.

45. Shephard, N. & Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* 84, 653-667.

46. Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* 61, 439-447.

47. White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1-25.

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