Observation Driven Models for Poisson Counts

Richard A. Davis, William T.M. Dunsmuir and Sarah B. Streett

December 24, 2001

Abstract

This paper is concerned with a general class of observation driven models for time series of counts whose conditional distributions given past observations and explanatory variables follow a Poisson distribution. These models provide a flexible framework for modeling a wide range of dependence structures. Conditions for stationarity and ergodicity of these processes are established from which the large sample properties of the maximum likelihood estimators can be derived. Simulations are provided to give additional insight into the finite sample behavior of the estimates. Finally an application to a regression model for daily counts of accident and emergency room presentations for asthma at several Sydney hospitals is described.

1 Introduction

In recent years there has been considerable development of models for non-Gaussian time series. In particular the special case of Poisson observations is of interest in a variety of applications including the modeling of the effects of environmental pollution on human health and the impact of policy controls on road deaths. Davis et al. (1999) provides a review of models for Poisson time series. There, the classification due to Cox (1981) of models into observation and parameter driven processes is described. In particular a new class of models, which we will refer to as generalized linear autoregressive moving average (GLARMA) models is introduced and its properties developed in part. The purpose of this paper is to develop these models more comprehensively.

In general terms, parameter driven models require considerable computational effort in order to obtain parameter estimates - see Durbin and Koopman (2000) and Jung and Liesenfeld (2001) for recent contributions to this topic. In addition, because they are built on a latent process, forecasting also requires considerable computational effort. Parameter driven models are, however, straightforward in their interpretation of the effects of covariates on the observed count process, an appealing point.

Observation driven models are sometimes referred to as transition models in the longitudinal data analysis literature (e.g., see Diggle, Liang and Zeger, 1994). Zeger and Qaqish (1988) review various observation driven models for count time series. In particular they imply various desirable properties that such models should possess. First, the marginal mean of Y_t should be approximated as

$$E(Y_t) = E(\mu_t) \approx \exp(\mathbf{x}_t^T \boldsymbol{\beta})$$

so that the regression coefficients can be interpreted as the proportional change in the marginal expectation of Y_t on the logarithm scale given a unit change in the regressor variables. This seems useful from the point of view of interpretation. Additionally, for stationary processes, both positive and negative serial dependence should be possible. Some of the models that are discussed in Zeger and Qaqish (1988) do not satisfy this last property and will admit stationary solutions only for negative serial dependence, a case that is less common in many applications than that of positive dependence.

To develop the class of models that are considered here, let

$$\mathcal{H}_t = (\mathbf{Y}^{(t-1)}, \mathbf{x}^{(t)})$$

be the past of the observed count process and the past and present of the regressor variables. Assume that the conditional distribution of $Y_t|\mathcal{H}_t$ is Poisson with mean μ_t. A simple and appealing way to build serial dependence in the model is to require the log-mean process to depend linearly on previous observations. That is,

$$\log(\mu_t) = \mathbf{x}_t^T \boldsymbol{\beta} + \sum_{i=1}^{p} \gamma_i Y_{t-i}. \tag{1}$$

Note that the Y_{t-i} enter without any form of mean correction or centering. Model (1) is applied to data in Fahrmier and Tutz (1994). Zeger and Qaqish (1988) point out that (1) cannot be stationary unless, at least in the case $p = 1$, $\gamma_1 \leq 0$ thereby excluding the possibility of positive dependence. This point is also acknowledged by Fahrmier and Tutz. It might be thought that the difficulties with model (1) could be overcome by subtracting the 'fixed effects' component of the mean from the Y_t to arrive at

$$\mu_t = \exp(\mathbf{x}_t^T \boldsymbol{\beta} + \sum_{i=1}^{p} \gamma_i (Y_{t-i} - \exp(\mathbf{x}_t^T \boldsymbol{\beta})), \tag{2}$$

but in fact this will not lead to a stationary process as can be seen by using a similar argument to that given by Zeger and Qaqish (1988). In an unpublished report Shephard (1995) extends (2) by using the standardized deviations $(Y_{t-i} - \exp(\mathbf{x}_t^T \boldsymbol{\beta}))/\exp(\mathbf{x}_t^T \boldsymbol{\beta})$ and by including lag structure that corresponds to rational functions similar to those for autoregressive moving average linear models. In a later unpublished paper Rydberg and Shephard (1998) consider models with exponential family distributions conditional on $\mathcal{H}_t$ which are analogous but for which the standardization uses conditional standard deviations rather than variances. Davis et al. (1999) consider a general class of models that allows for normalization to occur with various powers of the conditional variance including those considered by Shephard (1995) and Rydberg and Shephard (1998). In line with terminology introduced by Shephard (1995) we will refer to this general class of models as GLARMA models.

Observation driven models are generally very easy to fit using conditional maximum likelihood. Here conditioning refers to conditioning on some initial values and not on random effects. Also forecasting of future observations is straightforward. However, because of the way in which past observations feed into the mean term, the interpretation of the effect of covariates can be confused to varying extents depending on the form of the model.

In many practical problems, the primary objective is to develop models that relate covariates, such as environmental or policy intervention variables, to the observed time series of counts, such as the daily number of asthma cases at a hospital. Often there are numerous covariates that may need to be considered for inclusion in the model. In some instances, many comparable time series of counts need to be modeled as part of a larger study. For example, in studying the impact of alcohol or traffic regulation policy interventions on road deaths or youth suicides, all regions or states in a country may need to be considered separately since the timing and nature of these variables may have regional variations. In these settings there is a distinct advantage to have methods available that are easy to implement and rapid to compute for investigating the impact of covariates on time series of counts which properly control for serial dependence. At the present time, and for realistically long and numerous time series arising in the areas we have described, the computationally intensive methods required for fitting parameter driven models are not yet routinely available. This paper develops a class of observation driven models that are straightforward to implement and are rapid to fit. In addition, these models adjust for serial dependence in the inference for fixed effects and allow reasonable interpretations of the effects of these covariates on the response variable.

Our observation driven model is introduced in Section 2, where general properties about the process are also given. In Section 3, we consider maximum likelihood estimates for these models with supporting simulation results. Section 4 contains an application of fitting these models to data consisting of the number of asthma presentations at a hospital in the Sydney metropolitan area. Technical results related to the the properties of the process and the asymptotic normality of the maximum likelihood estimator are given in Section 5.

2 Observation Driven Models

2.1 The Basic Model

To introduce our model, assume that the observation Y_t given the past history $\mathcal{F}_{t-1} = \sigma(Y_s, s \leq t-1)$ is Poisson with mean μ_t which will be denoted by

$$Y_t|\mathcal{F}_{t-1} \sim P(\mu_t).$$

It is further assumed that the *state* process $\log(\mu_t)$ follows a linear model in the explanatory variables with residuals that have a moving average structure. The noise driving the moving average will be a martingale difference sequence generated from the data and hence the name observation driven model. Formally, the state process is given by

$$W_t := \log(\mu_t) = \mathbf{x}_t^T \boldsymbol{\beta} + \sum_{i=1}^{q} \theta_i e_{t-i},$$

where

$$e_t = (Y_t - \mu_t)/\mu_t^{\lambda}, \quad \lambda \geq 0.$$

Since the conditional mean $E(Y_t|\mathbf{Y}^{(t-1)})$ depends on the whole past, the process $\{Y_t\}$ is no longer Markov. However, the mean process $\log(\mu_t)$ is q^{th} order Markov. Unless $\mathbf{x}_t^T\boldsymbol{\beta}$ is constant, $\log(\mu_t)$ is not a time-homogeneous process.

Properties of the Basic Model ($q = 1,\ \mathbf{x}_t^T\boldsymbol{\beta} = \beta$)

In this simple but illuminating case, the state process reduces to

$$W_t = \beta + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}.$$

Under this formulation, the process $\{W_t\}$ as well as $\{\mu_t = e^{W_t}\}$ possesses many desirable properties. For example, the process $\{W_t\}$ is a Markov process with mean

$$\mathrm{E}(W_t) = \mathrm{E}[\mathrm{E}(W_t|W_{t-1})] = \beta,$$

and variance

$$\mathrm{Var}(W_t) = \mathrm{Var}(\mathrm{E}(W_t|W_{t-1})) + \mathrm{E}\mathrm{Var}(W_t|W_{t-1}) = \gamma^2\mathrm{E}[\exp((1-2\lambda)W_{t-1})].$$

It follows that for $\lambda = 0.5$, $\mathrm{Var}(W_t) = \gamma^2$, while for $\lambda = 1$,

$$\mathrm{Var}(W_t) = \gamma^2\mathrm{E}\left[e^{-W_{t-1}}\right] \geq \gamma^2 e^{-\mathrm{E}[W_{t-1}]} = \gamma^2 e^{-\beta}.$$

The state space for the conditional distribution of W_t given W_{t-1} has the following form:

$$W_t \geq \beta - \gamma e^{(1-\lambda)W_{t-1}}, \ \text{ if } \gamma \geq 0,$$

and

$$W_t \leq \beta - \gamma e^{(1-\lambda)W_{t-1}}, \ \text{ if } \gamma \leq 0.$$

While the range of W_t does not depend on the value of W_{t-1} for $\lambda = 1$, the range does depend on W_{t-1} for values of $\lambda < 1$ which severely complicates the analysis.

Another important property is that the process $\{W_t\}$ is uniformly ergodic for the case $\lambda = 1$ (see Appendix 5.1). Hence, there exists a unique stationary distribution for the log-mean process in this case. For $1/2 \leq \lambda < 1$, there exists a stationary distribution, yet the uniqueness of such a distribution is currently unknown (see Appendix 5.1). For $\lambda < 1/2$, existence of a stationary distribution has not been established as of yet.

The conditions on the state process translate into the following property on the mean process:

$$\mathrm{E}(\mu_t|\mu_{t-1}) = \mathrm{E}[\exp(\beta + \gamma(Y_{t-1} - \mu_{t-1})/\mu_{t-1}^{\lambda})].$$

Using the moment generating function for the Poisson distribution, we obtain

$$\begin{aligned}
\mathrm{E}(\mu_t|\mu_{t-1}) &= \exp(\beta)\exp(-\gamma\mu_{t-1}^{1-\lambda})\mathrm{E}(\exp(Y_{t-1}\gamma/\mu_{t-1}^{\lambda})) \\
&= \exp(\beta)\exp(-\gamma\mu_{t-1}^{1-\lambda})\exp(\mu_{t-1}(e^{\gamma/\mu_{t-1}^{\lambda}} - 1)) \\
&= \exp(\beta)\exp(-\gamma\mu_{t-1}^{1-\lambda} + \mu_{t-1}(e^{\gamma/\mu_{t-1}^{\lambda}} - 1)) \\
&= \exp(\beta)\exp(\mu_{t-1}(\sum_{k=2}^{\infty}\frac{(\gamma/\mu_{t-1}^{\lambda})^k}{k!})) \\
&= \exp(\beta)\exp(\mu_{t-1}(e^{\gamma/\mu_{t-1}^{\lambda}} - 1 - \gamma\mu_{t-1}^{-\lambda})). \qquad (3)
\end{aligned}$$

If $\lambda = 0$, equation (3) becomes

$$\mathrm{E}(\mu_t|\mu_{t-1}) = \exp(\beta)\exp(\mu_{t-1}(e^{\gamma} - 1 - \gamma)),$$

so that if $\gamma \geq 0$ the conditional means will evolve in an unstable fashion. Thus, when $\lambda = 0$, $\mathrm{E}(\mu_t|\mu_{t-1})$ will grow without bound whenever μ_t becomes positive. In contrast, for $\lambda = 1$, equation (3) becomes

$$\mathrm{E}(\mu_t|\mu_{t-1}) = \exp(\beta - \gamma)\exp(\mu_{t-1}(e^{\gamma/\mu_{t-1}} - 1)),$$

which is bounded as $\mu_t \to \infty$. For other values, $0 < \lambda < 1$, the stability properties of the process are less clear.

2.2 The GLARMA Model

Extensions to autoregressive-moving average filters applied to past values of e_t can also be made to the basic model. Let $\{U_t\}$ be the ARMA(p, q) process with noise given by the martingale difference sequence $\{e_t\}$, i.e.,

$$U_t = \phi_1 U_{t-1} + \cdots + \phi_p U_{t-p} + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}\,,$$

where the AR and MA polynomials, $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$, respectively, have all their zeros outside the unit circle. Then the best predictor of U_t based on the infinite past $\{U_{t-1}, U_{t-2}, \ldots\}$ is

$$\hat{U}_t = \sum_{i=1}^{\infty} \tau_i e_{t-i}\,,$$

where

$$\begin{aligned}\sum_{i=1}^{\infty} \tau_i z^j &= (1 - \sum_{i=1}^{p} \phi_i z^i)^{-1}(1 + \sum_{i=1}^{q} \theta_i z^i) - 1 \\ &= \phi(z)^{-1}\theta(z) - 1.\end{aligned}$$

The model for $\log(\mu_t)$ is then

$$W_t = \mathbf{x}_t^T \boldsymbol{\beta} + Z_t = \mathbf{x}_t^T \boldsymbol{\beta} + \sum_{i=1}^{\infty} \tau_i e_{t-i}\,,$$

where $Z_t = \hat{U}_t$ We refer to such models for the data $\{Y_t\}$ as generalized linear ARMA of order (p, q) or GLARMA(p, q). In the model fitting stage, Z_t is computed using the ARMA recursions. Specifically, for $t \leq 0$ set $e_t = 0$ and $Z_t = 0$ and for $t > 0$, the following recursions are applied:

$$\begin{aligned}\hat{Z}_t &= \phi_1(\hat{Z}_{t-1} + e_{t-1}) + \cdots + \phi_p(\hat{Z}_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} \\ W_t &= \mathbf{x}_t^T \boldsymbol{\beta} + \hat{Z}_t \\ e_t &= (Y_t - e^{W_t})e^{-\lambda W_t}\,.\end{aligned}$$

The structure of this model is similar to the one formulated in Shephard (1995, unpublished). He presents an argument, based on a Taylor series linearization of the link function, for using $\lambda = 1$ in the definition of e_t at least in the Poisson case. The inclusion of explanatory variables in the model is also handled in a different fashion than that above.

Properties of the Extended Model

Under the initial conditions above, $e_s = 0$ and $Y_s = 0$ for $s \leq 0$, $\mathcal{F}^e_{s-1} = \{e_t : t \leq s-1\}$ and $\mathcal{F}_{s-1} = \{Y_t : t \leq s-1\}$ generate the same $\sigma-$fields and hence it follows that the e_t form a martingale difference sequence, i.e.,

$$\mathrm{E}(e_s|\mathcal{F}^e_{s-1}) = 0 \quad \text{for } s \geq 1.$$

Hence, the e_t have zero mean and variance

$$\mathrm{E}(e_t^2) = \mathrm{E}[\mathrm{E}(e_t^2|\mu_t)] = \mathrm{E}[\mu_t^{1-2\lambda}],\ t \geq 1,$$

which is equal to 1 for $\lambda = 0.5$. It also follows from the martingale difference property that $\mathrm{Cov}(e_t, e_s) = 0$ for $t \neq s$. From the above properties we have, for any λ,

$$\mathrm{E}(W_t) = \mathbf{x}_t^T\boldsymbol{\beta},$$

$$\mathrm{Var}(W_t) = \sum_{i=1}^{\infty} \tau_i^2 \mathrm{E}[\mu_{t-i}^{1-2\lambda}],$$

and for $s = t+h$, $h > 0$,

$$\mathrm{Cov}(W_t, W_{t+h}) = \sum_{i=1}^{\infty} \tau_i \tau_{i+h} \mathrm{E}[\mu_{t-i}^{1-2\lambda}].$$

If $\lambda = 0.5$, then the covariances do not depend on time t even if $\{\mu_t\}$ is not strictly stationary.

While the process $\{W_t\}$ has mean $\mathbf{x}_t^T\boldsymbol{\beta}$, the process $\{\mu_t\}$ has mean greater than $e^{\mathbf{x}_t^T\boldsymbol{\beta}}$. Nevertheless, we have

$$\begin{aligned} W_t &= \mathbf{x}_t^T\boldsymbol{\beta} + Z_t \\ &\approx \mathbf{x}_t^T\boldsymbol{\beta} + U_t', \end{aligned}$$

in the sense that the distributions will be similar, where U_t' is a Gaussian stationary sequence with zero mean and variances and covariances matched to those for Z_t. Roughly speaking, $\{U_t'\}$ is a proxy for a latent process. Hence, using results obtained for latent processes, we have, again for the case $\lambda = 0.5$,

$$\begin{aligned} \mathrm{E}(e^{W_t}) &\approx e^{\mathbf{x}_t^T\boldsymbol{\beta} + \mathrm{Var}(Z_t)} \\ &= e^{\mathbf{x}_t^T\boldsymbol{\beta} + \frac{1}{2}\sum_{i=1}^{\infty}\tau_i^2}. \end{aligned}$$

Thus, in practice, the bias of $\mathrm{E}(\mu_t)$ as an estimate of $e^{\mathbf{x}_t^T\boldsymbol{\beta}}$ can be approximately adjusted for and, perhaps most importantly, the regression coefficients are then approximately interpretable as the amount by which the mean of Y_t on the log-scale would change for a unit change in the regressors.

While the distribution of $\{e_t\}$ is not normal, the linear combination $Z_t = \sum_{i=1}^{\infty} \tau_i e_{t-i}$ will have a distribution which may be closely approximated by a sequence of correlated normal random variables. The extent to which the joint distribution of the sequence $\{e_t\}$ differs from a process of independent Gaussian random variables with zero mean and unit variance will govern the extent to which the approximation

$$\mathrm{E}(e^{W_t}) \approx e^{\mathbf{x}_t^T\boldsymbol{\beta}+\frac{1}{2}\sum_{i=1}^{\infty}\tau_i^2}$$

holds.

Another advantage of the above formulation is that an approximately *unbiased* plot of μ_t can be generated by

$$\hat{\mu}_t = \exp(\hat{W}_t - \frac{1}{2}\sum_{i=1}^{\infty}\hat{\tau}_i^2),$$

where estimates have been used throughout. Thus, it is easy to predict with this model. In fact $\hat{\mu}_t$ could be used as the one step ahead forecast of Y_t, given a value for $\mathbf{x}_t$ or a reliable forecast of it.

3 Estimation and Inference for the Model

3.1 Maximum Likelihood Estimation

The likelihood and its first and second derivatives can easily be computed recursively and used in a Newton-Raphson update procedure. Standard errors for the parameter estimates that properly account for serial dependence are also readily available. The details follow.

Let $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ and define $L_t(\boldsymbol{\delta}) = \log f(y_t|\mathcal{F}_{t-1})$, where f is the conditional Poisson density of Y_t given $\mathcal{F}_{t-1}$. The log-likelihood can then be written as $\sum_{t=1}^n L_t(\boldsymbol{\delta})$ which, upon ignoring terms which do not involve the parameters, becomes

$$L(\boldsymbol{\delta}) = \sum_{t=1}^{n}\left(Y_t W_t(\boldsymbol{\delta}) - e^{W_t(\boldsymbol{\delta})}\right),$$

where

$$\log(\mu_t) = W_t(\boldsymbol{\delta}) = \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{\infty}\tau_i(\boldsymbol{\gamma})e_{t-i}(\boldsymbol{\delta}) \qquad (4)$$

and

$$e_t(\boldsymbol{\delta}) = (Y_t - \mu_t)/\mu_t^{\lambda}.$$

For brevity, we will often suppress the dependence of e_t on $\boldsymbol{\delta}$. The first and second derivatives of L are given by the following expressions

$$\frac{\partial L}{\partial \boldsymbol{\delta}} = \sum_{t=1}^{n}(Y_t - \mu_t)\frac{\partial W_t}{\partial \boldsymbol{\delta}} = \sum_{t=1}^{n} e_t\mu_t^{\lambda}\frac{\partial W_t}{\partial \boldsymbol{\delta}}$$

and

$$\begin{aligned}\frac{\partial^2 L}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} &= \sum_{t=1}^{n}\left[(Y_t - \mu_t)\frac{\partial^2 W_t}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} - \mu_t\frac{\partial W_t}{\partial \boldsymbol{\delta}}\frac{\partial W_t}{\partial \boldsymbol{\delta}^T}\right] \\ &= \sum_{t=1}^{n}\left[e_t\mu_t^{\lambda}\frac{\partial^2 W_t}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} - \mu_t\frac{\partial W_t}{\partial \boldsymbol{\delta}}\frac{\partial W_t}{\partial \boldsymbol{\delta}^T}\right].\end{aligned}$$

The remaining expressions needed to calculate these derivatives are given in Appendix 5.2. Asymptotic results for these estimates are given in Appendix 5.3 for the basic model with $\lambda = 1, p = 1$ and $\mathbf{x}_t\boldsymbol{\beta} = \beta$. In this case, the asymptotic distribution of the maximum likelihood estimates is $N(0, V^{-1})$, where

$$V = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} e^{W_t(\delta_0)} \dot{W}_t \dot{W}_t^T, \tag{5}$$

with $\dot{W}_t = \frac{\partial W_t(\delta_0)}{\partial \delta}$.

To initialize the Newton Raphson recursions we have found that using the GLM estimates without the autoregressive moving average terms together with zero initial values for $e_t,\ t \le 0$, gives reasonable starting values. Convergence in the majority of cases reported below (in which the first derivatives were less than 10^{-6}) occurred within 10 iterations from these starting conditions. The covariance matrix of the estimators is estimated by

$$\hat{\Omega} = -\left(\frac{\partial^2 L(\hat{\theta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T}\right)^{-1} \tag{6}$$

3.2 Simulation Results.

To illustrate the asymptotic properties of the parameter estimates, we simulate from two models and compare the results with the derived theory (see Appendix 5.2). These models correspond to constant and linear trends, i.e.,

$$W_t = \beta_0 + \gamma \left(Y_{t-1} - e^{W_{t-1}}\right) e^{-W_{t-1}}, \tag{7}$$

and

$$W_t = \beta_0 + \beta_1 t/n + \gamma \left(Y_{t-1} - e^{W_{t-1}}\right) e^{-W_{t-1}}\,. \tag{8}$$

Table 1 contains the results for the model given by (7) for two choices of β_0 and γ (denoted δ_1 and δ_2) with a sample size of $n = 250$ and $N = 1000$ replications. A burn-in period of length 100 was used in the simulation of the realizations. Table 2 contains the results for the model given by (8) for two combinations of $(\beta_0, \beta_1, \gamma) = (\delta_1, \delta_2, \delta_3)$, where again the sample size is 250 and the number of replications is 1000. In both tables $\hat{\mu}_{\hat{\delta}_j}$ is the average of the N estimates of δ_j, $\hat{\sigma}_{\hat{\delta}_j}$ is the sample standard deviation of the N estimates of δ_j, $s_{\hat{\delta}_{j,i}}$ is the estimate of the standard error of $\delta_{j,i}$ as computed by (6), and $\hat{\mu}_{s_{\hat{\delta}_j}}$ is the average of the $s_{\hat{\delta}_{j,i}}$, where $\boldsymbol{\delta} = (\boldsymbol{\beta}^T, \gamma)^T$.

In all cases, the "true" parameter value δ_j is very close to the estimated value, $\hat{\mu}_{\hat{\delta}_j}$. A comparison can also be made to evaluate the accuracy of the estimates of standard error. We estimate $V_{j,j}$ as defined in equation (5) by $\hat{\sigma}_{\hat{\delta}_j}$ and compare its value with the average of the N estimates of standard error, $\hat{\mu}_{s_{\hat{\delta}_j}}$. Again, these values are very close, supporting the theory for the maximum likelihood estimated derived in Appendix 5.2.

To further illustrate the theoretical properties of the parameter estimates, Figure 3.2 contains plots of the estimated densities along with the appropriate normal density for one

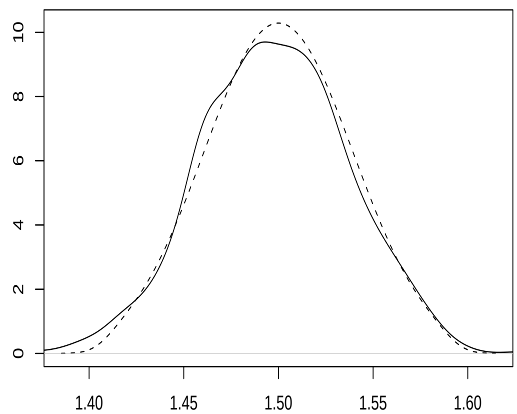
parameters	$\hat{\mu}_{\hat{\delta}_j}$	$\hat{\sigma}_{\hat{\delta}_j}$	$\hat{\mu}_{\hat{\delta}_j} \pm 1.96\hat{\sigma}_{\hat{\delta}_j}/\sqrt{N}$	$\hat{\sigma}_{\hat{\delta}_j}\sqrt{1 \pm 1.96\sqrt{(2/n)}}$	$\hat{\mu}_{s_{\hat{\delta}_j}}$
$\delta_1 = 1.5$	1.4978	0.0387	(1.4954, 1.5001)	(0.0352, 0.0420)	0.0374
$\delta_2 = 0.25$	0.2470	0.0582	(0.2434, 0.2506)	(0.0529, 0.0631)	0.0583
$\delta_1 = 1.5$	1.4990	0.0531	(1.4957, 1.5023)	(0.0483, 0.0576)	0.0660
$\delta_2 = 0.75$	0.7435	0.0386	(0.7411, 0.7459)	(0.0351, 0.0418)	0.0318
$\delta_1 = 3.0$	3.0001	0.0170	(2.9990, 3.0012)	(0.0155, 0.0185)	0.0176
$\delta_2 = 0.25$	0.2483	0.0618	(0.2445, 0.2521)	(0.0562, 0.0670)	0.0613
$\delta_1 = 3.0$	3.0000	0.0252	(2.9984, 3.0015)	(0.0229, 0.0273)	0.0244
$\delta_2 = 0.75$	0.7349	0.0392	(0.7325, 0.7374)	(0.0356, 0.0425)	0.0404

Table 1: Simulations, no trend; n=250, N=1000

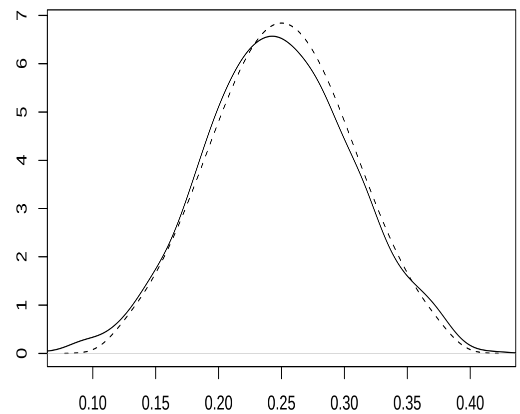
parameters	$\hat{\mu}_{\hat{\delta}_j}$	$\hat{\sigma}_{\hat{\delta}_j}$	$\hat{\mu}_{\hat{\delta}_j} \pm 1.96\hat{\sigma}_{\hat{\delta}_j}/\sqrt{N}$	$\hat{\sigma}_{\hat{\delta}_j}\sqrt{1 \pm 1.96\sqrt{(2/n)}}$	$\hat{\mu}_{s_{\hat{\delta}_j}}$
$\delta_1 = 1$	0.9951	0.1290	(0.9871, 1.0031)	(0.1172, 0.1399)	0.1307
$\delta_2 = 0.5$	0.5044	0.1313	(0.4962, 0.5125)	(0.1193, 0.1424)	0.1324
$\delta_3 = 0.25$	0.2448	0.0593	(0.2411, 0.2485)	(0.0539, 0.0643)	0.0586
$\delta_1 = 1$	0.9887	0.1609	(0.9788, 0.9987)	(0.1461, 0.1744)	0.1669
$\delta_2 = -0.15$	-0.1424	0.1746	(-0.1533, -0.1316)	(0.1586, 0.1893)	0.1784
$\delta_3 = 0.25$	0.2476	0.0559	(0.2441, 0.2510)	(0.0508, 0.0606)	0.0550

Table 2: Simulations, trend; n=250, N=1000

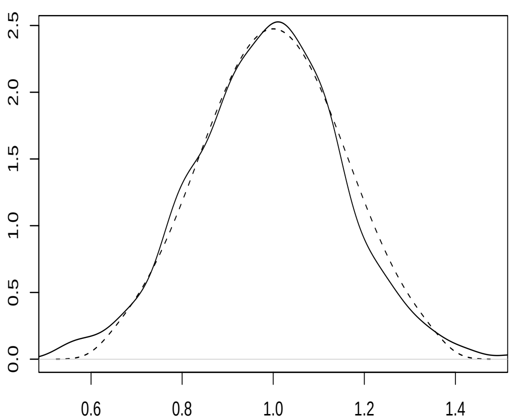
set of parameters from each of the two models. For the first model, the example $\beta_0 = 1.5,\ \gamma = 0.25$ is depicted and for the linear trend model, the example $\beta_0 = 1,\ \beta_1 = -0.15,\ \gamma = 0.25$ is shown. The illustrated normal densities have mean δ_j and variance $\hat{\sigma}_{\hat{\delta}_j}$. As seen from these plots, the estimated and asymptotic densities are in very good agreement for both examples.



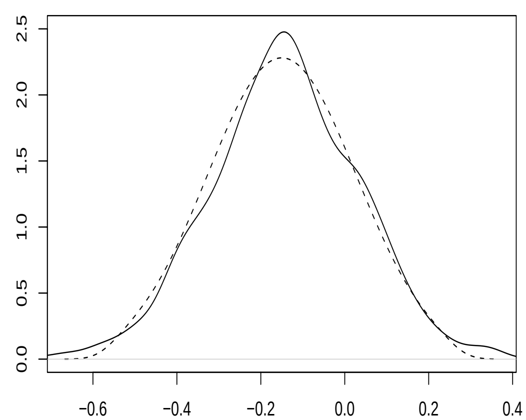
Model (7), $\beta_0 = 1.5$



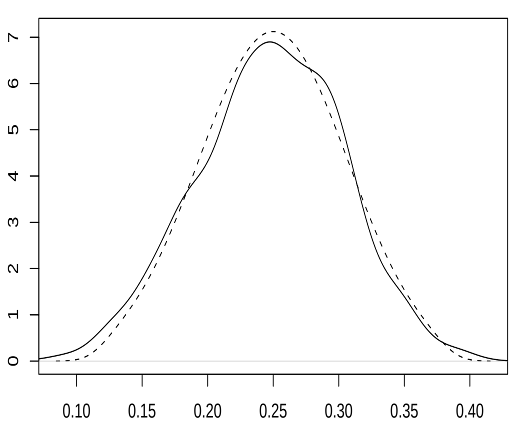
Model (7), $\gamma = 0.25$



Model (8), $\beta_0 = 1$



Model (8), $\beta_1 = -0.15$



Model (8), $\gamma = 0.25$

4 Applications.

4.1 Review of Previous Examples.

Davis et al. (1999) analyze the polio data introduced by Zeger (1988) using various GLARMA models. A summary of other analyses, including those based on parameter driven models, is also given by Davis et al. (1999). In addition, they analyze the UK sudden infant death syndrome series considered by Campbell (1994). Using a GLARMA model and other test statistics for serial dependence in count time series they conclude that the serial dependence effects are not required, so that $p = q = 0$ in the GLARMA model. This conclusion has recently been confirmed by Jung and Liesenfeld (2001) using an approximation to maximum likelihood estimation in a parameter driven model.

Davis et al. (1999, 2000) also report on a preliminary analysis of a series of daily counts of patients presenting at the accident and emergency department of Campbelltown Hospital located in the southwest metropolitan area of Sydney, Australia. Here we extend that analysis with a more comprehensive model for the seasonal effects and the pollution series. This results in a reasonably large number of covariates. This is one of several hospitals where similar analyses can be performed and serves as an illustrative example for which we believe that the use of observation driven models is particularly well suited.

The analysis to be presented here modifies and extends the model considered for the Campbelltown asthma series in Davis et al. (1999). This previous model included explanatory variables for a Sunday effect, a Monday effect, an increasing linear trend in time and a seasonal pattern. The latter was modeled using Fourier series terms consisting of $\cos(2\pi kt/365)$ and $\sin(2\pi kt/365)$ for $k = 1, 2, 3, 4$. To model the remaining serial dependence a GLARMA model with nonzero coefficients at lags 1,3,7 and 10 for the AR component and no moving average component was used. After fitting the model with these terms some slight overdispersion not explained by the lagged AR component of this observation driven model remained. This points to the need for additional covariates or possibly a more flexible seasonal pattern as well as a more complex serial dependence structure.

4.2 Analysis of Sydney Asthma Time Series.

The GLARMA model fit in Davis et al. (1999) did not address a major practical question for which the data was originally collected. Of interest was the role of air pollution levels on the number of daily asthma cases. Because meteorological conditions can be expected to play an important role in the pollution process, temperature can have a direct effect on asthma occurrences, and, because the growth of fungal spores and dust mite level can be related to humidity and temperature, it is reasonable to also consider inclusion of meteorological variables in the model. See Samet et al. (1998) for a discussion of the potential for meteorological conditions to play a role here. At the time of the analysis in Davis et al (1999), only partial data was available on relevant pollution series. Most importantly a series on particulate levels had not been compiled. We now describe in detail the terms investigated in the model and the sources of appropriate data.

4.2.1 Pollution measurements.

The New South Wales EPA provided all available pollution measurements for the Sydney metropolitan region commencing prior to January 1, 1990 (the date at which the asthma data commenced) and up to 1999. Unfortunately, for the time period of our data the network coverage was rather sparse for the southwest region of Sydney. After analyzing all available records for completeness we selected the observations from the Lidcombe observing site for ozone and NO_2 measurements. Lidcombe is to the northeast of Campbelltown but was considered to be sufficiently close so as to give representative readings for these pollutants. The other hospitals we analyzed were at Liverpool and Lidcome and are located even closer to the Lidcombe station. Unfortunately, nepholometer readings of particulate concentrations were not available at Lidcombe during the relevant time period. The two most complete records were at Rozelle and Kensington both of which are considerably closer to downtown Sydney and closer to the coastline. Particulate readings from these two locations were averaged to produce a single series. In addition, two types of pollution series were used: daily average and daily maximum readings based on hourly measurements.

4.2.2 Meteorological Data.

Meteorological data was obtained from the Australian Bureau of Meteorology at Liverpool and was considered to be relevant to the three hospitals considered. While there is spatial variability in meteorological conditions across the Sydney basin the temporal variability (of most relevance to this analysis) is substantially larger and so using a single representative location for these data is reasonable. We were particularly interested in the effects of moisture on asthma levels. In exploratory analysis rainfall did not appear to play a major role, whereas humidity did with an approximate lag of 14 days. Details of the construction and statistical significance of this variable are provided in Davis et al. (2000).

4.2.3 Seasonal Effects

Figure 1 of Davis et al. (2000) shows evidence of a triple peak seasonal pattern in the time series of hospital admission counts. In Davis et al. (1999, 2000) the seasonal variation was modeled using a harmonic regression with several harmonics for the annual frequency. This model may not be appropriate because the intensity of seasonal peaks appears to vary considerably from year to year. In the analysis to follow we propose an alternative representation of the seasonal behavior that allows us to test the constancy from year to year.

The timing of the peaks appears to line up with the terms in the K-12 school year. At that time in Sydney there were four terms per year with the break between terms one and two occurring at varying times due to the timing of the Easter vacation. In the revised seasonal model we assume that there is a broad annual seasonal pattern that is the same in all years and is modeled using annual harmonics $\cos(2\pi t/365)$ and $\sin(2\pi t/365)$. To model the peaks, we used a beta function as follows:

$$p(x) = \frac{1}{B(a,b)}(x)^a(1-x)^b, \quad x \in [0,1]$$

and $p(x) = 0$ if $x \notin [0, 1]$, with $a = 2.5$ and $b = 5$. These parameters were chosen based on a preliminary data analysis comparing the shapes of the peaks in all years and at three locations. Let T_{ij} be the start time for the jth term in year i where time is chosen from $t = 1, \ldots, 1461$, the days in our sample numbered sequentially. Then the peak function for the jth term in year i is

$$P_{ij}(t) = p\left(\frac{t - T_{ij}}{100}\right).$$

In this formulation there are sixteen (one each for four terms in four years) such functions, each of which enters into the regression model with an individual coefficient.

4.2.4 Other Model Terms

Additional explanatory variables used in the model included an overall linear trend over the four year period and the aforementioned indicator variables for Sunday and Monday effects. The inclusion of a linear trend allows for the testing of the hypothesis of increasing asthma rates.

As explained in Davis et al. (2000) there is a clear need in these data (with a fixed annual seasonal cycle) for serial dependence effects. Of interest to us is whether or not inclusion of a more flexible seasonal model, based on the school term peak components, will decrease the serial dependence effects. We investigated a number of options for specifying the autoregressive and moving average effects in the GLARMA model. In the final model a moving average component at lag 7 was all that was required.

4.2.5 Fitting the model

A number of models were fit to the data to investigate the effects of the regression variables summarized above. Based on preliminary analysis it was clear that several variables could be dropped. We retained the two annual harmonic terms, the Sunday and Monday effects, the linear trend, the minimum temperature (same day), the lagged composite humidity variable H_t, and the three same day pollution variables: maximum Ozone, N02 and Particulates. Additionally, we included the sixteen individual term peak components. In this model the term peak components for school terms 3 and 4 (the latter half of the calendar year) were not individually significant at the 5% level and were dropped from the model. We then performed a likelihood ratio test, based on GLARMA likelihoods, for the constancy of the term peak effects for terms 1 and 2 across all four years. The likelihood ratio statistic was 29.9, which assuming an approximating chi-square distribution on 6 degrees of freedom under the null hypothesis, gives a P-value of 0.00004. Accordingly we retained the more flexible representation in which each year allowed variation in the size of the term peaks. Proceeeding from this point the coefficients for Tmin and Trend were not individually significant and were dropped from the model. We next investigated the impact of the pollution variables. The coefficients of the same day values of maximum Ozone and Particulates were not significant (at the 5% level) while that of NO_2 was significant at the 5% level. We also investigated the one day lag effects of the three maximum pollution measurements and none of these were significant. The final model is summarized in Table 3.

Variable	Coefficient	S.E.	T-ratio
Intercept	0.583	0.062	9.46
Sunday	0.197	0.056	3.53
Monday	0.230	0.055	4.20
Annual Cosine	-0.214	0.039	-5.54
Annual Sine	0.176	0.040	4.35
Term 1, 1990	0.200	0.056	3.54
Term 2, 1990	0.132	0.057	2.31
Term 1, 1991	0.087	0.066	1.32
Term 2, 1991	0.172	0.057	2.99
Term 1, 1992	0.254	0.055	4.66
Term 2, 1992	0.308	0.049	6.31
Term 1, 1993	0.439	0.050	8.77
Term 2, 1993	0.116	0.061	1.91
Humidity $H_t/20$	0.169	0.055	3.09
NO_2 max	-0.104	0.033	-3.16
MA, lag 7	0.042	0.018	2.32

Table 3: Analysis of Sydney Asthma Time Series

The fitted values from the model are shown in Figure 2 along with the actual counts. In this final model, various test statistics reviewed in Davis et al. (1999, 2000) for the presence of a latent process and the degree of autocorrelation indicated that there was no need to include additional autoregressive or moving average terms.

4.2.6 Discussion

The expanded and revised model for the Campbelltown asthma series allows for seasonal patterns to be aligned with the school term dates and to vary in intensity from year to year. These differences are highly statistically significant. Virally induced asthma occurrences might be synchronized in part with the school terms and would not necessarily occur with the same intensity in the same terms across different years or across terms in the same year.

The use of a more flexible seasonal model leads to a simplification of the lag dependence structure compared with that in previous analyses. However, moving average dependence at lag 7 is positive and significant. Inferences about the key pollution and weather variables are adjusted for this serial dependence in the above analysis.

The same analysis was repeated at two other locations: Liverpool and Lidcome hospitals. Similar results were obtained for these two sites. However, at these places none of the pollution variables were statistically significant.

5 Appendix

This section provides theoretical complements to Sections 2 and 3. In particular, we establish existence of stationary solutions to the GLARMA model and give a derivation for the

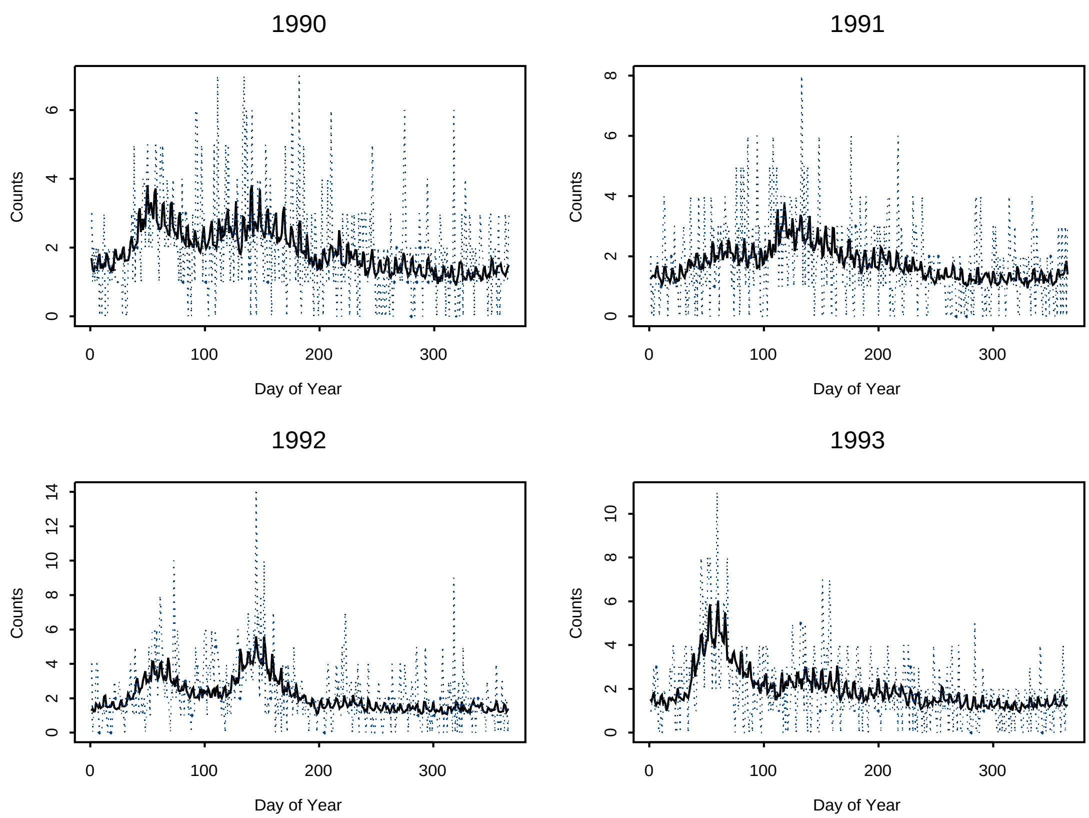


Figure 2: Asthma Counts at Campbelltown Hospital with conditional means.

asymptotic normality of maximum likelihood estimators in some reduced cases.

5.1 Existence of Stationary Solutions.

In this section, we establish the existence of a stationary solution for the process $\{W_t\}$ under the basic model with $1/2 \leq \lambda \leq 1$ and $\mathbf{x}_t\boldsymbol{\beta} = \beta$ given by,

$$W_t = \beta + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}. \tag{9}$$

The result is first stated for a general Markov chain and then shown to hold for the process given by (9) with $1/2 \leq \lambda \leq 1$.

Additionally, for the special case $\lambda = 1$, we will prove that the stationary distribution is unique using techniques from Meyn and Tweedie (1993). The remainder of this section is divided into these two goals.

Existence: $1/2 \leq \lambda \leq 1$.
We begin by stating the existence results for a general Markov chain.

Proposition 5.1 *If* $\{\mathbf{X}_n\}$ *is a weak Feller chain and if for any* $\epsilon > 0$ *there exists a compact set* $C \subset X$ *such that*

$$P(x,\ C^c) < \epsilon, \quad \textit{for all}\ x \in X,$$

then $\{\mathbf{X}_n\}$ *is bounded in probability; thus, there exists at least one stationary distribution for the chain.*

Proof: Assume that for any $\epsilon > 0$ there exists a compact set $C \subset X$ such that $P(x,\ C^c) < \epsilon$ for all $x \in X$. If $P^k(x, \cdot)$ denotes the k-step transition probability of the chain starting from state x then,

$$\begin{aligned} P^k(x,\ C^c) &= \int P(y,\ C^c)P^{k-1}(x, dy) \\ &< \epsilon. \end{aligned}$$

Thus, the chain is bounded in probability. In fact, the tightness of the k-step transition probabilities holds uniformly in x. It follows that the chain is bounded in probability on average and hence, by Theorem 12.0.1(i) of Meyn and Tweedie (1993), there exists a stationary distribution.□

Proposition 5.2 *Let*

$$Y_t \sim \text{Poisson}(e^{W_t}),$$

where

$$W_t = \ \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}},$$

$1/2 \leq \lambda \leq 1$. *Then the chain is bounded in probability, and therefore, admits an invariant measure.*

Proof: First note that the chain is weak Feller. Define $C := [-c,\ c]$. Then,

$$\begin{aligned} P(x,\ C^c) &= P(W_t \in C^c \mid W_{t-1} = x) \\ &= P\big[\gamma\big(Y_{t-1} - e^{W_{t-1}}\big)e^{-\lambda W_{t-1}} \in [-c,\ c]^c \ \big|\ W_{t-1} = x\big], \end{aligned}$$

which, by Markov's inequality,

$$\begin{aligned} &\leq \begin{cases} (\gamma/c)^2 e^{-2\lambda x}\mathrm{Var}\big[Y_t \mid W_{t-1} = x\big], & x \geq 0 \\ (\gamma/c)e^{-\lambda x}E\big[\big|Y_{t-1} - e^x\big| \ \big|\ W_{t-1} = x\big], & x < 0 \end{cases} \\ &\leq \begin{cases} (\gamma/c)^2 e^{(1-2\lambda)x}, & x \geq 0 \\ 2(\gamma/c)e^{(1-\lambda)x}, & x < 0 \end{cases} \\ &\leq \begin{cases} (\gamma/c)^2, & x \geq 0 \\ 2(\gamma/c), & x < 0. \end{cases} \end{aligned}$$

Thus, given $\epsilon > 0$ choose c large such that $\max\big(2(\gamma/c), (\gamma/c)^2\big) < \epsilon$. The result follows from Proposition 5.1.□

Uniqueness: $\lambda = 1$.
Under this assumption on λ, we are able to establish uniqueness of the stationary distribution. To accomplish this we shall show that the process $\{W_t\}$ is aperiodic and satisfies Doeblin's condition. It then follows from Theorem 16.2.3 of Meyn and Tweedie (1993) that $\{W_t\}$ is uniformly ergodic. We first begin with a statement of Doeblin's condition:

There exists a probability measure ν with the property that for some $m \geq 1, \epsilon > 0$, and $\delta > 0$

$$\nu(B) > \epsilon \Longrightarrow P^m(x, B) \geq \delta, \tag{10}$$

for every $x \in X$.

Proposition 5.3 *The process $\{W_t\}$ given in equation (9) satisfies Doeblin's condition and is strongly aperiodic. Hence, the process is uniformly ergodic.*

Proof: In order to establish Doeblin's Condition, we consider the two cases $\gamma < 0$ and $\gamma > 0$.
Case 1: $\gamma < 0$
¿From (9) with $\lambda = 1$, one can see that $W_t = \beta - \gamma + \gamma Y_{t-1}e^{W_{t-1}} \leq \beta - \gamma$. Define the measure ν to have unit point mass at $\{\beta - \gamma\}$. In order to verify (10), it suffices to only consider Borel sets B with $\beta - \gamma \in B$. Then, for all $x \leq \beta - \gamma$,

$$\begin{aligned} P(x, B) &= P\left(W_t \in B \mid W_{t-1} = x\right) \\ &\geq P\left(W_t = \beta - \gamma \mid W_{t-1} = x\right) \\ &= P\left(Y_{t-1} = 0 \mid W_{t-1} = x\right) \\ &= e^{-e^x} \\ &\geq e^{-e^{\beta-\gamma}}, \end{aligned}$$

which yields (10) with $m = 1$.
Case 2: $\gamma > 0$
For $\gamma > 0$, W_t has a lower bound of $\beta - \gamma$ and hence, the state space for W_t is a subset of

$[\beta-\gamma,\infty)$. As in Case 1, we will take the measure ν to have unit mass at $\{\beta-\gamma\}$. Let $C=[\beta-\gamma,\max(\epsilon,\beta+\gamma)]$, where $\epsilon>0$. Then, for all $x\in C$ and Borel sets B containing $\beta-\gamma$,

$$\begin{aligned}
P(x,B) &= P\left(W_t\in B \mid W_{t-1}=x\right)\\
&\geq P\left(W_t=\beta-\gamma \mid W_{t-1}=x\right)\\
&= P\left(Y_{t-1}=0 \mid W_{t-1}=x\right)\\
&= e^{-e^x}\\
&\geq e^{-e^{\max(\epsilon,\beta+\gamma)}} := \delta_1,
\end{aligned}$$

and

$$\begin{aligned}
P^2(x,B) &\geq P(W_{t+1}=\beta-\gamma, W_t=\beta-\gamma \mid W_{t-1}=x)\\
&\geq \delta_1^2.
\end{aligned}$$

On the other hand if $x\notin C$, then $x>\max(\epsilon,\beta+\gamma)$ and we have

$$\begin{aligned}
P(x,C) &= P\left(W_t\in C \mid W_{t-1}=x\right)\\
&\geq P\left(\beta-\gamma\leq W_t\leq\beta+\gamma \mid W_{t-1}=x\right)\\
&= P\left(|W_t-\beta|\leq\gamma \mid W_{t-1}=x\right)\\
&\geq 1-\gamma^{-2}Var\left(W_t \mid W_{t-1}=x\right) \text{ (by Chebyshev's Inequality)}\\
&= 1-\gamma^{-2}\gamma^2 e^{-x}\\
&\geq 1-e^{-\max(\epsilon,\beta+\gamma)} := \delta_2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
P^2(x,B) &= P(W_t\in B \mid W_{t-2}=x)\\
&\geq P(W_t\in B, W_{t-1}\in C \mid W_{t-2}=x)\\
&= \sum_{y\in C} P(W_t\in B, W_{t-1}=y \mid W_{t-2}=x)\\
&= \sum_{y\in C} P(W_t\in B \mid W_{t-1}=y)P(W_{t-1}=y \mid W_{t-2}=x)\\
&\geq \delta_1 P(W_{t-1}\in C \mid W_{t-2}=x)\\
&\geq \delta_1\delta_2.
\end{aligned}$$

Thus, Doeblin's condition is also satisfied for the case $\gamma\in(0,1]$.

For both cases, i.e., $0<|\gamma|\leq 1$, the chain $\{W_t\}$ is also strongly aperiodic since

$$\begin{aligned}
P(\beta-\gamma,\beta-\gamma) &= P(W_t=\beta-\gamma \mid W_{t-1}=\beta-\gamma)\\
&= P(Y_{t-1}=0 \mid W_{t-1}=\beta-\gamma)\\
&= e^{-e^{\beta-\gamma}}\\
&> 0.
\end{aligned}$$

As remarked earlier, we conclude that $\{W_t\}$ must be uniformly ergodic. $\square$

The result given above for $\lambda = 1$ extends to the case

$$W_t = \beta + \sum_{i=1}^{p} \gamma_i (Y_{t-i} - e^{W_{t-i}}) e^{-W_{t-i}}$$

by considering the $p-$variate Markov chain $(W_t, W_{t-1}, \ldots W_{t-p+1})$.

5.2 Maximum Likelihood Calculations

In this section, we derive the remaining expressions needed for the maximum likelihood calculations of Section 3.1. Recall,

$$\frac{\partial L}{\partial \boldsymbol{\delta}} = \sum_{t=1}^{n} (Y_t - \mu_t) \frac{\partial W_t}{\partial \boldsymbol{\delta}} = \sum_{t=1}^{n} e_t \mu_t^{\lambda} \frac{\partial W_t}{\partial \boldsymbol{\delta}},$$

and

$$\frac{\partial^2 L}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} = \sum_{t=1}^{n} \left[e_t \mu_t^{\lambda} \frac{\partial^2 W_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} - \mu_t \frac{\partial W_t}{\partial \boldsymbol{\delta}} \frac{\partial W_t}{\partial \boldsymbol{\delta}^T} \right].$$

First we note that

$$\frac{\partial e_t}{\partial \boldsymbol{\delta}} = -[e^{(1-\lambda)W_t} + \lambda e_t] \frac{\partial W_t}{\partial \boldsymbol{\delta}}.$$

Also

$$\frac{\partial W_t}{\partial \boldsymbol{\delta}} = \mathbf{x}_t^T \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\delta}} + \frac{\partial Z_t}{\partial \boldsymbol{\delta}},$$

where

$$\begin{aligned} Z_t &= \sum_{i=1}^{\infty} \tau_i e_{t-i} \\ &= (\phi(B)^{-1} \theta(B) - 1) e_t, \end{aligned}$$

so that

$$Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q} \theta_i e_{t-i}.$$

It follows that

$$\begin{aligned} \frac{\partial Z_t}{\partial \boldsymbol{\delta}} &= \sum_{i=1}^{p} \frac{\partial \phi_i}{\partial \boldsymbol{\delta}} (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{p} \phi_i \left(\frac{\partial Z_{t-i}}{\partial \boldsymbol{\delta}} + \frac{\partial e_{t-i}}{\partial \boldsymbol{\delta}} \right) \\ &\quad + \sum_{i=1}^{q} \frac{\partial \theta_i}{\partial \boldsymbol{\delta}} e_{t-i} + \sum_{i=1}^{q} \theta_i \frac{\partial e_{t-i}}{\partial \boldsymbol{\delta}}. \end{aligned}$$

In particular:

$$\frac{\partial Z_t}{\partial \beta_a} = \sum_{i=1}^{p} \phi_i \left(\frac{\partial Z_{t-i}}{\partial \beta_a} + \frac{\partial e_{t-i}}{\partial \beta_a} \right) + \sum_{i=1}^{q} \theta_i \frac{\partial e_{t-i}}{\partial \beta_a},$$

$$\frac{\partial Z_t}{\partial \phi_a} = Z_{t-a} + e_{t-a} + \sum_{i=1}^{p} \phi_i \left(\frac{\partial Z_{t-i}}{\partial \phi_a} + \frac{\partial e_{t-i}}{\partial \phi_a} \right) + \sum_{i=1}^{q} \theta_i \frac{\partial e_{t-i}}{\partial \phi_a}$$

and

$$\frac{\partial Z_t}{\partial \theta_a} = \sum_{i=1}^{p} \phi_i \left(\frac{\partial Z_{t-i}}{\partial \theta_a} + \frac{\partial e_{t-i}}{\partial \theta_a} \right) + e_{t-a} + \sum_{i=1}^{q} \theta_i \frac{\partial e_{t-i}}{\partial \theta_a}.$$

The second derivatives are then

$$\begin{aligned} \frac{\partial^2 e_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} &= -[e^{(1-\lambda)W_t} + \lambda e_t] \frac{\partial^2 W_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} \\ & \quad - \left[\frac{\partial W_t}{\partial \boldsymbol{\delta}} (1-\lambda) e^{(1-\lambda)W_t} + \lambda \frac{\partial e_t}{\partial \boldsymbol{\delta}} \right] \frac{\partial W_t}{\partial \boldsymbol{\delta}^T} \end{aligned}$$

and

$$\frac{\partial^2 W_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} = \frac{\partial^2 \beta^T}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} x_t + \frac{\partial \boldsymbol{\delta}^2 Z_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} = \frac{\partial \boldsymbol{\delta}^2 Z_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T},$$

in which

$$\begin{aligned} \frac{\partial^2 Z_t}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} &= \sum_{i=1}^{p} \left[\frac{\partial \phi_i}{\partial \boldsymbol{\delta}} \left(\frac{\partial Z_{t-i}}{\partial \boldsymbol{\delta}^T} + \frac{\partial e_{t-i}}{\partial \boldsymbol{\delta}^T} \right) + \left(\frac{\partial Z_{t-i}}{\partial \boldsymbol{\delta}} + \frac{\partial e_{t-i}}{\partial \boldsymbol{\delta}} \right) \frac{\partial \phi_i}{\partial \boldsymbol{\delta}^T} \right] \\ & \quad + \sum_{i=1}^{p} \phi_i \left(\frac{\partial^2 Z_{t-i}}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} + \frac{\partial^2 e_{t-i}}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} \right) + \sum_{i=1}^{q} \left[\frac{\partial \theta_i}{\partial \boldsymbol{\delta}} \frac{\partial e_{t-i}}{\partial \boldsymbol{\delta}^T} + \frac{\partial e_{t-i}}{\partial \boldsymbol{\delta}} \frac{\partial \theta_i}{\partial \boldsymbol{\delta}^T} \right] \\ & \quad + \sum_{i=1}^{q} \theta_i \frac{\partial^2 e_{t-i}}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T}. \end{aligned}$$

5.3 Asymptotic Distribution of MLE

In this section we establish asymptotic properties of the MLEs given in Section 3.1 for the specific case of the basic model with $\lambda = 1, p = 1$ and $\mathbf{x}_t^T \boldsymbol{\beta} = \beta$. Uniform ergodicity (as established in Proposition 5.3) and stationarity of $\{W_t\}$ are the key ingredients of the argument.

First replace $W_t(\boldsymbol{\delta})$ by

$$W_t^{\dagger}(\boldsymbol{\delta}) = W_t(\boldsymbol{\delta}_0) + (\boldsymbol{\delta} - \boldsymbol{\delta}_0)^T \dot{W}_t,$$

where $\dot{W}_t = \frac{\partial W_t(\boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}}$ and define a linearized form of the likelihood as

$$L^{\dagger}(\boldsymbol{\delta}) = \sum_{t=1}^{n} \left(Y_t W_t^{\dagger}(\boldsymbol{\delta}) - e^{W_t^{\dagger}(\boldsymbol{\delta})} \right).$$

Unless otherwise indicated, W_t and $\dot{W}_t$ are evaluated at $\boldsymbol{\delta}_0$. Now, re-parameterizing with the transformation $u = n^{1/2}(\boldsymbol{\delta} - \boldsymbol{\delta}_0)$, we have

$$
\begin{aligned}
R_n^\dagger(u) &:= L^\dagger(\boldsymbol{\delta}_0) - L^\dagger(\boldsymbol{\delta}_0 + un^{-1/2}) \\
&= -u^T n^{-1/2} \sum_{t=1}^n Y_t \dot{W}_t + \sum_{t=1}^n e^{W_t} \left(e^{u^T n^{-1/2} \dot{W}_t} - 1 \right) \\
&= -u^T n^{-1/2} \sum_{t=1}^n \left(Y_t - e^{W_t} \right) \dot{W}_t + \sum_{t=1}^n e^{W_t} \left(e^{u^T n^{-1/2} \dot{W}_t} - 1 - u^T n^{-1/2} \dot{W}_t \right). \quad (11)
\end{aligned}
$$

Note that $R_n^\dagger(u)$ is a convex function of u. The first term in (11) can be written as $-u^T H_n$ where

$$H_n := n^{-1/2} \sum_{t=1}^n e_t e^{W_t} \dot{W}_t$$

and $e_t = \left(Y_t - e^{W_t} \right) / e^{W_t}$. Now this is a sum of a triangular array of vector martingale differences

$$\eta_{nt} = n^{-1/2} e_t b_t,$$

where

$$b_t = \dot{W}_t e^{W_t} = \dot{W}_t \mu_t.$$

In order to apply a martingale central limit theorem, it suffices to show (see Corollary 3.1 of Hall and Heyde [8]) that

$$\sum_{t=1}^n E(\eta_{nt} \eta_{nt}^T \mid \mathcal{F}_{t-1}) \xrightarrow{P} V(\boldsymbol{\delta}_0), \quad (12)$$

where $\mathcal{F}_t = \sigma(Y_s, s \leq t)$, and, for all $\epsilon > 0$,

$$\sum_{t=1}^n E\left(\eta_{nt} \eta_{nt}^T I(|\eta_{nt}| > \epsilon) \mid \mathcal{F}_{t-1} \right) \xrightarrow{P} 0. \quad (13)$$

We then have

$$H_n \xrightarrow{d} N(0, V),$$

where

$$V = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^n \frac{\partial L_t(\boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}} \frac{\partial L_t(\boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}^T} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^n e_t^2 e^{2W_t} \dot{W}_t \dot{W}_t^T.$$

The second term in (11) is

$$u^T \left[(2n)^{-1} \sum_{t=1}^n e^{W_t} \dot{W}_t \dot{W}_t^T \right] u + O_p \left(n^{-3/2} \sum_{t=1}^n e^{W_t} (u^T \dot{W}_t)^3 \right)$$

in which the second term converges to zero. Hence

$$R_n^\dagger(u) \xrightarrow{d} R(u) := -u^T N(0, V) + u^T V u / 2,$$

where

$$V = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} e^{W_t} \dot{W}_t \dot{W}_t^T.$$

It then follows that $\hat{u}_n^\dagger$ =argmin$R_n^\dagger(u) \xrightarrow{d} \hat{u}$ =argmin$R(u)$. From the form of $R(u)$, we see that $\hat{u} = V^{-1}N(0,V) \sim N(0,V^{-1})$.

Next, we pass the convergence of $R_n^\dagger(u)$ onto $R_n(u) := L(\boldsymbol{\delta}_0) - L(un^{-1/2} + \boldsymbol{\delta}_0)$. Specifically, it suffices to show that $L(un^{-1/2} + \boldsymbol{\delta}_0) - L^\dagger(un^{-1/2} + \boldsymbol{\delta}_0) \xrightarrow{P} 0$ uniformly for $|u| \le K$. Writing $\boldsymbol{\delta} = un^{1/2} + \boldsymbol{\delta}_0$, we have

$$\begin{aligned}
& L(\boldsymbol{\delta}) - L^\dagger(\boldsymbol{\delta}) \\
= & \sum_{t=1}^{n} Y_t W_t(\boldsymbol{\delta}) - \sum_{t=1}^{n} e^{W_t(\boldsymbol{\delta})} - \sum_{t=1}^{n} Y_t \left(W_t + u^T n^{-1/2} \dot{W}_t \right) + \sum_{t=1}^{n} e^{W_t + u^T n^{-1/2} \dot{W}_t} \\
= & \sum_{t=1}^{n} Y_t \left(W_t(\boldsymbol{\delta}) - W_t - u^T n^{-1/2} \dot{W}_t \right) - \sum_{t=1}^{n} \left(e^{W_t(\boldsymbol{\delta})} - e^{W_t + u^T n^{-1/2} \dot{W}_t} \right) \\
= & \sum_{t=1}^{n} \left(Y_t - e^{W_t} \right) \left(W_t(\boldsymbol{\delta}) - W_t - u^T n^{-1/2} \dot{W}_t \right) \\
& - \sum_{t=1}^{n} \left[e^{W_t(\boldsymbol{\delta})} - e^{W_t + u^T n^{-1/2} \dot{W}_t} - e^{W_t} \left(W_t(\boldsymbol{\delta}) - W_t - u^T n^{-1/2} \dot{W}_t \right) \right]. \qquad (14)
\end{aligned}$$

The first term in equation (14) is

$$\begin{aligned}
A_n &= \sum_{t=1}^{n} \left(Y_t - e^{W_t} \right) \left(W_t(\boldsymbol{\delta}) - W_t - u^T n^{-1/2} \dot{W}_t \right) \\
&= u^T (2n)^{-1} \left[\sum_{t=1}^{n} \left(Y_t - e^{W_t} \right) \ddot{W}_t + \sum_{t=1}^{n} \left(Y_t - e^{W_t} \right) \left(\ddot{W}_t(\boldsymbol{\delta}^*) - \ddot{W}_t \right) \right] u.
\end{aligned}$$

Since $(Y_t - e^{W_t})\ddot{W}_t$ is stationary and $E[(Y_t - e^{W_t})\ddot{W}_t] = 0$, $A_n \to 0$ uniformly for $|u| \le K$ and for all $K < \infty$, where $\|\boldsymbol{\delta}^* - \boldsymbol{\delta}_0\| \le \|\boldsymbol{\delta} - \boldsymbol{\delta}_0\|$ assuming $\ddot{W}_t(\boldsymbol{\delta}^*) - \ddot{W}_t \xrightarrow{P} 0$. The second term is

$$B_n = -\sum_{t=1}^{n} \left[e^{W_t(\boldsymbol{\delta})} - e^{W_t + u^T n^{-1/2} \dot{W}_t} - e^{W_t} \left(W_t(\boldsymbol{\delta}) - W_t - u^T n^{-1/2} \dot{W}_t \right) \right],$$

which after expanding $e^{W_t(\boldsymbol{\delta})}, e^{u^T n^{-1/2}\dot{W}_t}$ and $W_t(\boldsymbol{\delta})$ in a Taylor series is

$$\begin{aligned}
&= -\sum_{t=1}^{n}\Big[e^{W_t}+u^T n^{-1/2}e^{W_t}\dot{W}_t+u^T(2n)^{-1}e^{W_t(\boldsymbol{\delta}_1^*)}\left(\dot{W}_t^2(\boldsymbol{\delta}_1^*)+\ddot{W}_t(\boldsymbol{\delta}_1^*)\right)u\\
&\quad -e^{W_t}\left(1+u^T n^{-1/2}\dot{W}_t+e^{c}(2n)^{-1}u^T\dot{W}_t^2 u\right)\\
&\quad -e^{W_t}\left(W_t+u^T n^{-1/2}\dot{W}_t+u^T(2n)^{-1}\ddot{W}_t(\boldsymbol{\delta}_2^*)u-W_t-u^T n^{-1/2}\dot{W}_t\right)\Big]\\
&= -u^T(2n)^{-1}\Bigg\{\sum_{t=1}^{n}\left(e^{W_t(\boldsymbol{\delta}_1^*)}-e^{W_t}\right)\left(\dot{W}_t^2(\boldsymbol{\delta}_1^*)+\ddot{W}_t(\boldsymbol{\delta}_1^*)\right)\\
&\quad +\sum_{t=1}^{n}e^{W_t}\left[\left(\dot{W}_t^2(\boldsymbol{\delta}_1^*)-e^{c}\dot{W}_t^2\right)+\left(\ddot{W}_t(\boldsymbol{\delta}_1^*)-\ddot{W}_t(\boldsymbol{\delta}_2^*)\right)\right]\Bigg\}u,
\end{aligned}$$

where $0\le c\le \frac{u^T}{2n}\dot{W}_t(\boldsymbol{\delta}_0)$ and $\left\|\boldsymbol{\delta}_j^*-\boldsymbol{\delta}_0\right\|\le\|\boldsymbol{\delta}-\boldsymbol{\delta}_0\|$ for $j=1,2$. Assuming each average in the above expression converges to a finite quantity in probability, we have that $B_n\to 0$ uniformly on compact subsets for u. Therefore, $L(\boldsymbol{\delta})-L^{\dagger}(\boldsymbol{\delta})\xrightarrow{P}0$ uniformly for $|u|\le K$, for all $K<\infty$ and we obtain the desired result:

$$R_n(u)\xrightarrow{d}R(u):=-u^T N(0,V)+u^T V u/2.$$

We now consider establishing conditions (12) and (13). ¿From (9) we see that

$$\begin{aligned}
\dot{W}_t &= \begin{bmatrix}\frac{\partial W_t}{\partial\gamma}\\ \frac{\partial W_t}{\partial\beta}\end{bmatrix}=\begin{bmatrix}\dot{W}_{t,1}\\ \dot{W}_{t,2}\end{bmatrix}\\
&= \begin{bmatrix}Y_{t-1}e^{-W_{t-1}}-1-\gamma Y_{t-1}e^{-W_{t-1}}\dot{W}_{t-1,1}\\ 1-\gamma y_{t-1}e^{-W_{t-1}}\dot{W}_{t-1,2}\end{bmatrix}\\
&= \begin{bmatrix}U_t+A_t\dot{W}_{t-1,1}\\ 1+A_t\dot{W}_{t-1,2}\end{bmatrix}=\begin{bmatrix}U_t+\sum_{i=1}^{\infty}A_t\cdots A_{t-i+1}U_{t-i}\\ 1+\sum_{i=1}^{\infty}A_t\cdots A_{t-i+1}\end{bmatrix},
\end{aligned}\tag{15}$$

where $U_t=Y_{t-1}e^{-W_{t-1}}-1$ and $A_t=-\gamma Y_{t-1}e^{-W_{t-1}}$. Since $\dot{W}_t$ is a function of $\{W_s, s\le t\}$, it also is a strictly stationary ergodic process. Now,

$$\sum_{t=1}^{n}E(\eta_{nt}\eta_{nt}^T\mid\mathcal{F}_{t-1})=\frac{1}{n}\sum_{t=1}^{n}e^{W_t}\dot{W}_t\dot{W}_t^T,$$

which is a function of two stationary ergodic processes, $\{W_t\}$ and $\{\dot{W}_t\}$. By the ergodic theorem we then have

$$\frac{1}{n}\sum_{t=1}^{n}e^{W_t}\dot{W}_t\dot{W}_t^T\xrightarrow{a.s.}V=E(e^{W_1}\dot{W}_1\dot{W}_1^T)$$

if $E|e^{W_t(\boldsymbol{\delta}_0)}\dot{W}_t\dot{W}_t^T|<\infty$. Conditions under which this holds will now be derived for a particular choice of parameter values of β and γ. It suffices to show $E|e^{W_t}\dot{W}_{t,i}^2|<\infty,\ i=1,2$. First we will consider the case $i=1$. Using $\|\cdot\|_2$ to denote the L_2 norm, we have from (15),

$$\|e^{W_t/2}\dot{W}_{t,1}\|_2\le\|e^{W_t/2}U_t\|_2+\sum_{i=1}^{\infty}\|e^{W_t/2}A_t\cdots A_{t-i+1}U_{t-i}\|_2.$$

Using properties of the moment generating function for a Poisson distributed random variable and the fact that the process W_t is bounded below by $\beta - \gamma$, we have

$$\begin{aligned}
\|e^{W_t/2}U_t\|_2^2 &= E\left[e^{\beta-\gamma}e^{\gamma Y_{t-1}e^{-W_{t-1}}}(Y_{t-1}e^{-W_{t-1}}-1)^2\right] \\
&= e^{\beta-\gamma}E\left[E\left(e^{\gamma Y_{t-1}e^{-W_{t-1}}}(Y_{t-1}^2e^{-2W_{t-1}}-2Y_{t-1}e^{-W_{t-1}}+1)\mid W_{t-1}\right)\right] \\
&= e^{\beta-\gamma}E\Big[e^{-W_{t-1}}e^{\gamma e^{-W_{t-1}}}e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)}+e^{2\gamma e^{-W_{t-1}}}e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)} \\
&\qquad -2e^{\gamma e^{-W_{t-1}}}e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)}+e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)}\Big] \\
&= e^{\beta-\gamma}E\left[e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)}\left[1+e^{\gamma e^{-W_{t-1}}}\left(e^{\gamma e^{-W_{t-1}}}+e^{-W_{t-1}}-2\right)\right]\right] \\
&\leq e^{\beta-\gamma}e^{e^{\beta-\gamma}(e^{\gamma e^{-(\beta-\gamma)}}-1)}\left[1+e^{\gamma e^{-(\beta-\gamma)}}\left(e^{\gamma e^{-(\beta-\gamma)}}+e^{-(\beta-\gamma)}-2\right)\right] := c_1^2, \\
E\left[e^{W_t}A_t^2\mid\mathcal{F}_{t-1}\right] &= E\left[\gamma^2e^{\beta-\gamma}Y_{t-1}^2e^{-2W_{t-1}}e^{\gamma Y_{t-1}e^{-W_{t-1}}}\mid W_{t-1}\right] \\
&= \gamma^2e^{\beta-\gamma}\left[e^{-W_{t-1}}e^{\gamma e^{-W_{t-1}}}e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)}+e^{2\gamma e^{-W_{t-1}}}e^{e^{W_{t-1}}(e^{\gamma e^{-W_{t-1}}}-1)}\right] \\
&\leq \gamma^2e^{\beta-\gamma}e^{e^{\beta-\gamma}(e^{\gamma e^{-(\beta-\gamma)}}-1)}e^{\gamma e^{-(\beta-\gamma)}}\left(e^{\gamma e^{-(\beta-\gamma)}}+e^{-(\beta-\gamma)}\right) := b_1^2, \\
E\left[A_t^2\mid\mathcal{F}_{t-1}\right] &= E\left[\gamma^2Y_{t-1}^2e^{-2W_{t-1}}\mid W_{t-1}\right] \\
&= \gamma^2\left(1+e^{-W_{t-1}}\right) \\
&\leq \gamma^2(1+e^{-(\beta-\gamma)})
\end{aligned}$$

and

$$\begin{aligned}
E\left[U_t^2\mid\mathcal{F}_{t-1}\right] &= E\left[Y_{t-1}^2e^{-2W_{t-1}}-2Y_{t-1}e^{-W_{t-1}}+1\mid W_{t-1}\right] \\
&= e^{-W_{t-1}} \\
&\leq e^{-(\beta-\gamma)} := b_2^2.
\end{aligned}$$

Applying these results, $\|e^{W_t/2}A_t\cdots A_{t-i+1}U_{t-i}\|_2^2$ may be calculated recursively:

$$\begin{aligned}
\|e^{W_t/2}A_t\cdots A_{t-i+1}U_{t-i}\|_2^2 &= E\left(e^{W_t}A_t^2\cdots A_{t-i+1}^2U_{t-i}^2\right) \\
&= E\left[E\left(e^{W_t}A_t^2\cdots A_{t-i+1}^2U_{t-i}^2\mid\mathcal{F}_{t-1}\right)\right] \\
&= E\left[A_{t-1}^2\cdots A_{t-i+1}^2U_{t-i}^2E\left(e^{W_t}A_t^2\mid\mathcal{F}_{t-1}\right)\right] \\
&\leq b_1^2E\left[E\left(A_{t-1}^2\cdots A_{t-i+1}^2U_{t-i}^2\mid\mathcal{F}_{t-2}\right)\right] \\
&= b_1^2E\left[A_{t-2}^2\cdots A_{t-i+1}^2U_{t-i}^2E\left(A_{t-1}^2\mid\mathcal{F}_{t-2}\right)\right] \\
&\leq b_1^2\gamma^2(1+e^{-(\beta-\gamma)})E\left[A_{t-2}^2\cdots A_{t-i+1}^2U_{t-i}^2\right] \\
&\vdots \\
&\leq b_1^2\left(\gamma^2(1+e^{-(\beta-\gamma)})\right)^{i-1}E\left[E\left(U_{t-i}\mid\mathcal{F}_{t-i-1}\right)\right] \\
&\leq b_1^2b_2^2\left(\gamma^2(1+e^{-(\beta-\gamma)})\right)^{i-1}.
\end{aligned}$$

Therefore,

$$\|e^{W_t/2}\dot{W}_{t,1}\|_2 \leq c_1 + c_2 \sum_{i=1}^{\infty} \gamma^{i-1}(1+e^{\gamma-\beta})^{(i-1)/2},$$

where $c_2 = b_1 b_2$.

Likewise,

$$\begin{aligned}\|e^{W_t/2}\dot{W}_{t,2}\|_2 &\leq \|e^{W_t/2}\|_2 + \sum_{i=0}^{\infty} \|e^{W_t/2}A_t \cdots A_{t-i}\|_2 \\ &\leq c_3 + c_4 \sum_{i=1}^{\infty} \gamma^{i-1}(1+e^{\gamma-\beta})^{(i-1)/2},\end{aligned}$$

where $c_3 = \left[e^{\beta-\gamma}e^{e^{\beta-\gamma}(e^{\gamma e^{-(\beta-\gamma)}}-1)}\right]^{1/2}$ and $c_4 = \left[\gamma^2 e^{\beta-\gamma}e^{e^{\beta-\gamma}(e^{\gamma e^{-(\beta-\gamma)}}-1)}e^{\gamma e^{-(\beta-\gamma)}}\left(e^{\gamma e^{-(\beta-\gamma)}}+e^{-(\beta-\gamma)}\right)\right]^{1/2}$. Therefore, $E|e^{W_t}\dot{W}_t\dot{W}_t^T|$ will be finite for $\gamma(1+e^{\gamma-\beta})^{1/2} < 1$.

The convergence required in condition (13) is easily established using condition (12) and the stationarity of $\{W_t\}$. Now,

$$\sum_{t=1}^{n} E\left(\eta_{nt}\eta_{nt}^T I(|\eta_{nt} > \epsilon)||\mathcal{F}_{t-1}\right)$$

$$\begin{aligned}&= \frac{1}{n}\sum_{t=1}^{n} E\left[(Y_{t-1}-e^{W_{t-1}})^2\dot{W}_t\dot{W}_t^T I(|(Y_{t-1}-e^{W_{t-1}})\dot{W}_t| > \epsilon\sqrt{n})|\mathcal{F}_{t-1}\right] \\ &\leq \frac{1}{n}\sum_{t=1}^{n} E\left[(Y_{t-1}-e^{W_{t-1}})^2\dot{W}_t\dot{W}_t^T I(|(Y_{t-1}-e^{W_{t-1}})\dot{W}_t| > M)|\mathcal{F}_{t-1}\right] \\ &\overset{n\to\infty}{\longrightarrow} E\left[(Y_1-e^{W_1})^2\dot{W}_1\dot{W}_1^T I(|(Y_1-e^{W_1})\dot{W}_1| > M)\right] \\ &\longrightarrow 0 \text{ as } M\to\infty.\end{aligned}$$

Therefore, the asymptotic distribution of the maximum likelihood estimates is $N(0, V^{-1})$ where

$$V = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} e^{W_t(\boldsymbol{\delta}_0)}\dot{W}_t\dot{W}_t^T.$$

References

[1] Campbell, M. J. (1994) "Time series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature", J. R. Statist. Soc. A, 157, 191-208.

[2] Cox, D.R. (1981). "Statistical Analysis of Time Series: Some Recent Developments." *Scandinavian Journal of Statistics*, **8**, 93–115.

[3] Davis, R.A., Dunsmuir, W.T.M., and Wang, Y. (1999). "Modelling Time Series of Count Data." *Asymptotics, Nonparametrics, and Time Series* (Subir Ghosh, editor) Marcel-Dekker, New York, 63–114.

[4] Davis, R.A., Dunsmuir, W.T.M., and Wang, Y. (2000). "On Autocorrelation in a Poisson Regression Model" (with W.T.M. Dunsmuir and Y. Wang). *Biometrika* **87** 491–506.

[5] Diggle, P.J., K-Y Liang and S.L. Zeger (1994) Analysis of Longitudinal Data, Oxford University Press, Oxford.

[6] Durbin, J. and Koopman, S.J. (2000) "Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesain perspectives", *J. R. Statist. Soc. B* **62**, p3–56.

[7] Fahrmeir, L. and Tutz, G. (1994) Multivariate Statistical Modeling Based on Generalised Linear Models, Springer-Verlag, New York.

[8] Hall, P. and Heyde, C.C. (1980) "Martingale limit theory and its application", Academic Press, New York.

[9] Jung, R.C. and Liesenfeld, R. (2001). "Estimating time series models for count data using efficient importance sampling." (To appear in *Allgemeines Statistisches Archive*, **85**.)

[10] Meyn, S.P. and Tweedie, R.L. (1994). "Markov Chains and Stochastic Stability". Springer-Verlag, New York.

[11] Rydberg, T.H. and Shephard, N. (1999). "BIN models for trade-by-trade data. Modelling the number of trades in a fixed interval of time." (Working paper, Oxford University).

[12] Samet, J., Zeger, S., Kelsall, J., Xu, J., and Kalkstein, L. (1988). "Does weather confound or modify the association of particulate air pollution with mortality? An analysis of the Philadelphia data, 1973-1980." Environ Res 77;1: 9-19

[13] Shephard, Neil (1995) "Generalized Linear Autoregressions". (Unpublished paper, Oxford University)

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

[15] Zeger, S. L. and Qaqish, B. (1988) "Markov regression models for time series: a quasi-likelihood approach", Biometrics **44**, 1019-1031.