Count Time Series Models

Konstantinos Fokianos

Department of Mathematics & Statistics, University of Cyprus
e-mail: fokianos@ucy.ac.cy

To appear in Handbook of Time Series Analysis
February 28th, 2011
Revised: 30th May 2011

Abstract
We review regression models for count time series. We discuss the approach which is based on generalized linear models and the class of integer autoregressive processes. The generalized linear models framework provides convenient tools for implementing model fitting and prediction using standard software. Furthermore, this approach provides a natural extension to the traditional ARMA methodology. Several models have been developed along these lines, but conditions for stationarity and valid asymptotic inference were given in the literature only recently. We review several of these facts. In addition, we consider integer autoregressive models for count time series and discuss estimation and possible extensions based on real data applications.

1 Introduction

Figure 1 motivates the study of appropriate models for the statistical analysis of count time series. The upper left plot shows a time series of claims—referred to as series C3—of short-term disability benefits made by cut injured workers in the logging industry, (Zhu and Joe (2006)). The lower left plot shows the usual sample autocorrelation function (acf) for these data. If the acf is adapted as a measure of correlation between pairs of observations, then the resulting plot points to weak correlation which decays fast, after a few lags. Therefore, to model these data, in the spirit of usual ARMA models (see Brockwell and Davis (1991) for instance), a few lagged variables entertained by a regression model will suffice to describe the correlation among the data. The right plots illustrate quite the opposite situation. The upper right plot shows the number of transactions for the stock Ericsson B, for one day period. The
lower right plot of the same figure shows the sample acf for these data. When compared to the previous data example, we note a distinct feature of these data; there exists a strong correlation among data which decays slowly. Hence, a few lagged variables in a regression model will not be sufficient to accommodate these particular features of the data. In other words, the modeling of these data raises analogous questions and poses similar challenges to the case of ARCH and GARCH models; Engle (1982), Bollerslev (1986). The main goal of this article is to discuss statistical inference for count time series and give some guidelines for inference in situations similar to the aforementioned data examples. By doing so, we also review some important probabilistic properties of such models and the associated statistical inference.

Figure 1: Left plot: Time series of claims (up) and their sample acf (bottom). Right plot: Number of transactions (up) and their sample acf (bottom).

Modeling counts of events can be found in all areas of statistics and econometrics, and throughout the social and physical sciences. Apart from the above cases, we can observe daily number of hospital admissions, monthly number of cases of some disease, weekly number of rainy days, and so on. For the regression analysis of count data, the ordinary linear model would not be applicable, since the response
variable assumes discrete values. However, a related counterpart is the Poisson regression model and it is a natural starting point to extend it to dependent count data. The Poisson regression model has been used in several applied areas to model counts. In fact, the Poisson model is a nonlinear, albeit straightforward and popular modeling tool, whose fitting is implemented by standard software. This chapter will survey models and methods for analyzing time series of counts, beginning with this basic tool.

The Poisson model provides the main instrument for modeling count time series data. However other distributional assumptions may be used instead; the most natural among other candidates being the negative binomial distribution. Regardless of the chosen distribution, we will review mostly models that fall under the framework of generalized linear models for time series. This class of models and the maximum likelihood theory provide a systematic framework for the analysis of quantitative as well as qualitative time series data. Indeed, estimation, diagnostics, model assessment, and forecasting are implemented in a straightforward manner, where the computation is carried out by a number of existing software packages. Experience with these models shows that both positive and negative association can be taken into account by a suitable parametrization of the model. These issues are addressed in the list of desiderata suggested in Davis et al. (1999) and Zeger and Qaqish (1988).

There are other alternative classes of regression models for count time series; the most prominent being the integer autoregressive models. These models are based on the notion of thinning operator. Accordingly, integer autoregressive models imitate the structure of the common autoregressive process in the sense that the thinning operation is applied instead of scalar multiplication.

This chapter surveys several of the above models. We discuss their properties, estimation methods and theory. Section 2 is introductory to Poisson regression modeling. Section 3 discusses, in detail, linear and log-linear models for count time series. The Poisson assumption is dropped in Section 4, where we study models using different distributional assumptions. Section 5 summarizes properties of the integer autoregressive models. Finally Section 6 concludes this work with other potential applications and further development of the methodology. For ease of presentation, we use slightly abused notation for the regression parameters and the error sequences. However, this does not affect the main concepts as will be clear from the context.

2 Poisson Regression Modeling

The Poisson distribution is commonly used to model rate of random events that occur (arrive) in some fixed time interval. If we assume that \( \lambda \) denotes the rate of arrivals, then the distribution of the random variable \( Y \), which denotes the number of arrivals in a fixed time interval, follows the Poisson distribution.
with probability mass function

\[ P[Y = y] = \frac{\exp(-\lambda)\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots \] (1)

It is an elementary exercise to show that the mean and variance of \( Y \) are both equal to \( \lambda \); \( E[Y] = \text{Var}[Y] = \lambda \). In fact, this property characterizes the Poisson distribution. A related property is that the cumulant generating function of a Poisson random variable is given by \( K_Y(t) = \log M_Y(t) = \lambda(\exp(t) - 1) \), where \( M_Y(t) \) is the moment generating function of \( Y \). This can be proved by simple calculations, but for this presentation, it is instructive to consider the Poisson distribution as a member of the natural exponential family of distributions.

Let \( f(x; \theta) \) denote the density function of the natural exponential family with parameter \( \theta \), i.e.

\[ f(x; \theta) = h(x) \exp(\theta x - b(\theta)), \quad x \in A \] (2)

where \( h(\cdot), b(\cdot) \) are known functions and \( A \) is a subset of \( \mathbb{R} \). Then, it is straightforward to show that the Poisson distribution is expressed as in (2) with \( \theta = \log \lambda, \ b(\theta) = \exp(\theta) \) and \( h(x) = 1/x! \). Using the fact that the cumulant generating function of (2) is equal to \( b(t + \theta) - b(\theta) \), the claim follows.

In most of the applications, count data are usually observed with some covariate information. For example, see McCullagh and Nelder (1989, Sec. 6.3.2) where the authors study the relation between the type of ship, its year of construction and its service period to the expected number of damage incidents using the logarithm of the aggregate months of service as an offset. (An offset is a continuous regression variable with corresponding known regression coefficient equal to 1). In general, assume that \( X_1, \ldots, X_p \) are \( p \) regression variables observed jointly with a count response variable \( Y \) that follows the Poisson distribution. A possible regression model for association between the regressors and the expected value of \( Y \) given \( X_1, \ldots, X_p \) is

\[ \lambda = \beta_0 + \sum_{i=1}^p \beta_i X_i. \] (3)

This is an ordinary linear model with unknown regression coefficients \( \beta_i, \ i = 0, \ldots, p \) to be estimated. Model (3) poses several difficulties for fitting, because the parameter \( \lambda \) has to be positive. Nevertheless, in the context of time series and when the correlation among successive observations is positive, models such as (3) are quite useful; recall Figure 1. A more natural choice for the regression modeling of count data is the so called log–linear model which is specified by

\[ \log \lambda = \beta_0 + \sum_{i=1}^p \beta_i X_i, \] (4)

where the notation is as in (3). Regardless of the chosen model, a fact that remains true is that both (3) and (4) belong to the class of generalized linear models as introduced by Nelder and Wedderburn (1972) and elaborated further in McCullagh and Nelder (1989). Recall, that a generalized linear model consists
of three components; the random component which belongs to the exponential family of distributions (2) with \( \text{E}[X] = \mu \), the systematic component \( \eta \) and the link function \( g(\cdot) \). The link function is a monotone twice differentiable function which is chosen by the user (or can be estimated). This function associates the random and systematic component via \( g(\mu) = \eta \). For the Poisson distribution, it is clear that both (3) and (4) introduce a generalized linear model with \( \eta = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \) and \( g(\lambda) = \lambda \) (for (3)) and \( g(\lambda) = \log \lambda \) (for (4)). Estimation and inference are based on the maximum likelihood theory—this topic has been described in several texts; see McCullagh and Nelder (1989) and Agresti (2002), for example. In the next section, we explore these ideas in the context of count time series.

3 Poisson Regression Models for Count Time Series

It is useful to consider the classical AR(1) process

\[
Y_t = b_1 Y_{t-1} + \epsilon_t, 
\]

where \( |b_1| < 1 \) and \( \{\epsilon_t\} \) is a sequence of independent and identically distributed (iid) normal random variables with zero mean and variance \( \sigma^2 \). This is a standard model used for the analysis of real valued time series. It implies that the value of the process at time \( t \) depends on the value of the process at time \( (t-1) \) plus a random error; e.g. Priestley (1981), Brockwell and Davis (1991, Ch.3) and Shumway and Stoffer (2006). It is enlightening to consider model \( (5) \) as a member of the family of generalized linear models for time series. Recalling the discussion at the end of the last section, note that the random component of the model (for the AR(1) model \( (5) \) the conditional probability density function of \( Y_t \) given its past is Gaussian) belongs to the exponential family of distribution. In addition, the systematic component is defined by \( \eta_t = b_1 Y_{t-1} \). If the link function \( g(\cdot) \) is chosen to be the identity, then \( g(\text{E}[Y_t | Y_{t-1}] = \eta_t \). Hence, the AR(1) process \( (5) \) falls within the framework of generalized linear models for time series, see Kedem and Fokianos (2002, Ch.1). This discussion motivates much of the following development.

3.1 Linear Models for Count Time Series

From this point on, assume that \( \{Y_t\} \) denotes a count time series; we will call this process the "response". Following the AR(1) paradigm we generalize model \( (5) \) in the context of Poisson autoregression by assuming that

\[
Y_t | F_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + b_1 Y_{t-1}, \quad t \geq 1, 
\]

with \( F_t = \sigma(Y_s, s \leq t) \), \( d, b_1 \) non-negative parameters and \( \{\lambda_t\} \) denoting the mean process of \( Y_t \) given its past. Positive \( d \) and \( b_1 \) ensure that \( \lambda_t > 0 \), since \( Y_t \) is a non-negative integer. With this notation, it is clear that model \( (6) \) falls within the framework of generalized linear models with the random component
being the Poisson distribution, the systematic component given by \( \eta_t = d + b_1 Y_{t-1} \) and the identity link. This is a situation quite analogous to (5). Model (6) implies the same dynamics of model (5), since

\[
Y_t = \lambda_t + (Y_t - \lambda_t) = d + b_1 Y_{t-1} + \epsilon_t, \quad t \geq 1, \tag{7}
\]

where the notation is obvious. The last line displays that the values of the process at time \( t \) depends on the value of the process at time \((t - 1)\) plus the term \( \{\epsilon_t\} \) which is white noise sequence; that is a sequence of uncorrelated random variables with zero mean and constant variance. Indeed, if we assume that the process \( \{Y_t\} \) is stationary, then we obtain the following results:

- **Constant mean:**
  \[
  E[\epsilon_t] = E\left( Y_t - \lambda_t \right) = E\left( Y_t - \lambda_t \mid F_{t-1} \right) = 0,
  \]

- **Constant variance:**
  \[
  \text{Var}[\epsilon_t] = \text{Var}\left( E\left( \epsilon_t \mid F_{t-1} \right) \right) + E\left[ \text{Var}\left( \epsilon_t \mid F_{t-1} \right) \right] = E[\lambda_t] = E[Y_t].
  \]
  This is independent of \( t \) since \( \{Y_t\} \) has been assumed to be stationary. The last equality follows from the fact that \( E[\epsilon_t] = 0 \).

- **Uncorrelated sequence:** For \( k > 0 \),
  \[
  \text{Cov}(\epsilon_t, \epsilon_{t+k}) = E[\epsilon_t \epsilon_{t+k}] = E[\epsilon_t E(\epsilon_{t+k} \mid F_{t+k-1})] = 0
  \]
  These results verify the claim that the sequence \( \{\epsilon_t\} \) is a white noise sequence. Because of the assumed stationarity, (7) shows that \( E[Y_t] = d + b_1 E[Y_{t-1}] \) and therefore \( \text{Var}[\epsilon_t] = E[Y_t] = d/(1 - b_1) \); a fact which illustrates that \( b_1 \) needs to be positive but less than 1.

To start investigating the second order properties of (6), we employ representation (7). By repeated substitution, we obtain that

\[
Y_t = d + b_1 Y_{t-1} + \epsilon_t \\
= d + b_1(d + b_1 Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\
= d(1 + b_1) + b_1^2 Y_{t-2} + b_1 \epsilon_{t-1} + \epsilon_t \\
= \ldots \\
= d(1 + b_1 + b_1^2 + \cdots + b_1^t) + \sum_{i=0}^{t} b_1^i \epsilon_{t-i}. \tag{8}
\]

Therefore, as in the case of the usual AR(1) model, assuming that \( 0 < b_1 < 1 \), we obtain by (8) for large \( t \), the useful representation

\[
Y_t = \frac{d}{1 - b_1} + \sum_{i=0}^{\infty} b_1^i \epsilon_{t-i}. \tag{9}
\]
in mean square sense. Standard arguments now show that the autocovariance function of model (6) is
given by

\[ \text{Cov}(Y_t, Y_{t+h}) = \frac{b_1}{1 - b_1^2} E[Y_t], \quad h \geq 0, \]  

(10)
a fact that yields the acf of model (6):

\[ \text{Corr}(Y_t, Y_{t+h}) = b_1^h, \quad h \geq 0. \]  

(11)

Note that unless \( b_1 = 0 \), the variance of \( \{Y_t\} \) is always greater than its expectation; i.e. model (6) takes
into account overdispersion. These results are straightforward consequences of (7) because it reveals that
(6) has identical second order properties to those of the AR(1) model (5). However, \( \text{Corr}(Y_t, Y_{t+h}) > 0, \) for all \( h > 0 \), because \( b_1 > 0 \); that is model (6) can be employed for positively correlated count time
series.

An empirical verification of these considerations is illustrated in the left plot of Figure 2. The upper
plot shows two hundred observations from (6) with \( d = 1 \) and \( b_1 = 0.6 \). The lower plot shows the
autocorrelation function of the same model. Quite clearly, as the lag \( h \) increases, the autocorrelation
function tends fast to smaller values; see equation (11) and compare this plot with the left hand plot of
Figure 1. (Further results about moments and cumulants of model (6) are given in Weiß (2010).)

The right plot of Figure 2 illustrates a different situation. It shows two hundred realizations of the
following model

\[ Y_t \big| \mathcal{F}^Y_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + a_1 \lambda_{t-1} + b_1 Y_{t-1}, \quad t \geq 1, \]  

(12)

where \( \mathcal{F}^Y_t \) the \( \sigma \)-field generated by \( \{Y_0, \ldots, Y_t, \lambda_0\} \), that is \( \mathcal{F}^Y_t = \sigma(Y_s, \lambda_s, s \leq t) \), and \( \{\lambda_t\} \) is a
Poisson intensity process, as before. The parameters \( d, a_1, b_1 \) are assumed to be positive and to satisfy
\( 0 < a_1 + b_1 < 1 \). Both starting values \( \lambda_0 \) and \( Y_0 \) are assumed to be random. When \( a_1 = 0 \), then model (6)
is recovered. Recall the lower right plot of Figure 2 and compare it with the corresponding plot of Figure
1. Apparently the persistence of large positive values of the autocorrelation function is a consequence of
the existence of the feedback mechanism \( \{\lambda_t\} \) introduced in (12). In principle, when count time series
are available and their autocorrelation function assumes relatively high values for large lags, then we
should expect a model of the form (6) to accommodate this fact by entertaining a large number of lagged regressor variables. However such an approach can be avoided when employing model (12); it simply
provides a parsimonious way to model this type of data, Fokianos et al. (2009).

Several results for model (12) have been reported in the literature, see Rydberg and Shephard (2000),
Streett (2000), Heinen (2003) and Ferland et al. (2006) who consider the following general model of order
\((p, q)\):

\[ Y_t \big| \mathcal{F}^Y_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, \quad t \geq \max(p, q), \]  

(13)
and show that it is second order stationary provided that $0 < \sum_{i=1}^p a_i + \sum_{j=1}^q b_j < 1$.

To study the properties of (12), it is instructive to consider again decomposition (7) and then use the second part of (12) to express the response process as
\[
  Y_t = d + (a_1 + b_1)Y_{t-1} + \varepsilon_t - a_1\varepsilon_{t-1},
\]
with some slight abuse of notation. In the last display, $\varepsilon_t = Y_t - \lambda_t$ and this sequence of random variables, although distinct from the corresponding sequence defined by means of (7), is still a white noise process; the proof of this fact is the same as to the case of the noise sequence that corresponds to model (6).

Furthermore, the last display can be rewritten as
\[
  \left( Y_t - \frac{d}{1 - (a_1 + b_1)} \right) = (a_1 + b_1) \left( Y_{t-1} - \frac{d}{1 - (a_1 + b_1)} \right) + \varepsilon_t - a_1\varepsilon_{t-1},
\]
which shows that (12) has exactly identical second order properties as those of a usual ARMA(1,1) model. Hence, when $0 < a_1 + b_1 < 1$, then there exists a stationary solution $\{Y_t\}$ of (12), with mean.
\[ E[Y_t] = E[\lambda_t] \equiv \mu = d / (1 - a_1 - b_1) \] and autocovariance function

\[
\text{Cov}[Y_t, Y_{t+h}] = \begin{cases} 
(1 - (a_1 + b_1)^2 + b_1^2)\mu, & h = 0, \\
\frac{b_1(1 - a_1(a_1 + b_1))(a_1 + b_1)^{h-1}\mu}{1 - (a_1 + b_1)^2}, & h \geq 1.
\end{cases}
\]

It is clear that the acf of model (12) is equal to

\[
\text{Corr}[Y_t, Y_{t+h}] = \frac{b_1(1 - a_1(a_1 + b_1))(a_1 + b_1)^{h-1}\mu}{1 - (a_1 + b_1)^2 + b_1^2}, \quad h \geq 1.
\]

This fact matches the right hand plots of Figures 1 and 2 and explains the slower decay of the corresponding acf.

For the Poisson distribution, \( E[Y_t | \mathcal{F}_{t-1}^{Y}] = \text{Var}[Y_t | \mathcal{F}_{t-1}^{Y}] = \lambda_t \). Therefore model (12) can be defined as an INGARCH(1,1); i.e. that is an integer GARCH model, since its structure is analogous to that of the customary GARCH model whereby volatility is regressed on past values of itself and squared responses. In fact, model (13) can be termed as an INGARCH(p, q) model. However (12), and more generally (13), specify a conditional mean relation to the past values of both \( \lambda_t \) and \( Y_t \). Observe that \( \text{Var}[Y_t] \geq E[Y_t] \) with equality when \( b_1 = 0 \). Thus, the inclusion of the past values of \( Y_t \) in the evolution of \( \lambda_t \) yields overdispersion–this is the same fact that holds for model (6). Furthermore, \( \text{Corr}[Y_t, Y_{t+h}] > 0 \) for model (12) like in the case of model (6).

By repeated substitution

\[
\lambda_t = d + a_1\lambda_{t-1} + b_1Y_{t-1} \\
= d + a_1(d + a_1\lambda_{t-2} + b_1Y_{t-2}) + b_1Y_{t-1} \\
= d + a_1d + a_1^2\lambda_{t-2} + a_1b_1Y_{t-2} + b_1Y_{t-1} \\
= \ldots \ldots \ldots \\
= d \frac{1 - a_1^h}{1 - a_1} + a_1^h\lambda_0 + b_1 \sum_{i=0}^{i-1} a_1^iY_{t-i-1}.
\]

The last display shows that the hidden process \( \{\lambda_t\} \) is determined by past functions of lagged responses and the initial value \( \lambda_0 \). Therefore model (12) belongs to the class of observation driven models in the sense of Cox (1981). Representation (15) explains further the reason that model (12) offers a parsimonious way of modeling count time series data whose acf decays slowly; see the example shown in the right plot of Figure 1. The process \( \{\lambda_t\} \) depends on a large number of lagged response values so it is expected to provide a more parsimonious model than a model of the form (6).

As a final remark, when \( a_1 + b_1 \) approaches 1, then the acf function of model (12) becomes unstable and the resulting model has similar properties to those of an integrated GARCH model; that is predictions
for $\lambda_t$ will reflect the most recent variation found in the data. Such models have not been studied in the literature.

### 3.2 Log-Linear Models for Count Time Series

The previous discussion shows that model (12) provides a satisfactory conceptual framework for modeling dependent count data. However, the model definition imposes implicitly some restrictions on the data. First recall that, $\text{Cov}[Y_t,Y_{t+h}] > 0$, because $0 < a_1 + b_1 < 1$. Therefore model (12) cannot be employed for modeling negative correlation among successive observations. An additional drawback of (12) is that it does not accommodate covariates in a straightforward way, because of the identity link function. However, as it was mentioned in Section 2, the choice of the logarithmic function is the most popular among the link functions for modeling count data. Hence, we resort to log-linear models for count time series; see Zeger and Qaqish (1988), Li (1994), MacDonald and Zucchini (1997), Brumback et al. (2000), Kedem and Fokianos (2002), Benjamin et al. (2003), Davis et al. (2003), Fokianos and Kedem (2004), Jung et al. (2006), Creal et al. (2008) and Fokianos and Tjøstheim (2011).

Suppose again that $\{Y_t\}$ denotes a count time series. We will be working with the so-called canonical link process $\nu_t \equiv \log \lambda_t$. We study the following family of log-linear autoregressive models

$$Y_t \mid \mathcal{F}^{Y,\nu}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a_1 \nu_{t-1} + b_1 \log(Y_{t-1} + 1), \quad t \geq 1.$$  \hspace{1cm} (16)

where $\mathcal{F}^{Y,\nu}_t$ the $\sigma$-field generated by $\{Y_0, \ldots, Y_t, \nu_0\}$, that is $\mathcal{F}^{Y,\nu}_t = \sigma(Y_s, \nu_0, s \leq t)$. In general, the parameters $d, a_1, b_1$ can be positive or negative but they need to satisfy certain conditions so that we obtain a stationary time series. Both $\nu_0$ and $Y_0$ are assumed again to be some random starting values. Note that the lagged observations of the response $Y_t$ are fed into the autoregressive equation for $\nu_t$ via the term $\log(Y_{t-1} + 1)$. This is a one-to-one transformation of $Y_{t-1}$ which is quite standard in coping with zero data values. Moreover, both $\lambda_t$ and $Y_t$ are transformed into the same scale. Covariates can be accommodated by model (16), by including them in the second equation of (16). An alternative modeling approach is based upon employing the transformation $\log(\max(Y_{t-1}, c))$, (cf. Zeger and Qaqish (1988)) for $c \in (0, 1]$, instead of $\log(Y_{t-1} + 1)$ in (16).

When $a_1 = 0$, we obtain the model

$$\nu_t = d + b_1 \log(Y_{t-1} + 1), \quad t \geq 1,$$  \hspace{1cm} (17)

which parallels the structure of (6). With this notation, it is clear that model (17) falls within the framework of generalized linear models with the random component being the Poisson distribution, the systematic component given by $\eta_t = d + b_1 \log(Y_{t-1} + 1)$ and the link function being the logarithmic. Figure 3 illustrates the same phenomenon as that observed in Figure 2, namely the inclusion of the feedback mechanism yields parsimony when the correlation decays slowly to zero.
Figure 3: Left plot: Two hundred observations (up) and their sample acf (bottom) from model (17) for $d = 0.1$, and $b_1 = 0.6$. Right plot: Two hundred observations (up) and their sample acf (bottom) from model (16) for $d = 0.1$, $a_1 = 0.3$ and $b_1 = 0.6$.

The log-intensity process of (16) can be rewritten as

$$\nu_t = d \frac{1 - a_1^t}{1 - a_1} + a_1^t \nu_0 + b_1 \sum_{i=0}^{t-1} a_1^i \log(1 + Y_{t-i-1}),$$

(18)
after repeated substitution. Hence, we obtain again that the hidden process $\{\nu_t\}$ is determined by past functions of lagged responses. Equivalently, the log–linear model (16) belongs to the class of observation driven models and possess similar properties to the linear model (12).

To motivate further the choice of the $\log(\cdot)$ function for the lagged values of the response, consider a model like (16), but with $Y_{t-1}$ included instead of $\log(Y_{t-1} + 1)$. In other words, set

$$Y_t \mid F_{t-1}^{Y_t} \sim \text{Poisson}(\lambda_t), \quad \nu_t = d + a_1 \nu_{t-1} + b_1 Y_{t-1}.$$ 

In this case

$$\lambda_t = \exp(d) \nu_{t-1}^a \exp(b_1 Y_{t-1}),$$

11
and therefore stability of the above system is guaranteed only when $b_1 < 0$. Otherwise, the process \{\lambda_t\} increases exponentially fast, see Wong (1986) and Kedem and Fokianos (2002, Ch.4) for more. Hence, only negative correlation can be introduced by such a model. However (16) yields both positive (respectively, negative) correlation by allowing the parameter $b_1$ to take positive (respectively, negative) values. It is a challenging problem to obtain an explicit expression for the autocorrelation function of model (16). This is easily seen by considering (18). Exponentiating both sides of this formula shows that

$$\lambda_t = \exp \left( d \frac{(1 - a_1^t)}{(1 - a_1)} \right) \lambda_0^{b_1} \prod_{i=0}^{t-1} \left( 1 + Y_{t-i+1} \right)^{b_1 a_1^i},$$

which demonstrates the complications of calculating first and second moments for model (16). However, by simulating a very long path of the series, we get a clue of the range of possible values of correlation obtained by (16). Table 3.2 illustrates the acf of model (16) at lags one and two. It is evident that the log-linear model takes into account both negative and positive correlations.

<table>
<thead>
<tr>
<th>$a_1$</th>
<th>$b_1$</th>
<th>$\rho(1)$</th>
<th>$\rho(2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.800</td>
<td>-0.430</td>
<td>-0.984</td>
<td>0.997</td>
</tr>
<tr>
<td>-0.500</td>
<td>-1.000</td>
<td>-0.519</td>
<td>0.613</td>
</tr>
<tr>
<td>-0.400</td>
<td>-0.350</td>
<td>-0.188</td>
<td>0.117</td>
</tr>
<tr>
<td>0.100</td>
<td>0.200</td>
<td>0.145</td>
<td>0.016</td>
</tr>
<tr>
<td>0.250</td>
<td>0.550</td>
<td>0.630</td>
<td>0.500</td>
</tr>
<tr>
<td>0.250</td>
<td>0.730</td>
<td>0.979</td>
<td>0.959</td>
</tr>
</tbody>
</table>

Table 1: Typical values of the autocorrelation function at lags 1 and 2 derived by model (16) for selected values of the parameters $a_1$ and $b_1$ when $d = 0.5$. Results are based on 10000 data points. Here $\rho(h) = \text{Corr}[Y_t, Y_{t+h}]$ for $h = 1, 2$.

An alternative log-linear model specification for count time series was studied by Davis et al. (2003). The model is given by the following

$$v_t = \beta_0 + \sum_{i=1}^{p} \beta_i \zeta_{t-i},$$

(19)

with

$$\zeta_t = \frac{Y_t - \lambda_t}{\lambda_t^d},$$

(20)

where $\beta_i, i = 0, \ldots, p$ (with $\beta_i \neq 0$ for $i = 1, 2, \ldots, p$) are unknown regression parameters and $\delta \in (0, 1]$. If $\delta = 1/2$ then (19) is a moving average model of the so called Pearson residuals; see definition (31). Under the above specification, we have the following results:

- The mean of the sequence \{\zeta_t\} is zero:
  \[ E[\zeta_t] = 0. \]

- The variance of the sequence \{\zeta_t\} is given by the following:
  \[ \text{Var}[\zeta_t] = E[\lambda_t^{1-2\delta}]. \]
• The mean and acf of the log-mean process \( \{v_t\} \) are:

\[
E[v_t] = \beta_0,
\]

and

\[
\text{Cov}[v_t, v_{t+h}] = \begin{cases} 
\sum_{i=1}^{p-h} \beta_i \beta_{i+h} \lambda_{i-1}^{1-2\delta}, & h \leq p, \\
0, & \text{otherwise.}
\end{cases}
\]

We note that when \( \delta = 1/2 \), all the above expressions do not depend on \( t \). In particular, the covariance between \( v_t \) and \( v_{t+h} \), for \( h > 0 \) reduces to the covariance of a standard moving average model of order \( p \).

**Remark 3.1** As it was already mentioned, one of the advantages of model (16) is that time dependent covariates can be easily introduced. To be more specific, suppose that \( \{X_t\} \) is some covariate time series. Then enlarging the \( \sigma \)-field to \( \mathcal{F}^Y_{t-1} = \sigma(Y_s, X_{s+1}, \lambda_0, s \leq t) \) we obtain the model

\[
Y_t \mid \mathcal{F}^Y_{t-1} \sim \text{Poisson}(\lambda_t), \quad v_t = d + a v_{t-1} + b \log(Y_{t-1} + 1) + c X_t, \quad t \geq 1,
\]

where \( c \) is, in general, a real valued parameter. Some remarks about the possible choices of the parameter \( c \) will be made in later sections. This remark, with obvious modifications, also applies to the case of model (19) as well.

### 3.3 Non-Linear Models for Count Time Series

A large class of models for the analysis of count time series is given by the following specification

\[
Y_t \mid \mathcal{F}^Y_{t-1}, \lambda_t = f(\lambda_{t-1}, Y_{t-1}), \quad t \geq 1,
\]

where \( f(\cdot) \) is a known function up to an unknown finite dimensional parameter vector. Moreover, \( f(\cdot) \) takes values on the positive real line, that is \( f : (0, \infty) \times \mathbb{N} \rightarrow (0, \infty) \) and the initial values \( Y_0 \) and \( \lambda_0 \) are assumed again to be random. An interesting example of a non-linear regression model for count time series analysis is given by the following specification

\[
f(\lambda, y) = d + (a_1 + c_1 \exp(-\gamma \lambda^2)) \lambda + b_1 y,
\]

where \( d, a_1, c_1, b_1, \gamma \) are positive parameters. The above model is rather similar to the traditional exponential autoregressive model, see Haggan and Ozaki (1981). In Fokianos et al. (2009), model (23) was
studied for the case \( d = 0 \). Note that the parameter \( \gamma \) introduces a perturbation of the linear model (12), in the sense that when \( \gamma \) tends either to 0 or infinity, then (23) approaches two distinct linear models. An obvious generalization of model (22) is given by the following specification of the mean process

\[
\lambda_t = f(\lambda_{t-1}, \ldots, \lambda_{t-p}, Y_{t-1}, \ldots, Y_{t-q}),
\]

where \( f(.) \) is a function such that \( f : (0, \infty)^p \times \mathbb{N}^q \to (0, \infty) \). Such examples are provided by the class of smooth transition autoregressive models of which the exponential autoregressive model is a special case (cf. Teräsvirta et al. (2010)). Further examples of non-linear time series models can be found in Tong (1990) and Fan and Yao (2003). These models have not been considered in the literature earlier in the context of generalized linear models for count time series, and they provide a flexible framework for studying dependent count data.

### 3.4 Inference

We illustrate conditional maximum likelihood inference for the linear model (12). The methodology is quite analogous for models (16) and (22), so it is omitted. However, for models such as (23), the presence of non-linear parameters requires larger sample sizes for accurate estimation. Recall now (12) and let \( \theta \) be the three dimensional vector of unknown parameters, that is \( \theta = (d, a_1, b_1)' \), and let the true value of the parameter be \( \theta_0 = (d_0, a_{1,0}, b_{1,0})' \). Then the conditional likelihood function for \( \theta \) based on model (12) and given a starting value \( \lambda_0 \) is given by

\[
L(\theta) = \prod_{t=1}^{n} \frac{\exp(-\lambda_t(\theta)) \lambda_t^{Y_t}(\theta)}{Y_t!}.
\]

Here we use the Poisson assumption, \( \lambda_t(\theta) = d + a_1 \lambda_{t-1}(\theta) + b_1 Y_{t-1} \) by (12) and \( \lambda_t = \lambda_t(\theta_0) \). Hence, the log–likelihood function is given up to a constant, by

\[
l(\theta) = \sum_{t=1}^{n} l_t(\theta) = \sum_{t=1}^{n} \left( Y_t \log \lambda_t(\theta) - \lambda_t(\theta) \right),
\]

and the score function is defined by

\[
S_n(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \sum_{t=1}^{n} \frac{\partial l_t(\theta)}{\partial \theta} = \sum_{t=1}^{n} \left( \frac{Y_t}{\lambda_t(\theta)} - 1 \right) \frac{\partial \lambda_t(\theta)}{\partial \theta},
\]

where \( \frac{\partial \lambda_t(\theta)}{\partial \theta} \) is a three-dimensional vector with components given by

\[
\frac{\partial \lambda_t}{\partial d} = 1 + a_1 \frac{\partial \lambda_{t-1}}{\partial d}, \quad \frac{\partial \lambda_t}{\partial a_1} = \lambda_{t-1} + a_1 \frac{\partial \lambda_{t-1}}{\partial a_1}, \quad \frac{\partial \lambda_t}{\partial b_1} = Y_{t-1} + a_1 \frac{\partial \lambda_{t-1}}{\partial b_1}.
\]

The solution of the equation \( S_n(\theta) = 0 \), if it exists, yields the conditional maximum likelihood estimator of \( \theta \) which is denoted by \( \hat{\theta} \). Furthermore, the Hessian matrix for model (12) is obtained by further
differentiation of the score equations (26),

\[
H_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta} = \sum_{i=1}^{n} \frac{Y_i}{\lambda_i^2(\theta)} \left( \frac{\partial \lambda_i(\theta)}{\partial \theta} \right) \left( \frac{\partial \lambda_i(\theta)}{\partial \theta} \right)' - \sum_{i=1}^{n} \left( \frac{Y_i}{\lambda_i(\theta)} - 1 \right) \frac{\partial^2 \lambda_i(\theta)}{\partial \theta \partial \theta}.
\]  

(28)

The conditional information matrix is defined by

\[
G_n(\theta) = \sum_{i=1}^{n} \text{Var} \left[ \frac{\partial l_i(\theta)}{\partial \theta} \mid \mathcal{F}_{t-1}^{\mathcal{C}} \right] = \sum_{i=1}^{n} \frac{1}{\lambda_i(\theta)} \left( \frac{\partial \lambda_i(\theta)}{\partial \theta} \right) \left( \frac{\partial \lambda_i(\theta)}{\partial \theta} \right)',
\]

(29)

and plays a crucial role in the asymptotic distribution of the MLE \( \hat{\theta} \). More specifically, under certain regularity conditions, it can be proved that \( \hat{\theta} \) is consistent and asymptotically normal, i.e.

\[
\sqrt{n}(\hat{\theta} - \theta_0) \overset{D}{\to} \mathcal{N}(0, G^{-1}),
\]

with the matrix \( G(\theta) \) defined by

\[
G(\theta) = E \left( \frac{1}{\lambda_i} \left( \frac{\partial \lambda_i}{\partial \theta} \right) \left( \frac{\partial \lambda_i}{\partial \theta} \right)' \right),
\]

where \( E[\cdot] \) is taken with respect to the stationary distribution. All the above quantities can be computed, and they are employed for constructing predictions, confidence intervals and so on. Although, the above formulae are given for the linear model (12), they can be modified suitably for the log-linear model (16) and the non-linear model (22).

3.5 On the Asymptotic Distribution of the MLE

It can be proved for regression models of the form (12), (16) and more generally for models like (22), that the MLE \( \hat{\theta} \) is asymptotically normally distributed, as it was mentioned before. This is an important fact, since inference is based on this approximation. However to study the asymptotic theory there is need to develop a central limit theory for the bivariate process \( \{ (Y_t, \lambda_t) \} \). We mention the approach taken by Neumann (2011), who studies model (22) and shows that the bivariate process \( \{ (Y_t, \lambda_t) \} \) has a unique stationary distribution and the response process is absolutely regular. In addition, Franke (2010) considers (24) and shows that the response process is weakly dependent with finite first moment; see Doukhan and Louhichi (1999) and the recent monograph by Dedecker et al. (2007) for definition of weak dependence and further examples. Using the general model (24), the essential condition assumed by both the above references, is that the function \( f(\cdot) \) is a contraction, that is, for any \( (\lambda_1, \ldots, \lambda_p, y_1, \ldots, y_q) \)
and \((\lambda'_1, \ldots, \lambda'_p, y'_1, \ldots, y'_q)\)

\[
|f(\lambda_1, \ldots, \lambda_p, y_1, \ldots, y_q) - f((\lambda'_1, \ldots, \lambda'_p, y'_1, \ldots, y'_q))| \leq \sum_{i=1}^{p} a_i |\lambda_i - \lambda'_i| + \sum_{j=1}^{q} \gamma_j |y_j - y'_j|,
\]

where \(\sum_{i=1}^{p} a_i + \sum_{j=1}^{q} \gamma_j < 1\). This is the same condition assumed by Fokianos et al. (2009), Fokianos and Tjøstheim (2010) whose approach is based on Markov chains theory.

Turning now to the questions regarding ergodicity and inference we note that these problems have been examined in detail by Fokianos et al. (2009), Fokianos and Tjøstheim (2011, 2010) (see also Woodard et al. (2010)) who also use a perturbation argument to prove geometric ergodicity of \(\{(Y_t, \lambda_t)\}\). This means that instead of proving geometric ergodicity of \(\{(Y_t, \lambda_t)\}\), the authors are considering a perturbed \(\{(Y'_t, \lambda'_t, U_t)\}\), where \(\{U_t\}\) is a sequence of iid uniform random variables. The strategy to study the properties of the bivariate process \(\{(Y_t, \lambda_t)\}\) is to prove geometric ergodicity of \(\{(Y'_t, \lambda'_t, U_t)\}\) and then to use this fact to obtain asymptotic normality for the likelihood estimators. Asymptotic normality of the likelihood estimates of the non perturbed model is proved by employing an approximation lemma which gives conditions for the proximity of the perturbed version to non perturbed version. Detailed exposition of the perturbation argument can be found in the references mentioned above. Here we list the following up to date known facts for models (12) and (16):

1. For the linear model (12):
   
   (a) Consider the perturbed linear model and suppose that \(0 < a_1 + b_1 < 1\). Then the process \(\{(Y'_t, \lambda'_t, U_t), t \geq 0\}\) is a \(V_{(Y,U,\lambda)}\)-geometrically ergodic Markov chain with \(V_{Y,U,\lambda}(Y,U,\lambda) = 1 + Y^k + \lambda^k + U^k\).

   (b) If \(0 < a_1 + b_1 < 1\), then the perturbed model can be made arbitrarily close to the unperturbed model.

   (c) If \(0 < a_1 + b_1 < 1\), then the conditional maximum likelihood estimators of \((d, a_1, b_1)\) are consistent and asymptotically normally distributed.

2. For the log–linear model (16), define \(\{(Y'_t, v'_t, U_t), t \geq 0\}\) as its perturbed version.
   
   (a) Suppose that \(|a_1| < 1\). In addition, assume that when \(b_1 > 0\), then \(|a_1 + b_1| < 1\), and when \(b_1 < 0\), then \(|a_1| |a_1 + b_1| < 1\). Then, the process \(\{(Y'_t, U_t, v'_t), t \geq 0\}\) is a \(V_{(Y,U,v)}\)-geometrically ergodic Markov chain with \(V_{Y,U,v}(Y,U,v) = 1 + \log^k(1 + Y) + v^{2k} + U^{2k}\), \(k\) being a positive integer.

   (b) If \(|a_1 + b_1| < 1\), whenever \(a_1\) and \(b_1\) have the same sign, and \(a_1^2 + b_1^2 < 1\) whenever \(a_1\) and \(b_1\) have different signs, then the perturbed log–linear model can be made arbitrarily close to the unperturbed log-linear model.
(c) If $|a_1 + b_1| < 1$, whenever $a_1$ and $b_1$ have the same sign, and $a_1^2 + b_1^2 < 1$ whenever $a_1$ and $b_1$ have different signs, then the conditional maximum likelihood estimators of $(d, a_1, b_1)$ are consistent and asymptotically normally distributed.

We clarify the above results by considering the first statement about the linear model (12). Result 1(a) implies that when $0 < a_1 + b_1 < 1$, then the perturbed model possesses moments of any order and any average of functions of $\{ (Y^m_t, \lambda^m_t, U_t), t \geq 0 \}$ will converge weakly to its expected value. This fact has important consequences since it allows the study of the maximum likelihood estimators derived by (25) given that the unperturbed model is close to the perturbed model under the same condition. For the log-linear model (16), the conditions for proving that the perturbed model approaches the unperturbed model are quite restrictive when compared to the conditions for geometric ergodicity. The same phenomenon occurs for model (19) for the case $p = 1$; see Davis et al. (2005) who prove asymptotic normality of the maximum likelihood estimators when $\delta = 1$ and $\beta_1 > 0$, such that $\beta_1 (1 + \exp (\beta_1 - \beta_0))^{1/2} < 1$. However, it was shown by Davis et al. (2003) that if $1/2 \leq \delta \leq 1$, then the chain $\{ \nu_t \}$ has a stationary distribution. In particular, when $\delta = 1$, then $\{ \nu_t \}$ is uniformly ergodic and has a unique stationary distribution.

To complement the presentation, ergodicity of model (22) has been proved by employing the contraction assumption (30) for $p = q = 1$ on $f(\cdot)$, see Fokianos and Tjøstheim (2010) (for its perturbed version), and Neumann (2011), Franke (2010) for the response process. Under such assumption Fokianos and Tjøstheim (2010) show the asymptotic normality of the MLE for model (22) showing that the perturbed version approximates the non perturbed version. We close this part by the following important remarks.

Remark 3.2 For a log-linear model which includes covariates, such as (21), the estimation problem is attacked along the lines described in Section 3.4. To study ergodicity and asymptotic normality of the MLE in this case, suppose that $\{ X_t \}$ a real-valued Markov chain which possess a density. Then we can construct a 2-dimensional Markov chain $\{ \nu_t, X_{t+1} \}$ and a corresponding 3-dimensional chain with $\{ Y_t \}$ included. If the transition mechanism of $\{ X_t \}$ does not depend on $\{ \nu_t, Y_t \}$, it is simple to find conditions for geometric ergodicity; see Fokianos and Tjøstheim (2011), for more.

Remark 3.3 We note, however, that the asymptotic theory concerning the maximum likelihood estimators for the regression parameters has been developed under the assumption of the Poisson distribution. Such an approach poses several robustness issues related to model misspecification. A possible venue to overcome this problem is the quasi–likelihood estimation method, see Heyde (1997) and Kedem and Fokianos (2002, Sec. 1.7), for instance. In this case, the score is determined by a mean regression equation and a working variance function. Such methods have been explored, for example, in the GARCH framework by Berkes et al. (2003) and it is worth studying their performance in the context of count time series regression models.
3.6 Data Examples

The above theory is applied to the real data examples discussed in the Introduction; recall Figure 1. For both time series, the mean is always less than their variance. In other words, the data exhibits overdispersion—a fact that holds for all the Poisson distributed models that were discussed so far. For the analysis of those time series we fit both the linear model (12) and the log-linear model (16). To model the data, set \( \lambda_0 = 0 \) and \( \partial \lambda_0 / \partial \theta = 0 \) for initialization of the recursions in the case of the linear model; see equations (27). For the log-linear model, the corresponding initializations are set to \( \nu_0 = 1 \) and \( \partial \nu_0 / \partial \theta = 0 \). Table 2 lists the results of the analysis. The numbers in parentheses, next to the estimators, correspond to the standard errors of the estimates. These are computed by using the so called robust sandwich matrix

\[
H_n(\hat{\theta})^{-1} H_n(\hat{\theta}) H_n(\hat{\theta})^{-1},
\]

where \( G_n(\hat{\theta}) \) has been defined by (29) and \( H_n(\hat{\theta}) \) is given by (28).

To examine the adequacy of the fit, consider the so called Pearson residuals (recall (20) with \( \delta = 1/2 \))

\[
e_t = \frac{Y_t - \lambda_t}{\sqrt{\lambda_t}}, \quad t \geq 1.
\]  

Under the true model, the process \( \{e_t\} \) is a white noise sequence with constant variance; see Kedem and Fokianos (2002, Sec. 1.6.3). To estimate the Pearson residuals, substitute \( \lambda_t \) by \( \hat{\lambda}_t \equiv \hat{\lambda}_t(\hat{\theta}) \). Comparison among the models, is implemented by calculating the mean square error (MSE) of the Pearson residuals which is given by

\[
\sum_{t=1}^{N} \hat{e}_t^2 / (N - p),
\]

where \( p \) denotes the number of estimated parameters; see Kedem and Fokianos (2002, Sec. 1.8) for more on diagnostics (see also Zhu and Wang (2010), for a recent contribution directly related to models of the form (6)).

Table 2 summarizes the findings of the data analysis. Consider first the C3 series. We note that both linear and log-linear models yield almost the same MSE and the estimators obtained for \( a_1 \) and \( b_1 \) are similar from both models. In fact, the feedback mechanism does not provide any improvement for the fit, since the estimator of \( a_1 \) is large when compared to its standard error, for both models. Figure 4 shows the results of the data analysis which point to the adequacy of the fit. Note, that the bottom plot shows the cumulative periodogram plot of the Pearson residuals which confirms that the sequence (31) is white noise. One practical aspect that arises in applications of model (16) is the choice of log \((Y_{t-1} + 1)\) in the regression equation. Here, we mention that in order to examine the sensitivity of the results, as a function of the log term in model (16), we can work as follows. Fit the following series of models to the log-mean processes

\[
\nu_t = d + av_{t-1} + b \log(Y_{t-1} + v),
\]

for both time series, where \( v \) is a constant which takes values from 1 to 10 (or some other bound) with step equal to 0.5. Then calculate the MSE of the Pearson residuals for all different model specifications obtained by varying the constant \( v \) and compare them. For the C3 series the sample variance of obtained MSE values is almost zero. In conclusion, we see that the choice of \( \log(Y_{t-1} + 1) \) does not affect the results of the analysis greatly, at least for the C3 series.
Turning now to the transactions data, we see again that both models (12) and (16) yield similar MSE values. Note that the sum of estimated coefficients is close to one for both linear and log-linear model. This corresponds to a frequently observed phenomenon for GARCH(1,1) models. Figure 5 demonstrates again the adequacy of the fit for the log-linear model. To examine the sensitivity of the results as a function of the log term in model (16), we repeat the previous exercise with a constant $v$ which takes values from 1 to 10 with step equal to 0.5. The MSE values have a range between 2.389 to 2.391. Therefore we observe again that choice of $\log(Y_{t-1} + 1)$ does not affect the results of the analysis greatly.

As a general remark, models like (12) and (16) will be more useful when applied to count time series data with strong correlation. The feedback mechanism allows for more parsimonious modeling. This is in accordance with the GARCH methodology whereby lagged values of volatility allow for parsimony. Furthermore, when the data are positively correlated, then models (12) and (16) will yield similar conclusions. It is anticipated that the log-linear model (16) provides a better fit when either there exists negative correlation among the data or when covariates need to be taken into account for the data analysis.

<table>
<thead>
<tr>
<th>Series C3</th>
<th>Linear Model</th>
<th>Log-linear Model Fit</th>
<th>Transactions Data</th>
<th>Linear Model</th>
<th>Log-linear Model Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$</td>
<td>$a_1$</td>
<td>$b_1$</td>
<td>MSE</td>
<td>$d$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>2.385 (0.533)</td>
<td>0.050 (0.088)</td>
<td>0.5603 (0.073)</td>
<td>1.285</td>
<td>0.476 (0.183)</td>
<td>0.080 (0.097)</td>
</tr>
<tr>
<td>0.581 (0.162)</td>
<td>0.744 (0.026)</td>
<td>0.198 (0.016)</td>
<td>2.367</td>
<td>0.105 (0.034)</td>
<td>0.746 (0.026)</td>
</tr>
</tbody>
</table>

Table 2: Data analysis results.

4 Other Regression Models for Count Time Series

The Poisson distribution is the most natural candidate among discrete distributions to model count data. However, the literature offers several alternatives to Poisson. In this section we discuss the case of Negative Binomial distribution and the Double Poisson distribution as alternative models for the analysis of count time series. We also survey other alternative regression based methods for count time series analysis.
Figure 4: (a) Series C3. The red line corresponds to the prediction $\lambda_t(\hat{\theta})$ obtained by fitting model (16). (b) Pearson Residuals obtained by fitting model (16). (c) Cumulative periodogram plot of the Pearson residuals.

### 4.1 Other Distributional Assumptions

Recall that if $Y$ is random variable which follows the negative binomial distribution with parameters $(r, \theta)$, where $\theta \in (0, 1)$ and $r$ an integer, then its probability mass function is given by

$$P[Y = y] = \binom{y + r - 1}{y} \theta^y (1 - \theta)^r, \quad y = 0, 1, 2, \ldots.$$  (32)

Accordingly, we denote $Y \sim \text{NegBin}(r, \theta)$. With this notation, it is well known that $E[Y] = r\theta / (1 - \theta)$ and $\text{Var}[Y] = r\theta / (1 - \theta)^2$.

Consider again $\{Y_t\}$ to be the response and assume the following model, see Zhu (2011):

$$Y_t | X_{t-1}^{Y, \lambda} \sim \text{NegBin}(r, \theta_t), \quad \lambda_t \equiv \frac{\theta_t}{1 - \theta_t} = d + a_1 \lambda_{t-1} + b_1 Y_{t-1}, \quad t \geq 1.$$  (33)
where the parameters $d, a_1, b_1$ are all non-negative and $\lambda_0, Y_0$ are some random starting values. The above model regresses the log–odds of $\theta_t$ to its past values and past values of the responses. More generally, we can study models of the form

$$\lambda_t = d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j}, \quad t \geq \max(p, q),$$

but we will insist on the simpler model (33) for ease of presentation (see Zhu (2011) for more). With the same notation as before, it is easily seen that this particular specification yields again

$$E[Y_t] = E\left( E[Y_t | F_{t-1}^{Y_{t-1}}] \right) = rE[A_t].$$

Therefore, assuming stationarity, we obtain, from (33), that

$$E[Y_t] = r \frac{d}{1 - a_1 - rb_1},$$

Figure 5: (a) Transactions data. The red line corresponds to the prediction $\lambda_t(\hat{\theta})$ obtained by fitting model (16). (b) Pearson Residuals obtained by fitting model (16). (c) Cumulative periodogram plot of the Pearson residuals.
provided that \( a_1 + rb_1 < 1 \). We will be working again as in the case of model (12) in order to understand the dynamics of model (33). Towards this goal, consider again the following representation

\[
Y_t = r\lambda_t + (Y_t - r\lambda_t) = rd + ra_1\lambda_{t-1} + rb_1Y_{t-1} + \epsilon_t,
\]

where the error term \( \{\epsilon_t\} \) is again a white noise sequence. This fact is proved next, by assuming the condition \( a_1 + rb_1 < 1 \). Furthermore, equation (34) implies that the observed process depends on its past values and on its past odds of the sequence of probabilities \( \{\theta_t\} \). The details for proving that the sequence \( \{\epsilon_t\} \) is white noise are as follows:

- **Constant mean:**
  
  \[ \mathbb{E}[\epsilon_t] = \mathbb{E}\left[(Y_t - r\lambda_t)\right] = \mathbb{E}\left[Y_t - r\lambda_t \mid \mathcal{F}_{t-1}^{\lambda}\right] = 0, \]

- **Constant variance:**
  
  \[
  \text{Var}[\epsilon_t] = \text{Var}\left[\mathbb{E}\left(\epsilon_t \mid \mathcal{F}_{t-1}^{\lambda}\right)\right] + \mathbb{E}\left[\text{Var}\left(\epsilon_t \mid \mathcal{F}_{t-1}^{\lambda}\right)\right] = r\mathbb{E}[\lambda_t(1 + \lambda_t)]
  \]
  \[
  = \frac{1 - (a_1 + rb_1)^2}{1 - (a_1 + rb_1)^2 - rb_1^2} \left( \mathbb{E}[Y_t] + \frac{\mathbb{E}^2[Y_t]}{r} \right),
  \]

  which is independent of \( t \) since \( \{Y_t\} \) is stationary. Equation (35) is proved in the Appendix.

- **Uncorrelated sequence:** For \( k > 0 \),
  
  \[
  \text{Cov}(\epsilon_t, \epsilon_{t+k}) = \mathbb{E}[\epsilon_t\epsilon_{t+k}] = \mathbb{E}\left[\epsilon_t\mathbb{E}\left(\epsilon_{t+k} \mid \mathcal{F}_{t+k-1}^{\lambda}\right)\right] = 0
  \]

To study the second order properties of (33), we employ representation (34) using the same technique as that which was employed for deriving (14). More specifically, equation (34) shows that the \( \{Y_t\} \) process can be expressed as

\[
\left( Y_t - \frac{rd}{1 - a_1 - rb_1} \right) = (a_1 + rb_1)\left( Y_{t-1} - \frac{rd}{1 - a_1 - rb_1} \right) + \epsilon_t - a_1\epsilon_{t-1}. \]

This is an ARMA(1,1) process and therefore when \( 0 < (a_1 + rb_1)^2 + rb_1^2 < 1 \) we have that \( \{Y_t\} \) is second order stationary with autocovariance function

\[
\text{Cov}\{Y_t, Y_{t+h}\} = \begin{cases} 
\frac{(1 - (a_1 + rb_1)^2 + r^2b_1^2)}{1 - (a_1 + rb_1)^2 - rb_1^2} \left( \mathbb{E}[Y_t] + \frac{\mathbb{E}^2[Y_t]}{r} \right), & h = 0, \\
\frac{rb_1(1 - a_1(a_1 + rb_1))}{1 - (a_1 + rb_1)^2 - rb_1^2} \left( \mathbb{E}[Y_t] + \frac{\mathbb{E}^2[Y_t]}{r} \right)(a_1 + rb_1)^{h-1}, & h \geq 1.
\end{cases}
\]
It is clear that the acf of model (33) is equal to (see Appendix)

$$\text{Corr}[Y_t, Y_{t+h}] = \frac{rb_1(1-a_1(a_1 + rb_1))}{1 - (a_1 + rb_1)^2 + r^2b_1^2} (a_1 + rb_1)^{h-1}, \quad h \geq 1. \quad (37)$$

Estimation for model (33) is based on the maximum likelihood method where a profiling procedure is employed. For a grid of values of $r$ (recall that $r$ is positive integer) the negative binomial log–likelihood function is maximized with respect to $(d,a_1,b_1)$. Then, we estimate $r$ by the value which maximizes all log-likelihood functions. For this choice of $r$, the regression parameters are estimated. This method implies that the standard errors of the parameter estimators need to be calculated by resampling methods because the methodology corresponds to a two–stage procedure. The problem of estimation of the parameter $r$ is challenging and the interpretation of its value is unclear. A better way to deal with this issue is to define the negative binomial probability mass function (32) by

$$P[Y = y] = \frac{\Gamma(y + k)}{y!\Gamma(k)} \left( \frac{k}{\lambda + k} \right)^y \left( \frac{\lambda}{\lambda + k} \right)^k, \quad y = 0, 1, 2, \ldots,$$

where $k > 0$. This is a plain consequence of the fact that the negative binomial distribution is a mixture of Poisson random variables. Then we employ model (33)–or its generalization–with obvious modifications. An alternative way to relax the Poisson distributional assumption is given by the double Poisson distribution; Efron (1986). The double Poisson distribution is an exponential combination of two Poisson densities, that is

$$f(y; \lambda, \theta) = C(\lambda, \theta) [\text{Poisson}(\lambda)]^\theta [\text{Poisson}(y)]^{1-\theta},$$

where $\theta$ is a dispersion parameter and $C(\lambda, \theta)$ is the normalizing constant. It can be shown that

$$\frac{1}{C(\lambda, \theta)} \approx 1 + \frac{1 - \theta}{12\theta\lambda} \left( 1 + \frac{1}{\theta\lambda} \right),$$

and that the mean and variance of the double Poisson distribution are approximately equal to $\lambda$ and $\lambda/\theta$, respectively. For the Double Poisson model, we can use models such as (12) to model the mean process, see Kedem and Fokianos (2002, Sec. 4.6, Problem 4) and Heinzel (2003). Properties of maximum likelihood estimators derived by imposing either the negative binomial distribution or the double Poisson distribution is a research topic which has not been addressed suitably in the literature. As a final remark, we note that several other alternative distributional assumptions can be adopted along the previous lines; for example data can be modeled by means of the zero–inflated Poisson model (Lambert (1992)), or the truncated Poisson model (Fokianos (2001)), and so on. However the likely gains of such approaches will depend, in general, upon the context of their application.

4.2 Parameter Driven Models

So far we have discussed models that fall under the framework of observation driven model. This implies that even though the mean process $\{\lambda_t\}$ is not observed directly, it can still be recovered explicitly as
function of the past responses, see equation (15) for example. However, a different point of view has been taken by Zeger (1988), who introduced a regression models for time series of counts by assuming that the observed process is driven by a latent (unobserved) process. To be more specific suppose that, conditional on an unobserved process \( \{ \xi_t, t \geq 1 \} \), \( \{ Y_t, t \geq 1 \} \), is a sequence of independent counts such that

\[
E[Y_t | \xi_t] = \text{Var}[Y_t | \xi_t] = \xi_t \exp(d + a_1 Y_{t-1}).
\]  

(38)

In the above we consider a simple model for illustration, but more complex models which include higher order lagged values of the response and any covariates can be included in (38). Assume that \( \{ \xi_t \} \) is a stationary process with \( E[\xi_t] = 1 \) and \( \text{Cov}[\xi_t, \xi_{t+h}] = \sigma^2 \rho_h \), for \( h \geq 0 \). Then, it can be proved that

\[
E[Y_t] = E[\exp(d + a_1 Y_{t-1})], \quad \text{Cov}[Y_t, Y_{t+h}] = \sigma^2 E[Y_t] E[Y_{t+h}] \rho_h.
\]

It is clear that the above formulation, although similar to a Poisson–loglinear model, reveals that the observed data are overdispersed. Estimation of all unknown parameters is discussed by Zeger (1988). Further detailed study of model (38) can be found in in Davis et al. (2000) where the authors address the problem of existence of the latent stochastic process \( \{ \xi_t \} \) and derive the asymptotic distribution of the regression coefficients when the latter exist. They also suggest adjustments for the estimators of \( \sigma^2 \) and of the autocovariance. In the context of negative binomial regression, the latent process model (38) has been extended by Davis and Wu (2009). See also Harvey and Fernandes (1989) for a state-space approach with conjugate priors for the analysis of count time series and Jørgensen et al. (1999) for multivariate count longitudinal data. More generally, state space models for count time series are discussed in West and Harrison (1997), Durbin and Koopman (2001) and Cappé et al. (2005) among others.

5 Integer Autoregressive Models

We discuss another class of models for integer valued time series. This class consists of the so-called integer autoregressive models which are constructed by means of the thinning operator. As we shall see, these models can be viewed as a special case of a branching process with immigration; see Kedem and Fokianos (2002, Ch.5) for a detailed account of integer AR and MA processes. The notation \( \{ Y_t \} \) still refers to the response process.

5.1 Branching Processes

An important model for integer-valued time series is the branching process with immigration, also known as the Galton-Watson process with immigration. It is defined by

\[
Y_t = \sum_{j=1}^{Y_{t-1}} X_{t,j} + I_t, \quad t = 1, \ldots,
\]  

(39)
where the initial value \( Y_0 \) is a nonnegative integer-valued random variable, and \( \sum_0^1 \equiv 0 \). The processes \( \{ X_{t,i} \} \) and \( \{ I_i \} \) drive the dynamics of the system and they are mutually independent, independent of \( Y_0 \), and each consisting of iid random variables. This defines a Markov chain \( \{ Y_t \} \) with nonnegative integer states. Model (39) was originally introduced and applied by Smoluchowski in 1916 for studying the fluctuations in the number of particles contained in a small volume in connection with the second law of thermodynamics; see Chandrasekhar (1943). Since then, the process has been applied extensively in biological, sociological and physical branching phenomena, see for instance Kedem and Chiu (1987), Franke and Seligmann (1993), Berglund and Brännäs (2001), Böckenholt (1999) and the review by McKenzie (2003), Weiß (2008) and Jung and Tremayne (2011).

Note that \( Y_t \) is the size of the \( t \)'th generation of a population, \( X_{t,1}, \ldots, X_{t,Y_{t-1}} \) are the offspring of the \( (t-1) \)'st generation, and \( I_i \) is the contribution of immigration to the \( t \)'th generation. An important role in the behavior of \( \{ Y_t \} \) is played by the mean \( m = E[X_{t,i}] \) of the offspring distribution, where the cases \( m < 1, m = 1, m > 1 \) are referred to as subcritical, critical, and supercritical, respectively. In the subcritical case \( \{ Y_t \} \) has a limiting stationary distribution, while in the supercritical case \( \{ Y_t \} \) explodes at an exponential rate. In the critical case the process is either null recurrent or transient.

The process (39) admits a useful autoregressive representation, similar to the model representation (34), for instance. Let \( \lambda = E[I_i] \), and let \( F_t \) be generated by the past information \( Y_0, Y_1, Y_2, \ldots, Y_t \). Then \( E[Y_t | F_{t-1}] = mY_{t-1} + \lambda \). Therefore, with \( \epsilon_t \equiv Y_t - E[Y_t | F_{t-1}] \), the stochastic equation (39) is transformed into a stochastic regression model,

\[
Y_t = mY_{t-1} + \lambda + \epsilon_t, \quad t = 1, \ldots,
\]

as before. The noise process \( \{ \epsilon_t \} \) consists of uncorrelated random variables such that \( E[\epsilon_t] = 0 \). However \( E[\epsilon_t^2 | F_{t-1}] = \text{Var}[X_{t,i}]Y_{t-1} + \text{Var}[I_i] \) is unbounded as \( Y_{t-1} \) increases.

As suggested by (40), the least squares estimators for \( m, \lambda \) are obtained by minimizing,

\[
\sum_{t=1}^n \epsilon_t^2 = \sum_{t=1}^n (Y_t - mY_{t-1} - \lambda)^2,
\]

and are given by

\[
\hat{m} = \frac{\sum Y_t \sum Y_{t-1} - n \sum Y_{t-1} Y_{t-1}}{(\sum Y_{t-1})^2 - n \sum Y_{t-1}^2}, \quad \hat{\lambda} = \frac{\sum Y_{t-1} Y_t \sum Y_{t-1} - \sum Y_{t-1}^2 \sum Y_t}{(\sum Y_{t-1})^2 - n \sum Y_{t-1}^2},
\]

where the summation limits are from \( t = 1 \) to \( t = n \). It turns out that \( \hat{m} \) is consistent in all three cases, while \( \hat{\lambda} \) is not consistent in the critical and supercritical cases. Improved estimators are obtained by weighted least squares. We write (40) as

\[
\frac{Y_t}{\sqrt{Y_{t-1} + 1}} = m\sqrt{Y_{t-1} + 1} + \frac{(\lambda - m)}{\sqrt{Y_{t-1} + 1}} + \frac{\epsilon_t}{\sqrt{Y_{t-1} + 1}},
\]

(41)
and estimate $m$ and $\lambda - m$ by minimizing $\sum \delta_i^2$ where $\delta_i = \epsilon_i / \sqrt{Y_{i-1} + 1}$ to obtain (see Winnicki (1986)),

$$\hat{m} = \frac{\sum Y_i \sum \frac{1}{Y_{i-1} + 1} - n \sum \frac{Y_i}{Y_{i-1} + 1}}{\sum(Y_{i-1} + 1) \sum \frac{1}{Y_{i-1} + 1} - n^2},$$

$$\hat{\lambda} = \frac{\sum Y_{i-1} \sum \frac{1}{Y_{i-1} + 1} - \sum Y_i \sum \frac{Y_i - 1}{Y_{i-1} + 1}}{\sum(Y_{i-1} + 1) \sum \frac{1}{Y_{i-1} + 1} - n^2},$$

where again the summation limits are from 1 to $n$. Then for $0 < m < \infty$, $\hat{m} \rightarrow m$ in probability. That is $\hat{m}$ is consistent in all cases, provided that $m > 0$. Furthermore, the limiting distribution of $\hat{m}$ is normal in noncritical cases and nonnormal in the critical case. On the other hand, $\hat{\lambda}$ is consistent for $m \leq 1$, but not for $m > 1$, and is asymptotically normal when $m < 1$ or $m = 1$ and $2\lambda > \text{Var}[Y_n]$; Wei and Winnicki (1990), Winnicki (1986).

### 5.2 Thinning Operator Based Models

In this section, we review models that are based on the thinning operator. The thinning operator is defined as follows; see Steutel and van Harn (1979). Suppose that $Y$ is a non-negative integer random variable and let $\alpha \in [0, 1]$. Then, the thinning operator, denoted by $\circ$, is defined as

$$\alpha \circ Y = \sum_{i=1}^{Y} X_i$$

where $\{X_i\}$ is a sequence of iid Bernoulli random variables--independent of $Y$--with success probability $\alpha$. The sequence $\{X_i\}$ is termed a counting series. The random variable $\alpha \circ Y$ counts the number of successes in a random number of Bernoulli trials where the probability of success $\alpha$ remains constant throughout the experiment. Therefore, given $Y = y$, the random variable $\alpha \circ Y$ follows the binomial distribution with parameters $y$ and $\alpha$.

It turns out that the thinning operator is quite useful for modeling count time series. Building a model for count time series is based on a typical autoregressive model where scalar multiplication is replaced by thinning operators, see McKenzie (1985, 1986, 1988), Al-Osh and Alzaid (1987); Alzaid and Al-Osh (1990), Du and Li (1991). Let us consider the simple integer autoregressive model of order 1, which is abbreviated by INAR(1). The INAR(1) model is a special case of the branching process with immigration (39). However, it deserves special consideration due to the thinning operation or calculus. Suppose that $a_1 \in (0, 1)$ and let $\{\epsilon_t\}$ be a sequence of iid nonnegative integer valued random variables with $E[\epsilon_t] = \mu$ and $\text{Var}[\epsilon_t] = \sigma^2$. The integer autoregressive process of order 1, $\{Y_t, t \geq 1\}$, is defined as

$$Y_t = a_1 \circ Y_{t-1} + \epsilon_t, \quad t \geq 1,$$
where \( a_1 \circ Y_{t-1} \) is the sum of \( Y_{t-1} \) Bernoulli random variables all of which are independent of \( Y_{t-1} \). It should be noted that the Bernoulli variables used in \( a_1 \circ Y_{t-1} \) are independent of those used in \( a_1 \circ Y_{t-2} \), and so on. This is the assumption imposed by Du and Li (1991) and used subsequently in the majority of all published work related to integer autoregressive processes. Clearly, (44) is a special case of (39).

Employing the same techniques that we used before (or by repeated substitution into (44) and use of the properties of the thinning operator), we obtain that the mean, variance and acf of the INAR(1) are given by

\[
E[Y_t] = \frac{\mu}{1 - a_1}, \quad \text{Var}[Y_t] = \frac{a_1 \mu + \sigma^2}{1 - a_1^2}, \quad \text{Cov}[Y_t, Y_{t+h}] = a_1^h, \quad h \geq 1. \tag{45}
\]

Note that the acf decays exponentially with the lag \( h \) as in AR(1) models, but unlike the autocorrelation of a stationary AR(1) process, it is always positive for \( a_1 \in (0, 1) \). Furthermore, under suitable conditions, it can be shown that \( Y_t \) has a discrete self-decomposable distribution. This, in turn, implies unimodality properties and characterization of the distribution of \( Y_t \) through the sequence \( \{\epsilon_t\} \). For instance, we can obtain the result that \( Y_t \) follows the Poisson distribution if and only if \( \epsilon_t \) follows the Poisson distribution; see Al-Osh and Alzaid (1987).

Estimation in INAR(1) means estimation in the branching process with immigration in the subcritical case, and this has already been discussed earlier. Still it is interesting to note a few facts regarding estimation in the Poisson INAR(1). Estimation procedures for the parameters \( a_1 \) and \( \mu \) of the INAR(1) model (44) assuming that the sequence \( \{\epsilon_t\} \) follows the Poisson distribution, has been discussed in Al-Osh and Alzaid (1987). Imposing the Poisson assumption on the distribution of the error sequence \( \{\epsilon_t\} \), and employing equation (45) yields a method of moments estimators for \( a_1 \) and \( \mu \) (in this case \( \mu = \sigma^2 \)), given by

\[
\hat{a}_1 = \frac{\sum_{t=0}^{n-1} (Y_t - \bar{Y})(Y_{t+1} - \bar{Y})}{\sum_{t=0}^{n-1} (Y_t - \bar{Y})^2}, \quad \hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} \hat{\epsilon}_t,
\]

where \( \hat{\epsilon}_t = Y_t - \hat{a}_1 Y_{t-1} \), for \( t = 1, \ldots, n \). Alternatively, we can consider the conditional least squares of the parameters \( a_1 \) and \( \mu \), i.e. the values that minimize the residual sum of squares

\[
\sum_{t=1}^{N} (X_t - a_1 X_{t-1} - \mu)^2.
\]

Asymptotic properties of the resulting estimators are deduced by using classical results from Klimko and Nelson (1978). In Ispány et al. (2003), the authors study the INAR(1) model where the autoregressive coefficient \( a_1 = a_{1n} \) satisfies \( a_{1n} = 1 - \gamma_n/n \) with \( \gamma_n \to \gamma > 0 \). Such a sequence is called nearly unstable. The authors show that it can be approximated, in an appropriate sense, by a Gaussian martingale and use this result to show that the conditional least squares estimator of \( a_1 \) is asymptotically normal with the rate of convergence \( n^{3/2} \). As a final remark we note that application of maximum likelihood estimation requires a full distributional assumption about the innovations. With the Poisson assumption, the
likelihood function of a time series $Y_0, Y_1, \ldots, Y_n$ from model (44) is

$$\left( \prod_{i=1}^{N} P_i(Y_i) \right) \left( \frac{\mu / (1 - a_1)}{Y_0!} \right) \exp \left( -\frac{\mu}{(1 - a_1)} \right),$$

$$P_i(y) = \exp \left( -\mu \right) \sum_{i=0}^{\min(Y_i, Y_{i-1})} \frac{\mu^{y-i}}{(y-i)!} \left( \frac{Y_{i-1}}{a_i} \right) a_i^i (1 - a_1)^{Y_{i-1} - i}, \quad t = 1, 2, \ldots, n.$$ 

More generally, the $p$'th order model, abbreviated by INAR($p$), is defined as

$$Y_t = \sum_{i=1}^{p} a_i \circ Y_{t-i} + \epsilon_t,$$

(46)

where $\{\epsilon_t\}$ is a sequence of iid nonnegative integer valued random variables with mean $\mu$ and variance $\sigma^2$ and all $p$ thinning operations are independent of each other; existence and generalizations of INAR($p$) are studied by Latour (1997, 1998) while unifying work based on convolution is presented by Joe (1996). A unique stationary and ergodic solution of (46) exists if

$$\sum_{i=1}^{p} a_i < 1.$$  

(47)

Estimation for INAR($p$) models is based on the same methods as those described for the INAR(1) model. However, in a recent contribution, Drost et al. (2009) consider the problem of semiparametric maximum likelihood estimation for INAR($p$) models. In other words, the authors estimate both the finite dimensional parameters of the model plus the unknown cumulative distribution of the residual process and obtain efficient estimators. See also Jung and Tremayne (2006), Neal and Subba Rao (2007), Bu et al. (2008) and McCabe et al. (2011) for further results on estimation and prediction.

5.3 Extensions of Thinning Operator Based Models

A generalization of binomial thinning is given in Joe (1996), who considers the following model

$$Y_t = A_t(Y_{t-1}; a) + \epsilon_t, \quad t = 1, 2, \ldots,$$

where $A_t(\cdot)$ is a random transformation, and $A_t(Y_{t-1}; a)$ and $\epsilon_t$ are independent. Based on this general thinning, a class of stationary moving average processes with margins in the class of infinitely divisible exponential dispersion models was introduced in Jørgensen and Song (1998).

Another generalization is that of the first-order conditional linear autoregressive process, abbreviated by CLAR(1),

$$m(Y_{t-1}) = a_1 Y_{t-1} + \mu,$$

where $m(Y_{t-1}) = E[Y_t | Y_{t-1}]$, and $a_1, \mu$ are real numbers. The CLAR(1) class includes many of the non-Gaussian AR(1) models proposed in the literature and allows various generalizations of previous
results; see Grunwald et al. (2000). Interestingly, when \(|a_1| < 1\), the acf of the CLAR(1) model is equal to \(a_1^h\), \(h = 1, 2, \ldots\), as in other first-order autoregressive processes including the branching process with immigration (39).

Non-negative integer-valued bilinear processes have been defined and studied by Doukhan et al. (2006); see also Latour and Truquet (2008). These processes are given by

\[
Y_t = \sum_{i=1}^{p} a_i \circ Y_{t-i} + \sum_{j=1}^{d} c_j \epsilon_{t-j} + \sum_{k=1}^{m} \sum_{l=1}^{n} b_{lk} \circ (Y_{t-k}\epsilon_{t-l}) + \epsilon_t,
\]

where all thinning operators are defined independently of each other and \(\{\epsilon_t\}\) is a sequence of iid nonnegative integer valued random variables. Furthermore, Drost et al. (2008) consider some special cases of the above model. Random coefficient integer-valued autoregressive models have been proposed in Zheng et al. (2006) and Zheng et al. (2007). For instance, the random coefficient model of order 1 is given by

\[
Y_t = a_{1t} \circ Y_{t-1} + \epsilon_t,
\]

where now \(\{a_{1t}\}\) is an sequence of independent and identically distributed random variables, independent of the noise \(\{\epsilon_t\}\). Multivariate INAR type of models have been considered by Franke and Rao (1995) who also discuss stationarity conditions and the properties of the maximum likelihood estimator for the first order multivariate model. For several new results regarding inference for multivariate models, see the thesis by Pedeli (2011).

5.4 Renewal Process Models

In a recent contribution by Cui and Lund (2009), the authors propose a new and simple model for stationary time series of integer counts. Their methodology does not resort to thinning operations. Instead they use a renewal process to generate a time-correlated sequence of Bernoulli trials. It turns out that superposition of iid such processes, yields stationary processes with binomial, Poisson, geometric, or any other discrete marginal distribution. Apparently, this new model class of non-Markov model offers parsimony and easily produces series with either short or long memory autocovariances. The model can be fitted with linear prediction techniques for stationary series.

6 Conclusions

In general, count time series refers to stochastic processes whose state space is a countable set. Although probabilistic properties of these processes are well understood, (e.g. Billingsley (1961) and Meyn and Tweedie (1993)), it is still not clear what conditions should be met for valid parametric modeling, especially when the time series is observed jointly with covariates or its behavior is driven by an unobserved
process. The parametric framework allows for estimation, model assessment and forecasting by employing existing statistical software. These facts make a strong case in favor of this approach and advocate the point of view that a successful approach towards the resolution of the aforementioned problems is via the theory of GLM as advanced by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). In this contribution, we have surveyed the most commonly used models related to the regression of count time series. The foundation for this class of models is Poisson regression. However, models like (12) can be developed within the context of other distributional assumptions, like the negative binomial or other discrete distributions. The literature, both applied and theoretical, on this subject is growing fast. For instance, Andersson and Karlis (2010) consider methods for estimating the parameters of the first-order integer-valued autoregressive model in the presence of missing data and Monteiro et al. (2010) study the periodic integer-valued autoregressive model of order one. In Fokianos and Fried (2010), the authors introduce the concept of intervention for the linear model (6) and discuss estimation of the intervention size as well as testing for its existence. Another problem of current interest is the analysis of multivariate count data, see Jung et al. (2011) who propose a dynamic factor model for the analysis of number of trades for five stocks from two industrial sectors.

Appendix

Proof of (35) and (37)

Consider the negative binomial regression model (33) and note that \( E[Y_t] = rd/(1 - a_1 - rb_1) \). Then assuming second order stationarity we obtain that

\[
\sigma^2 = \text{Var}[e_t] = rE[\lambda_t + \lambda_t^2].
\]

But \( E[\lambda_t] = E[Y_t]/r \). Hence, we need to calculate \( E[\lambda_t^2] = \mu^{(2)}_\lambda \). Consider the state equation of (33) to obtain

\[
\mu^{(2)}_\lambda = E\left(d + a_1 \lambda_{t-1} + b_1 Y_{t-1}\right)^2
= E\left(d + (a_1 + rb_1) \lambda_{t-1} + b_1 (Y_{t-1} - r \lambda_{t-1})\right)^2
= \mu^{(2)}_\lambda + d^2 + (a_1 + rb_1)^2 \mu^{(2)}_\lambda + 2d(a_1 + rb_1)E[\lambda_t]
= \left((a_1 + rb_1)^2 + rb_1^2\right) \mu^{(2)}_\lambda + d^2 + \left(rb_1^2 + 2d(a_1 + rb_1)\right)E[\lambda_t].
\]

Therefore, we obtain that

\[
\mu^{(2)}_\lambda = \frac{d^2 + \left(rb_1^2 + 2d(a_1 + rb_1)\right)E[\lambda_t]}{1 - (a_1 + rb_1)^2 - rb_1^2}.
\]
Plugging this expression into the definition of $\sigma^2$ and using the fact that $d = (1 - a_1 - rb_1)E[\lambda]$, yields to (35). Formula (37) is proved by employing representation (36) and well known results about the acf.

**Acknowledgements**

Part of this work was completed while K. Fokianos was visiting the Department of Mathematics, University of Cergy-Pontoise. The hospitality of P. Doukhan and G. Lang is greatly appreciated. Special thanks to V. Christou, M. Neumann, B. Kedem, S. Kitromilidou and an anonymous reviewer for several suggestions which improved the presentation.

**References**


