2

Markov Models for Count Time Series

Harry Joe

CONTENTS

2.1 Introduction ................................................................. 29
2.2 Models for Count Data ...................................................... 30
2.3 Thinning Operators .......................................................... 32
  2.3.1 Analogues of Gaussian AR(p) ........................................ 34
  2.3.2 Analogues of Gaussian Autoregressive Moving Average .... 36
  2.3.3 Classes of Generalized Thinning Operators ..................... 36
  2.3.4 Estimation ............................................................... 37
  2.3.5 Incorporation of Covariates ......................................... 38
2.4 Operators in Convolution-Closed Class .............................. 38
2.5 Copula-Based Transition ................................................. 41
2.6 Statistical Inference and Model Comparisons ...................... 44
References ........................................................................... 47

2.1 Introduction

The focus of this chapter is on the construction of count time series models based on thinning operators or a joint distribution on consecutive observations, and comparison of the properties of the resulting models.

The models for count time series considered here are mainly intended for low counts with the possibility of 0. If all counts are large and “far” from 0, the models considered here can be used as well as models that treat the count response as continuous.

Count data are often overdispersed relative to Poisson. There are many count regression models with covariates, examples are regression models with negative binomial (NB), generalized Poisson (GP), zero-inflated Poisson, zero-inflated NB, etc.

If the count data are observed as a time series sequence, then the count regression model can be adapted in two ways: (1) add previous observations as covariates and (2) make use of some models for stationary count time series. Methodology for case (1) is covered in Davis et al. (2000) and Fokianos (2011), and here, we discuss the quite different methodology for case (2). The advantage of a time series regression model with univariate margins corresponding to a count regression model is that predictions as a function of covariates can be made with or without preceding observations. That is, this is useful if one is primarily interested in regression but with time-dependent observations.
Common parametric models for count regression are NB and GP, and these include Poisson regression at the boundary. In this chapter, we use these two count regression models for concreteness, but some approaches, such as copula-based models, can accommodate other count distributions.

The remainder of the chapter is organized as follows. Section 2.2 summarizes some count regression models and contrasts some properties of count time series models constructed under different approaches. Sections 2.3 through 2.5 provide some details for count time series models based, respectively, on thinning operators, multivariate distributions with random variables in a convolution-closed infinitely divisible class, and copulas for consecutive observations. Section 2.6 compares the fits of different models for one data set.

Some conventions and notation that are subsequently used are as follows: $f$ is used for probability mass functions (pmf) and $F$ is used for cumulative distribution functions (cdfs) with the subscript used to indicate the margin or random vector; $\sum_{i=1}^{y} k_i = 0$ when $y = 0$; $N_0$ is the set of nonnegative integers; $\epsilon_t$ is used for the innovation at time $t$ (that is, $\epsilon_t$ is independent of random variables at time $t-1, t-2, \ldots$ in the stochastic representation).

### 2.2 Models for Count Data

In this section, we show how NB and GP distributions have been used for count regression and count time series.

NB and GP regression models are nonunique in how regression coefficient $\beta$s are introduced into the univariate parameters. If the mean is assumed to be loglinear in covariates, there does not exist a unique model because the mean involves the convolution parameter and a second parameter that links to overdispersion.

Brief details are summarized as follows, with $F_{NB}$ and $F_{GP}$ denoting the cdfs and $f_{NB}$ and $f_{GP}$ denoting the pmfs:

1. (NB): $\theta$ convolution parameter, $\pi$ probability parameter, $\xi = (\pi^{-1} - 1) \geq 0$; mean $\mu = \theta \xi = \theta (1 - \pi)/\pi$, variance $\sigma^2 = \mu(1 + \xi) = \theta (1 - \pi)/\pi$, and
   
   $$f_{NB}(y; \theta, \xi) = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} \frac{\xi^y}{(1 + \xi)^{\theta + y}}, \quad y = 0, 1, 2, \ldots, \quad \theta > 0, \xi > 0.$$  

   If $\theta \to \infty$, $\xi \to 0$ with $\theta \xi$ fixed, the Poisson distribution is obtained.

2. (GP): $\theta$ convolution parameter, second parameter $0 \leq \eta < 1$; mean $\mu = \theta/(1 - \eta)$, variance $\sigma^2 = \theta/(1 - \eta)^3$, and
   
   $$f_{GP}(y; \theta, \eta) = \frac{\theta(\theta + \eta y)^{y-1}}{y!} e^{-\theta - \eta y}, \quad y = 0, 1, 2, \ldots, \quad \theta > 0, 0 \leq \eta < 1.$$  

   If $\eta = 0$, the Poisson distribution is obtained.

Cameron and Trivedi (1998) present the NB$k(\mu, \gamma)$ parametrization where $\theta = \mu^{2-k}\gamma^{-1}$ and $\xi = \mu^{k-1}\gamma$, $1 \leq k \leq 2$. For the NB$k$ model, $\log \mu = z^T \beta$ depends on the covariate vector $z$, and either or both $\theta$ and $\xi$ are covariate dependent. For the NB1 parametrization:
\( k = 1, \theta = \mu \gamma^{-1}, \xi = \gamma \); that is, \( \theta \) depends on covariates, and the dispersion index \( \xi = \gamma \) is constant. For the NB2 parametrization: \( k = 2, \theta = \gamma^{-1}, \xi = \mu \gamma \) and this is the same as in Lawless (1987); that is, \( \theta \) is constant and \( \xi \) is a function of the covariates, and the dispersion index varies with the covariates. For \( 1 < k < 2 \), one could interpolate between these two models using the NB\( k \) parametrization. Similarly, GP1 and GP2 regression models can be defined.

Next, we consider stationary time series \( \{Y_t: t = 1, 2, \ldots\} \), where the stationary distribution is NB, GP, or general FY.

A Markov order 1 time series can be constructed based on a common joint distribution \( F_{12} \) for \( (Y_{t-1}, Y_t) \) for all \( t \) with marginal cdfs \( F_1 = F_2 = F_Y = F_{NB}(\cdot; \theta, \xi) \) or \( F_{GP}(\cdot; \theta, \eta) \) (or another parametric univariate margin). Let \( f_{12} \) and \( f_Y \) be the corresponding bivariate and univariate pmfs. The Markov order 1 transition probability is

\[
Pr(Y_t = y_{\text{new}} | Y_{t-1} = y_{\text{prev}}) = \frac{f_{12}(y_{\text{prev}}, y_{\text{new}})}{f_Y(y_{\text{prev}})}
\]

A Markov order 2 time series can be constructed based on a common joint distribution \( F_{123} \) for \( (Y_{t-2}, Y_{t-1}, Y_t) \) for all \( t \) with univariate marginal cdfs \( F_1 = F_2 = F_3 \) and bivariate margins \( F_{12} = F_{23} \). The ideas extend to higher-order Markov. However for count time series with small counts, simpler models are generally adequate for forecasting.

There are two general approaches to obtain the transition probabilities; the main ideas can be seen with Markov order 1.

1. Thinning operator for Markov order 1 dependence: \( Y_t = R_t(Y_{t-1}; \alpha) + \epsilon_t(\alpha), \) \( 0 \leq \alpha \leq 1, \) where \( R_t \) are independent realizations of a stochastic operator, the \( \epsilon_t \) are appropriate innovation random variables, and typically \( E[R_t(y; \alpha)|Y_{t-1} = y] = \alpha y \) for \( y = 0, 1, \ldots \).

2. Copula-based transition probability from \( F_{12} = C(F_Y, F_Y; \delta) \) for a copula family \( C \) with dependence parameter \( \delta \).

The review paper McKenzie (2003) has a section entitled “Markov chains” but copula-based transition models were not included. Copulas are multivariate distributions with \( U(0,1) \) margins and they lead to flexible modeling of multivariate data with the dependence structure separated from the univariate margins. References for use of copula models are Joe (1997) and McNeil et al. (2005).

Some properties and contrasts are summarized below, with details given in subsequent sections. Weiß (2008) has a survey of many thinning operators for count time series models, and Fokianos (2012) has a survey of models based on thinning operators and conditional Poisson. Some references where copulas are used for transition probabilities are Joe (1997) (Chapter 8), Escarela et al. (2006), Biller (2009), and Beare (2010).

For thinning operators, the following hold:

- The stationary margin is infinitely divisible (such as NB, GP).
- The serial correlations are positive.
- The operator is generally interpretable and the conditional expectation is linear.
- For extension to include covariates (and/or time trends), the ease depends on the operator; covariates can enter into a parameter for the innovation distribution, but in this way, the marginal distribution does not necessarily stay in the same family.
• For extension to higher Markov orders, there are “integer autoregressive” models, such as INAR($p$) in Du and Li (1991) or GINAR($p$) in Gauthier and Latour (1994), and constructions as in Lawrance and Lewis (1980), Alzaid and Al-Osh (1990), and Zhu and Joe (2006) to keep margins in a given family. Without negative serial correlations, the range of autocorrelation functions is not as flexible as the Gaussian counterpart.

• Numerical likelihood inference is simple if the transition probability has closed form; for some operators, only the conditional probability generating function (pgf) has a simple form and then the approach of Davies (1973) can be used to invert the pgf. Conditional least squares (CLS) and moment methods can estimate mean parameters but are not reliable for estimating the overdispersion parameter.

• There are several different thinning operators for NB or GP and these can be differentiated based on the conditional heteroscedasticity $\text{Var}(Y_t|Y_{t-1} = y)$.

• Although thinning operators can be used in models that are analogues of Gaussian AR($p$), MA($q$), and ARMA models, not as much is known about probabilistic properties such as stationary distributions.

For copula-based transition, the following hold:

• The stationary margin can be anything, and positive or negative serial dependence can be attained by choosing appropriate copula families.

• The conditional expectation is generally nonlinear and different patterns are possible. The tail behavior of the copula family affects the conditional expectation and variance for large values.

• It is easier to combine the time series model with covariates in a univariate count regression model.

• The extension from Markov order 1 to higher-order Markov is straightforward.

• Theoretically, the class of autocorrelation functions is much wider than those based on thinning operators.

• Likelihood inference is easy if the copula family has a simple form.

• The Gaussian copula is a special case; for example, autoregressive-to-anything (ARTA) in Biller and Nelson (2005).

• As a slight negative compared with thinning operators, for NB/GP, the copula approach does not use any special univariate property.

### 2.3 Thinning Operators

This section has more detail on thinning operators. Operators are initially presented and discussed without regard to stationarity and distributional issues. Notation similar to that in Jung and Tremayne (2011) is used.

A general stochastic model is

$$Y_t = R_t(Y_{t-1}, Y_{t-2}, \ldots) + \epsilon_t,$$
where $\epsilon_t$ is the innovation random variable at time $t$ and $R_t$ is a random variable that depends on the previous observations. In order to get a stationary model with margin $F_Y$, the choice of distribution for $\{\epsilon_t\}$ depends on the distribution of $R_t(Y_{t-1}, Y_{t-2}, \ldots)$. If one is not aiming for a specific stationary distribution, there is no constraint on the distribution of $\{\epsilon_t\}$.

If $R_t(Y_{t-1}, Y_{t-2}, \ldots) = R_t(Y_{t-1})$, then a Markov model of order 1 is obtained and if $R_t(Y_{t-1}, Y_{t-2}, \ldots) = R_t(Y_{t-1}, Y_{t-2})$, then a Markov model of order 2 is obtained, etc.

For Markov order 1 models, the conditional pmf of $[R(Y)|Y = y]$ is the same as the pmf of $R(y)$ and the unconditional pmf of $R(Y)$ is

$$f_{R(Y)}(x) = \sum_{y=0}^{\infty} f_{R(y)}(x) f_Y(y).$$

The following classes of operators are included in the review in Weiβ (2008):

- Binomial thinning (Steutel and Van Harn 1979): $R(y) = \alpha \circ y \sim \sum_{i=1}^{y} I_i(\alpha)$ has a $\text{Bin}(y, \alpha)$ distribution, where $I_1(\alpha), I_2(\alpha), \ldots$ are independent Bernoulli random variables with mean $\alpha \in (0, 1)$. Hence $E[R(y)] = \alpha y$ and $\text{Var}(R(y)) = \alpha(1 - \alpha)y$.

- Expectation or generalized thinning (Latour, 1998; Zhu and Joe, 2010a): $R(y) = K(\alpha) \circ y \sim \sum_{i=1}^{y} K_i(\alpha)$ where $K_1(\alpha), K_2(\alpha), \ldots$ are independent random variables that are replicates of $K(\alpha)$ which has support on $\mathbb{N}_0$ and satisfies $E[K(\alpha)] = \alpha \in [0, 1]$. For the boundary case $K(0) \equiv 0$ and $K(1) \equiv 1$. Hence $E[R(y)] = \alpha y$ and $\text{Var}(R(y)) = y \text{Var}[K(\alpha)]$.

- Random coefficient thinning (Zheng et al., 2006, 2007): $R(y) = A \circ y \sim \sum_{i=1}^{y} I_i(A)$ where $A$ has support on $(0, 1)$ and given $A = a$, $I_i(a)$ are independent Bernoulli(a) random variables. Hence $\text{Pr}(A \circ y = j) = \int_{0}^{1} \binom{y}{j} a^j (1 - a)^{y-j} dF_A(a)$. If $A$ has mean $\alpha$, then $E[R(y)] = E[Ay] = \alpha y$ and $\text{Var}(R(y)) = E[A(1 - A)y] + \text{Var}(Ay) = \alpha(1 - \alpha)y + y(y - 1) \text{Var}(A)$.

Interpretations are provided later, with a subscript on the thinning operator to emphasize that thinnings are performed at time $t$.

- Time series based on binomial thinning:

$$Y_t = \alpha \circ_t Y_{t-1} + \epsilon_t = \sum_{i=1}^{Y_{t-1}} I_i(\alpha) + \epsilon_t,$$

where the $I_i(\alpha)$ are independent over $t$ and $i$. It can be considered that $\alpha \circ_t Y_{t-1}$ consists of the “survivors” (continuing members) from time $t - 1$ to time $t$ (with each individual having a probability $\alpha$ of continuing), and $\epsilon_t$ consists of the “newcomers” (innovations) at time $t$.

- Time series based on generalized thinning:

$$Y_t = K(\alpha) \circ_t Y_{t-1} + \epsilon_t = \sum_{i=1}^{Y_{t-1}} K_i(\alpha) + \epsilon_t$$ (2.2)
where the $K_{tji}(\alpha)$ are independent over $t$ and $i$. This can be viewed as a dynamic system so that $K(\alpha) \circ_t Y_{t-1}$ is a sum where each countable unit at time $t - 1$ may be absent, present, or split into more than one new unit at time $t$, and $\epsilon_t$ consists of the new units at time $t$. Also $K(\alpha) \circ y$ can be considered as a compounding or branching operator, and the time series model can be considered as a branching process model with immigration.

- Time series based on random coefficient thinning: this is random binomial thinning, where the chance of survival to the next time is a random variable that depends on $t$.

$$Y_t = A_t \circ Y_{t-1} + \epsilon_t = \sum_{i=1}^{Y_{t-1}} I_{ti}(A_t) + \epsilon_t, \quad (2.3)$$

A beta-binomial thinning operator based on the construction in Section 2.4 fits within this class.

Because all of the above models have a conditional expectation that is linear in the previous observation, they have been called integer-autoregressive models of order 1, abbreviated INAR(1). The models are not truly autoregressive in the sense of linear in the previous observations (because such an operation would not preserve the integer domain).

### 2.3.1 Analogues of Gaussian AR($p$)

An extension of (2.3) to a higher-order Markov time series model is given in Section 2.4 for one special case. Otherwise, binomial thinning is a special case of generalized thinning. We next extend (2.2) to higher-order Markov:

$$Y_t = \sum_{j=1}^{p} K(\alpha_j) \circ_t Y_{t-j} + \epsilon_t = \sum_{j=1}^{p} \sum_{i=1}^{Y_{t-j}} K_{tji}(\alpha_j) + \epsilon_t, \quad (2.4)$$

where $0 \leq \alpha_j \leq 1$ for $j = 1, \ldots, p$ and the $K_{tji}(\alpha_j)$ are independent over $t, j$ and $i$, and $\epsilon_t$ is the innovation at time $t$. This is called GINAR($p$) in (Gauthier and Latour 1994; Latour 1997, 1998). It can also be interpreted as a branching process model with immigration, where a unit at time $t$ has independent branching at times $t + 1, \ldots, t + p$. The most common form of INAR($p$) in the statistical literature involves the binomial thinning operator; see Du and Li (1991). For the binomial thinning operator, Alzaid and Al-Osh (1990) define INAR($p$) in a different way from the above with a conditional multinomial distribution for $(\alpha_1 \circ Y_t, \ldots, \alpha_p \circ Y_t)$. Because the survival/continuation interpretation for (2.1) does not extend to second and higher orders, it is better to consider (2.4) with more general thinning operators; if the $K_{tji}$ are Bernoulli random variables, this can still be interpreted as a branching process model with immigration (with limited branching).

More specifically, for a GINAR(2) model based on compounding, unit $i$ at time $t'$ contributes $K_{t'+1,i}(\alpha_1)$ units to the next time and $K_{t'+2,i}(\alpha_2)$ units in two time steps. That is, at time $t$, the total count comes from branching of units at times $t - 1$ and $t - 2$ plus the innovation count.
It will be shown below that the GINAR($p$) model has an overdispersion property if \( \{K(\alpha)\} \) satisfies \( \text{Var}[K(\alpha)] = a_k \alpha (1 - \alpha) \) where \( a_k \geq 1 \). A sufficient condition for this is the self-generalizability of \( \{K(\alpha)\} \) (defined below in Section 2.3.3).

Let \( R_t(y_{t-1}, \ldots, y_{t-p}) = \sum_{j=1}^{p} y_{t-j} K_{tj} (\alpha_j) \). Then its conditional mean and variance are \( \sum_{j=1}^{p} \alpha_j y_{t-j} \) and \( \sum_{j=1}^{p} \text{Var}[K(\alpha_j)] \), respectively. That is, this GINAR($p$) model has linear conditional expectation and variance, given previous observations. The mean is the same for all \( \{K(\alpha)\} \) that have \( \text{E}[K(\alpha)] = \alpha \), but the conditional variance depends on the family of \( \{K(\alpha)\} \). With a self-generalized family, the conditional variance is \( a_k \sum_{j=1}^{p} y_{t-j} \alpha_j (1 - \alpha_j) \) so that different families of \( \{K(\alpha)\} \) lead to differing amounts of conditional heteroscedasticity, and a larger value of \( a_k \) leads to more heteroscedasticity.

The condition for stationarity of (2.4) is \( \sum_{j=1}^{p} \alpha_j < 1 \). In this case, in stationary state, the equations for the mean and variance lead to

\[
\mu_Y = \mu_Y \sum_{j=1}^{p} \alpha_j + \mu_{\epsilon}, \quad \mu_Y = \frac{\mu_{\epsilon}}{(1 - \alpha_1 - \cdots - \alpha_p)}, \tag{2.5}
\]

and

\[
\sigma_Y^2 = \sigma_Y^2 \left[ \sum_{j=1}^{p} \alpha_j^2 + 2 \sum_{1 \leq j < k \leq p} \rho_{k-j} \alpha_j \alpha_k \right] + \mu_Y \sum_{j=1}^{p} \sigma_{\epsilon}(\alpha_j) + \sigma_{\epsilon}^2, \tag{2.6}
\]

where \( \rho_{\ell} \) is the autocorrelation at lag \( \ell \). If the innovation is overdispersed relative to Poisson (that is, \( \sigma_{\epsilon}^2 / \mu_{\epsilon} \geq 1 \)), then we show that the stationary distribution of \( Y \) is also overdispersed. From (2.5) and (2.6), and assuming \( \sigma_{\epsilon}(\alpha) = a_k \alpha (1 - \alpha) \),

\[
\frac{\sigma_Y^2}{\mu_Y} = \frac{a_k \sum_{j=1}^{p} \alpha_j (1 - \alpha_j) + \sigma_{\epsilon}^2 \left[ 1 - \sum_{j=1}^{p} \alpha_j \right]}{1 - \sum_{j=1}^{p} \alpha_j^2 - 2 \sum_{1 \leq j < k \leq p} \rho_{k-j} \alpha_j \alpha_k} \geq \frac{\sum_{j=1}^{p} \alpha_j (1 - \alpha_j) + \left[ 1 - \sum_{j=1}^{p} \alpha_j \right]}{1 - \sum_{j=1}^{p} \alpha_j^2 - 2 \sum_{1 \leq j < k \leq p} \rho_{k-j} \alpha_j \alpha_k}
\]

\[
= \frac{1 - \sum_{j=1}^{p} \alpha_j^2}{1 - \sum_{j=1}^{p} \alpha_j^2 - 2 \sum_{1 \leq j < k \leq p} \rho_{k-j} \alpha_j \alpha_k} \geq 1,
\]

because \( \rho_{\ell} \geq 0 \) and the denominator is positive. The inequality is strict for \( p > 1 \) with \( \rho_1 > 0 \). If the innovation is Poisson with \( \sigma_{\epsilon}^2 / \mu_{\epsilon} = 1 \) and \( a_k = 1 \) for binomial thinning, then one still has \( \sigma_Y^2 / \mu_Y > 1 \) for \( p > 1 \), so that the stationary distribution cannot be Poisson. A Markov model of order \( p \) with stationary Poisson marginal distributions and Poisson innovations is developed in Section 2.4.

With \( p = 1 \) and \( \alpha_1 = \alpha \), the above becomes \( D = \sigma_Y^2 / \mu_Y = (1 + \alpha)^{-1}[a_k \alpha + \sigma_{\epsilon}^2 / \mu_{\epsilon}] \). For a GINAR($p$) stationary time series without a self-generalized family \( \{K(\alpha)\} \), no general overdispersion property can be proved.
2.3.2 Analogues of Gaussian Autoregressive Moving Average

To define analogues of Gaussian moving-average (MA) and ARMA-like models, let \( \{ \epsilon_t : \ldots, -1, 0, 1, \ldots \} \) be a sequence of independent and identically distributed random variables with support on \( \mathbb{N}_0; \epsilon_t \) is an innovation random variable at time \( t \). In a general context, the extension of moving average of order \( q \) becomes \( q \)-dependent where observations more than \( q \) apart are independent.

The model, denoted as INMA(1), is

\[
Y_t = \epsilon_t + K(\alpha') \circ \epsilon_{t-1} = \epsilon_t + \sum_{i=1}^{\epsilon_{t-1}} K_{ji}(\alpha'),
\]  

with independent \( K_{ji}(\alpha') \) over \( t, i \), and the model denoted as INMA(\( q \)) is

\[
Y_t = \epsilon_t + \sum_{j=1}^{q-1} \sum_{i=1}^{\epsilon_{t-j}} K_{tji}(\alpha'_j),
\]

with independent \( K_{tji}(\alpha'_j) \) over \( t, j, i \). The model denoted as INARMA(1, \( q \)), with a construction analogous to the Poisson ARMA(1, \( q \)) in McKenzie (1986), is the following:

\[
Y_t = W_{t-q} + \sum_{j=0}^{q-1} K(\alpha'_j) \circ \epsilon_{t-j} = W_{t-q} + \sum_{j=0}^{q-1} \sum_{i=1}^{\epsilon_{t-j}} K_{tji}(\alpha'_j),
\]

\[
Y_t = \sum_{j=1}^{\epsilon_{t-j}} K_{s\ell}(\alpha) + \epsilon_s,
\]

with independent \( K_{tji}(\alpha'_j), K_{s\ell}(\alpha) \) over \( t, j, i, s, \ell \). If \( \alpha = 0 \), then \( W_s = \epsilon_s \) and \( Y_t = \epsilon_{t-q} + \sum_{j=0}^{q-1} K(\alpha'_j) \circ \epsilon_{t-j} \) is \( q \)-dependent (but not exactly the same as (2.8)).

2.3.3 Classes of Generalized Thinning Operators

In this subsection, some classes of generalized thinning operators and known results about the stationary distribution for GINAR series with \( p = 1 \) are summarized. The following definitions are needed:

Definition 2.1 (Generalized discrete self-decomposability and innovation).

(a) A nonnegative integer-valued random variable \( Y \) is generalized discrete self-decomposable (GDSD) with respect to \( (K(\alpha)) \) if and only if (iff)

\[
Y \overset{d}{=} K(\alpha) \circ Y + \epsilon(\alpha) \quad \text{for each } \alpha \in [0, 1].
\]

In this case, \( \epsilon(\alpha) \) has pgf \( G_Y(s)/G_K(G_K(s; \alpha)) \).
(b) Under expectation thinning compounding and a GDSD marginal (with pgf $G_Y(s)$), the stationary time series model is (2.2), where the innovation $\epsilon_t$ has pgf $G_Y(s)/G_Y(G_K(s; \alpha))$.

**Definition 2.2** (Self-generalized $\{K(\alpha)\}$). Consider a family of $K(\alpha) \sim F_K(\cdot; \alpha)$ with $E[K(\alpha)] = \alpha$ and pgf $G_K(s; \alpha) = E[s^{K(\alpha)}], \alpha \in [0, 1]$. Then $\{F_K(\cdot; \alpha)\}$ is self-generalized iff

$$G_K(G_K(s; \alpha); \alpha') = G_K(s; \alpha \alpha'), \quad \forall \alpha, \alpha' \in (0, 1).$$

For binomial thinning, the class of possible margins is called the discrete self-decomposable (DSD) class. Note that unless $Y$ is Poisson and $\{K(\alpha)\}$ corresponds to binomial thinning, the distribution of the innovation is in a different parametric family than $F_Y$.

The terminology of self-generalizability is used in Zhu and Joe (2010b), and the concept is called a semigroup operator in Van Harn and Steutel (1993). Zhu and Joe (2010a) show that (1) $\text{Var}[K(\alpha)] = \sigma_K^2(\alpha) = a_K \alpha(1 - \alpha)$, where $a_K \geq 1$ for a self-generalized family $\{K(\alpha)\}$ and (2) that generalized thinning operators without self-generalizability lack some closure properties. Also self-generalizability is a nice property for embedding into a continuous-time process.

For NB, Zhu and Joe (2010b) show that NB$(\theta, \xi)$ is GDSD for three self-generalizable thinning operators that are given below. For NB$(\theta, \xi)$, with parametrization as given in Section 2.2, the pgf is $G_{NB}(s; \theta, \xi) = (1 - (1 - \pi)s)\theta, \text{ for } s > 0, \theta > 0$ and $\xi > 0$.

Three types of thinning operators based on $\{K(\alpha)\}$ are given below in terms of the pgf, together with $\text{Var}[K(\alpha)]$; the second operator (I2) has been used by various authors in several different parametrizations; the specification is simplest via pgfs. The different $\{K(\alpha)\}$ families allow different degrees of conditional heteroscedasticity.

1. **(I1)** (binomial thinning) $G_K(s; \alpha) = (1 - \alpha) + \alpha s$, with $\text{Var}[K(\alpha)] = \alpha(1 - \alpha)$.
2. **(I2)** $G_K(s; \alpha; \gamma) = \frac{(1 - \alpha) + (\alpha - \gamma)\xi}{(1 - \alpha)\gamma - (1 - \alpha)\gamma s}, 0 \leq \gamma \leq 1$, with $\text{Var}[K(\alpha)] = \alpha(1 - \alpha)(1 + \gamma)/(1 - \gamma)$.
   Note that $\gamma = 0$ implies $G_K(s; \alpha) = (1 - \alpha) + \alpha s$.
3. **(I3)** $G_K(s; \alpha; \gamma) = \gamma^{-1}[1 + \gamma - (1 + \gamma - \gamma s)^{\alpha}], 0 \leq \gamma$, with $\text{Var}[K(\alpha)] = \alpha(1 - \alpha)(1 + \gamma)$.
   Note that $\gamma \rightarrow 0$ implies $G_K(s; \alpha) = (1 - \alpha) + \alpha s$.

For NB$(\theta, \xi)$, GDSD with respect to **I2**$(\gamma)$ holds for $0 \leq \gamma \leq 1 - \pi = \xi/(1 + \xi)$, and GDSD with respect to **I3**$(\gamma)$ holds for $0 \leq \gamma \leq (1 - \pi)/\pi = \xi$. For GP$(\theta, \eta)$, the property of GDSD is shown in Zhu and Joe (2003), and it can be shown that GP$(\theta, \eta)$ is GDSD with respect to **I2**$(\gamma(\eta))$, where $\gamma(\eta)$ increases as the overdispersion $\eta$ increases. Note that the GP distribution does not have a closed-form pgf.

### 2.3.4 Estimation

For parameter estimation in count time series models, a common estimation approach is CLS. This involves the minimization of $\sum_{i=2}^{n}(y_i - E[Y_i | y_{i-1}, y_{i-2}, \ldots])^2$ for a time series of length $n$. For a stationary model, it is straightforward to get point estimators of $\mu_Y$ and some autocorrelation parameters. One problem with conditional least squares (CLS) is that it cannot distinguish overdispersed Poisson models for $\epsilon_t$ and $Y_t$. For example, if a NB or GP time series is assumed with one of the above generalized thinning operators, then the overdispersion cannot be reliably estimated with an extra moment equation after CLS.
We next mention what can be done for computations of pmfs and the likelihood for binomial thinning and generalized thinning.

1. Zhu and Joe (2006) have an iterative method for computing pmfs with binomial thinning and a DSD stationary margin.

2. The pgf of $K \ast y$ has closed form if the pgf of $K(\alpha)$ has closed form and the pgf of the innovation has closed form if the pgf of $Y$ has closed form. In this case, Zhu and Joe (2010b) invert a characteristic function for the pgf of $G_{K(\alpha) \ast Y} G_{\epsilon(\alpha)}$ using an algorithm of Davies (1973) to compute the conditional pmf of $Y_t$ given $Y_{t-1} = Y_{t-1}$. Let $\varphi_W(s) = E(e^{sW}) = G_W(e^{sY})$ for a nonnegative integer random variable $W$ and define $a(w) := \frac{1}{2} - (2\pi)^{-1} \int_{-\pi}^{\pi} \Re \left( \frac{\varphi_W(u) e^{-iuw}}{1 - e^{-iw}} \right) du$. Then $\Pr(W < w) = a(w)$. The pmf of $W$ is

$$f_W(0) = \Pr(W < 1) = a(1), \quad f_W(w) = a(w + 1) - a(w), \quad w = 1, 2, \ldots.$$ 

This works for NB but not GP because the latter does not have a closed-form pgf.

2.3.5 Incorporation of Covariates

For a NB$(\theta, \xi)$ stationary INAR(1) model, the pdf of the innovation is

$$G_{NB}(s; \theta, \xi) = \frac{G_{NB}(s; \theta, \xi)}{G_{NB}(s; \alpha; \theta, \xi)}.$$

For a time-varying $\theta_t$ that depends on covariates with fixed $\xi$ (fixed overdispersion index), suppose the innovation $\epsilon_t$ has pgf

$$G_{NB}(s; \theta_t, \xi) = \frac{G_{NB}(s; \theta_t, \xi)}{G_{NB}(s; \alpha; \theta_t, \xi)}.$$ 

(2.10)

An advantage of this assumption is that a NB stationary margin results when $\theta_t$ is constant.

More generally, for GINAR($p$) series where the stationary distribution does not have a simple form, the simplest extension to accommodate covariates is to assume an overdispersed distribution for $\epsilon_t$ and absorb a function of covariates into the mean of $\epsilon_t$. Alternatively, other parameters can be made into functions of the covariates.

2.4 Operators in Convolution-Closed Class

The viewpoint in this section is to construct a stationary time series of order $p$ based on a joint pmf $f_{1 \ldots (p+1)}$ for $(Y_{t-p}, \ldots, Y_t)$, where marginal pmfs satisfy $f_{1 \ldots m} = f_{(i+1) ; (m+i)}$ for $i = 1, \ldots, p + 1 - m$ and $m = 2, \ldots, p$. Suppose the univariate marginal pmfs of $f_{1 \ldots (p+1)}$ are all $f_Y = f_{Y_t}$. From this, one has a transition probability $f_{p+1 \ldots p} = f_{Y_t | Y_{t-p}, \ldots, Y_{t-1}}$. For $p = 1$, $f_{2 | 1}$ leads to a stationary Markov time series of order 1. For $p = 2$, $f_{3 | 12}$ leads to a stationary Markov time series of order 2 if $f_{123}$ has bivariate marginal pmfs $f_{12} = f_{23}$. 

© 2016 by Taylor & Francis Group, LLC
There is some theory that covers several count time series models when \( f_Y \) is convolution-closed and infinitely divisible (CCID), because there is a way to construct a joint multivariate distribution based on these properties and they lead to thinning operators. This theory provides a bridge between the thinning operator approach of Section 2.3 and the general Markov approach with copulas in Section 2.5.

The operators have been studied in specific discrete cases by McKenzie (1985, 1986, 1988), Al-Osh and Alzaid (1987), and Alzaid and Al-Osh (1993), and in a more general framework in Joe (1996) and Jørgensen and Song (1998).

The general operator is presented first for the Markov order 1 case and then it is mentioned how it can be extended to higher-order Markov or \( q \)-dependent, etc. series. Also it will be mentioned how covariates can be accommodated. For this construction, Markov order 1 implies linear conditional expectation but not Markov orders of 2 or higher.

Let \( \{F(\cdot; \theta) : \theta > 0\} \) be a CCID parametric family such that \( F(\cdot; \theta_1) \ast F(\cdot; \theta_2) = F(\cdot; \theta_1 + \theta_2) \), where \( \ast \) is the convolution operator; \( F(\cdot; 0) \) corresponds to the degenerate distribution at 0. For \( X_j \sim F(\cdot; \theta_1) \), \( j = 1, 2 \), with \( X_1, X_2 \) independent, let \( H(\cdot; \theta_1, \theta_2, y) \) be the distribution of \( X_1 + X_2 = y \). Let \( R(\cdot) = R(\cdot; \alpha, \theta) \) \( (0 < \alpha \leq 1) \) be a random operator such that \( R(Y) \) given \( Y = y \) has distribution \( H(\cdot; \alpha \theta, (1 - \alpha) \theta, y) \), and \( R(Y) \sim F(\cdot; \alpha \theta) \) when \( Y \sim F(\cdot; \theta) \).

A stationary time series with margin \( F(\cdot; \theta) \) and autocorrelation \( 0 < \alpha < 1 \) (at lag 1) can be constructed as

\[
Y_t = R_t(Y_{t-1}) + \varepsilon_t, \quad R_t(y_{t-1}) \sim H(\cdot; \alpha \theta, (1 - \alpha) \theta, y_{t-1}),
\]

since \( F(\cdot; \theta) = F(\cdot; 0 \alpha) \ast F(\cdot; \theta(1 - \alpha)) \), when the innovations \( \varepsilon_t \) are independent and identically distributed with distribution \( F(\cdot; (1 - \alpha) \theta) \). Note that \( \{R_t : t \geq 1\} \) are independent replications of the operator \( R \).

The intuitive reasoning is as follows. A consecutive pair \( (Y_{t-1}, Y_t) \) has a common latent or unobserved component \( X_{12} \) through the stochastic representation:

\[
Y_{t-1} = X_{12} + X_1, \quad Y_t = X_{12} + X_2,
\]

where \( X_{12}, X_1, X_2 \) are independent random variables with distributions \( F(\cdot; \alpha \theta), F(\cdot; (1 - \alpha) \theta), F(\cdot; (1 - \alpha) \theta) \), respectively. The operator \( R_t(Y_{t-1}) \) “recovers” the unobserved common component \( X_{12} \); hence the distribution of \( R_t(y) \) given \( Y_{t-1} = y \) must be the same as the distribution of \( X_{12} \) given \( X_{12} + X_1 = y \).

Examples of CCID operations for the infinite divisible distributions of Poisson, NB and GP are given below.

1. If \( F(\cdot; \theta) \) is Po\((\theta)\), then \( H(\cdot; \alpha \theta, (1 - \alpha) \theta, y) \) is Bin\((y, \alpha)\). The resulting operator is binomial thinning.
2. If \( F(\cdot; \theta) = F_{NB}(\cdot; \theta, \xi) \) with fixed \( \xi > 0 \), then \( H(\cdot; \alpha \theta, (1 - \alpha) \theta, y) \) or \( Pr(X_1 = x \mid X_1 + X_2 = y) \) with \( X_j \) independently NB\((\theta_j, \xi)\), is Beta-binomial\((y, \alpha \theta, (1 - \alpha) \theta)\) independent of \( \xi \). The pmf of \( H \) is

\[
h(x; \theta_1, \theta_2, y) = \binom{y}{x} \frac{B(\theta_1 + x, \theta_2 + y - x)}{B(\theta_1, \theta_2)}, \quad x = 0, 1, \ldots, y,
\]
The operator matches the random coefficient thinning in Section 2.3, but not binomial thinning or generalized thinning. This first appeared in McKenzie (1986). For (2.11) based on this operator $E[Y_t|Y_{t-1} = y] = \alpha y + (1 - \alpha)\theta \xi$, and

$$
\text{Var}(Y_{t}|Y_{t-1} = y) = \frac{(1 - \alpha)\theta \xi(1 + \xi) + y(\theta + y)\alpha(1 - \alpha)}{(\theta + 1)}.
$$

The conditional variance is quadratically increasing in $y$ for large $y$, and hence this process has more conditional heteroscedasticity than those based on compounding operators in Section 2.3.

3. If $F(\cdot;\theta) = F_{GP}(\cdot;\theta,\eta)$ with $0 < \eta < 1$ fixed, then $H(\cdot;\alpha \theta, (1 - \alpha)\theta, y)$ or $\text{Pr}(X_1 = x | X_1 + X_2 = y)$ with $X_j$ independently $\text{GP}(\theta_j,\eta)$ is a quasi-binomial distribution with parameters $\pi = \theta_1/(\theta_1 + \theta_2)$, $\zeta = \eta/(\theta_1 + \theta_2)$. The quasi-binomial pmf is:

$$
h(x;\pi,\zeta, y) = \binom{y}{x} \frac{\pi(1 - \pi)}{1 + \zeta y} \left[ \frac{\pi + \zeta x}{1 + \zeta y} \right]^{x-1} \left[ \frac{1 - \pi + \zeta(y - x)}{1 + \zeta y} \right]^{y-x-1},
$$

for $x = 0, 1, \ldots, y$. For (2.11) with this operator, $E[Y_t|Y_{t-1} = y] = \alpha y + (1 - \alpha)\theta/(1 - \eta),$

$$
\text{Var}[Y_t|Y_{t-1} = y] = \alpha(1-\alpha) \left[ y^2 - \sum_{j=0}^{y-2} \frac{y!\zeta^j}{(y-j-2)!(1+y\zeta)^{j+1}} \right] + \frac{(1 - \alpha)\theta}{(1 - \eta)^2}, \quad \zeta = \frac{\eta}{\theta}.
$$

see Alzaid and Al-Osh (1993). Numerically this is superlinear and asymptotically $O(y^2)$ as $y \to \infty$.

These operators can be used for INMA($q$) and INARMA($1,q$) models in an analogous manner to the models in (2.8) and (2.9).

Next, we present the Markov order $2$ extension of Joe (1996) and Jung and Tremayne (2011). Consider the following model for three consecutive observations

$$
\begin{align*}
Y_{t-2} &= X_{123} + X_{12} + X_{13} + X_1 \\
Y_{t-1} &= X_{123} + X_{12} + X_{23} + X_2, \\
Y_t &= X_{123} + X_{23} + X_{13} + X_3,
\end{align*}
$$

(2.12)

where $X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X_{123}$ have distributions in the family $F(\cdot;\theta)$ with respective parameters $\theta_1' = \theta - \theta_0 - \theta_1 - \theta_2$, $\theta_2' = \theta - \theta_0 - 2\theta_1$, $\theta_1', \theta_1, \theta_2, \theta_1, \theta_0$ ($\theta$ is defined so that $\theta_1', \theta_2'$ are nonnegative). The conditional probability $\text{Pr}(Y_t = y_{\text{new}} | Y_{t-1} = y_{\text{prev1}}, Y_{t-2} = y_{\text{prev2}})$ does not lead to a simple operator for Markov order $1$, so that computationally one can just use

$$
\text{Pr}(Y_t = w_3 | Y_{t-1} = w_2, Y_{t-2} = w_1) = \frac{\text{Pr}(Y_{t-2} = w_1, Y_{t-1} = w_2, Y_t = w_3)}{\text{Pr}(Y_{t-2} = w_1, Y_{t-1} = w_2)}.
$$
The numerator involves a quadruple sum:

\[
\begin{align*}
\sum_{x_{123}=0} & \sum_{x_{12}=0} \sum_{x_{23}=0} \sum_{x_{13}=0} f(x_{123}, \theta_0) f(x_{12}; \theta_1) f(x_{23}; \theta_1) f(w_1 - x_{123} - x_{12} - x_{13}; \theta'_1) \\
& \cdot f(w_2 - x_{123} - x_{12} - x_{23}; \theta'_2) f(w_3 - x_{123} - x_{23} - x_{13}; \theta'_3).
\end{align*}
\]

For a model simplication, let \( \theta_{13} = 0 \) so that \( X_{13} = 0 \); then the numerator becomes a triple sum. Letting \( X_{13} = 0 \) is sufficient to get a one-parameter extension of Markov order 1; this Markov order 2 model becomes the Markov order 1 model when \( \theta_0 = \alpha^2 \theta, \theta_1 = \alpha(1 - \alpha) \theta, \theta'_1 = (1 - \alpha) \theta, \theta'_2 = (1 - \alpha^2) \theta \).

When \( \theta_{13} = 0 \), and \( \alpha_1 \geq \alpha_2 \geq 0 \) are the autocorrelations at lags 1 and 2, the time series model has a stochastic representation:

\[
Y_t = R_t(Y_{t-1}, Y_{t-2}) + \epsilon_t, \quad \epsilon_t \sim F(\cdot; (1 - \alpha_1) \theta),
\]

(2.13)

where via (2.12), \( R_t(y_{t-1}, y_{t-2}) \) has the the conditional distribution of \( X_{123} + X_{12} + X_{23} \) given \( Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2} \) and the convolution parameters of \( X_{123}, X_{12}, X_1, X_2 \) are, respectively, \( \alpha_2 \theta, (\alpha_1 - \alpha_2) \theta, (1 - \alpha_1) \theta, (1 - 2\alpha_1 + \alpha_2) \theta \) with \( \alpha_2 \geq 2\alpha_1 - 1 \). If \( \theta_{13} > 0 \), then the convolution parameters of \( X_{123}, X_{12}, X_{13}, X_1, X_2 \) are, respectively, \( \theta_0, \theta_1, \theta_2, \theta'_1, \theta'_2 \) with \( \alpha_1 = (\theta_1 + \theta_0)/\theta, \alpha_2 = (\theta_2 + \theta_0)/\theta \).

The pattern extends to higher-order Markov but numerically the transition probability becomes too cumbersome because the most general \( p \)-dimensional distribution of this type involves \( 2^p - 1 \) independent \( X_S \) for \( S \) being a nonempty subset of \( \{1, \ldots, p\} \). As mentioned by Jung and Tremayne (2011), the autocorrelation structure of this Markov model for \( p \geq 2 \) with Poisson, NB, or GP margins does not mimic the Gaussian counterpart, because of a nonlinear conditional mean function.

Because the distribution of the innovation is in the same family as the stationary marginal distribution, the models can be extended easily so that the convolution parameter of \( Y_t \) is \( \theta_t \), which depends on time-varying covariates. For example, for the Markov order 1 model, with \( Y_t \sim F(\cdot; \theta_t), R_t(Y_{t-1}) \sim F(\cdot; \alpha \theta_{t-1}) \) and \( \epsilon_t \sim F(\cdot; \zeta_t) \) with \( \zeta_t = \theta_t - \alpha \theta_{t-1} \geq 0 \) (Joe 1997, Section 8.4.4). For NB and GP, this means the univariate regression models are NB1 and GP1, respectively.

### 2.5 Copula-Based Transition

The copula modeling approach is a way to get a joint distribution for \( (Y_{t-p}, \ldots, Y_t) \) without an assumption of infinite divisibility. Hence univariate margins can be any distribution in the stationary case. However, the property of linear conditional expectation for the Markov order 1 process will be lost. For a \((p + 1)\)-variate copula \( C_{1:(p+1)} \), then \( F_{1:(p+1)} = C_{1:(p+1)}(F_Y, \ldots, F_Y) \) is a model for the multivariate discrete distribution of \((Y_{t-p}, \ldots, Y_t)\). For stationarity, marginal copulas satisfy \( C_{1:m} = C_{1+1:(m+i)} \) for \( i = 1, \ldots, p + 1 - m \) and \( m = 2, \ldots, p \). The resulting transition probability \( \Pr(Y_t = y_t \mid Y_{t-p} = y_{t-p}, \ldots, Y_{t-1} = y_{t-1}) \) can be computed from \( F_{1:(p+1)} \). If \( Y \) were a continuous random variable, there is a simple
stochastic representation for the copula-based Markov model in terms of $U(0, 1)$ random variables, but this is not the case for $Y$ discrete.

If there are time-varying covariates $z_i$ so that $F_Y = F(y; \beta, z_i)$, then one can use $F_{1:(p+1)} = C_{1:(p+1)}(F_{Y_1:p}, \ldots, F_{Y_i})$ for the distribution of $(Y_{1:p}, \ldots, Y_i)$ with Markov dependence and a time-varying parameter in the univariate margin.

For $q$-dependence, one can get a time series model $\{F_Y^{-1}(U_t)\}$ with stationary margin $F_Y$ if $\{U_t\}$ is a $q$-dependent sequence of $U(0, 1)$ random variables. For mixed Markov/$q$-dependent, a copula model that combines features of Markov and $q$-dependence can be defined. Chapter 8 of Joe (1997) has the copula time series models for Markov dependence and $1$-dependence.

More specific details of parametric models are given for Markov order 1, followed by brief mention of higher-order Markov, $q$-dependent and mixed Markov/$q$-dependent.

For a stationary time series model, with stationary univariate distribution $F_Y$, let $F_{12} = C(F_Y, F_Y; \delta)$ be the distribution of $(Y_{t-1}, Y_t)$ where $C$ is a bivariate copula family with dependence parameter $\delta$. Then the transition probability $\Pr(Y_t = y_t|Y_{t-1} = y_{t-1})$ is

$$f_{21}(y_t|y_{t-1}) = \frac{F_{12}(y_{t-1}, y_t) - F_{12}(y_{t-1}', y_t) - F_{12}(y_{t-1}, y_t') + F_{12}(y_{t-1}', y_t')}{f_Y(y_{t-1})},$$

where $y_i'$ is shorthand for $y_i - 1$ for $i = t - 1$ and $t$.

Below are a few examples of one-parameter copula models that include independence, perfect positive dependence, and possibly an extension to negative dependence. Different tail behavior of the copula leads to different asymptotic tail behavior of the conditional expectation and variance, but the conditional expectation is roughly linear in the middle. If a copula $C$ is the distribution of a bivariate uniform vector $(U_1, U_2)$, then the distribution of the reflection $(1 - U_1, 1 - U_2)$ is $\hat{C}(u_1, u_2) := u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$. The copula $C$ is reflection symmetric if $C = \hat{C}$. Otherwise for a reflection asymmetric bivariate copula $C$, one can also consider $\hat{C}$ as a model with the opposite direction of tail asymmetry.

The bivariate Gaussian copula can be considered as a baseline model from which other copula families deviate from in tail behavior. Based on Jeffreys’ and Kullback–Leibler divergences of $Y_1, Y_2$ that are NB or GP, the bivariate distribution $F_{12}$ from the binomial thinning operator or the beta/quasi-binomial operators are very similar, with typically a sample size of over 500 needed to distinguish the models when the (lag 1) correlation is moderate ($0.4–0.7$).

Below is a summary of bivariate copula families with different tail properties and hence different tail behavior of the conditional mean $E(Y_{t|Y_{t-1} = y})$ and variance $\text{Var}(Y_{t|Y_{t-1} = y})$ as $y \to \infty$, when $F_Y = F_{NB}$ or $F_{GP}$.

1. **Bivariate Gaussian**: reflection symmetric, with $\Phi, \Phi_2$ being the univariate and bivariate Gaussian cdf with mean 0 and variance 1, $C(u_1, u_2; \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho)$, $-1 < \rho < 1$. The conditional mean is asymptotically slightly sublinear and the conditional variance is asymptotically close to linear.

2. **Bivariate Frank**: reflection symmetric, $C(u_1, u_2; \delta) = -\delta^{-1} \log[1 - (1 - e^{-\delta u_1})(1 - e^{-\delta u_2})/(1 - e^{-\delta})]$, $-\infty < \delta < \infty$. Because the upper tail behaves like $1 - u_1 - u_2 + C(u_1, u_2) \sim \zeta(1 - u_1)(1 - u_2)$ for some $\zeta > 0$ as $u_1, u_2 \to 1^-$, the conditional mean and variance are asymptotically flat.
3. Bivariate Gumbel: reflection asymmetric with stronger dependence in the joint upper tail. $C(u_1, u_2; \delta) = \exp\{-(\log u_1)^{\delta} + (\log u_2)^{\delta}\}^{1/\delta}$ for $\delta \geq 1$. The conditional mean is asymptotically linear and the conditional variance is asymptotically sublinear.

4. Reflected or survival Gumbel: reflection asymmetric with stronger dependence in the joint lower tail. $C(u_1, u_2; \delta) = u_1 + u_2 - 1 + \exp\{-(\log(1 - u_1))^{\delta} + (\log(1 - u_2))^{\delta}\}^{1/\delta}$ for $\delta \geq 1$. The conditional mean and variances are asymptotically sublinear.

The Gumbel or reflected Gumbel copula can be recommended when there is some tail asymmetry relative to Gaussian. The Gumbel copula can be recommended when it is expected that there is some clustering of large values (exceeding a large threshold). The Frank copula is the simple copula that is reflection symmetric and can allow negative dependence. However its bivariate joint upper and lower tails are lighter than the Gaussian’s copula, and this has implication that the conditional expectation $E(Y_t|Y_{t-1} = y)$ converges to a constant for large $y$. For Gaussian, Gumbel, and reflected Gumbel, $E(Y_t|Y_{t-1} = y)$ is asymptotically linear or sublinear for large $y$ for $\{Y_t\}$ with a stationary distribution that is exponentially decreasing (like NB and GP). Some of these results can be proved with the techniques in Hua and Joe (2013).

For second order Markov chains, one just needs a trivariate copula that satisfies a good choice is the trivariate Gaussian copula with lag 1 and lag 2 latent correlations being $\rho_1, \rho_2$ respectively. If closed-form copula functions are desired, a class to consider has form

$$C_{\psi,H}(u_1, u_2, u_3) = \psi \left( \sum_{j \in \{1,3\}} \left[ -\log H(e^{-0.5\psi^{-1}(u_j)} e^{-0.5\psi^{-1}(u_2)}) + \frac{1}{2} \psi^{-1}(u_j) \right] \right),$$

where $\psi$ is the Laplace transform of a positive random variable, $H$ is a bivariate permutation symmetric max-infinite divisible copula; it has bivariate margins:

$$C_2 = \psi \left( -\log H(e^{-0.5\psi^{-1}(u_j)} e^{-0.5\psi^{-1}(u_2)}) + \frac{1}{2} \psi^{-1}(u_j) + \frac{1}{2} \psi^{-1}(u_2) \right), \quad j = 1, 3,$$

and $C_3(u_1, u_3) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_3))$. This $C_{\psi,H}$ is a suitable copula, with closed-form cdf, for the transition of a stationary time series of Markov order 2, when there is more dependence for measurements at nearer time points. If a model with clustering of large values is desired, then one can take $H$ to be the bivariate Gumbel copula and $\psi$ to be the positive stable Laplace transform, and then $C_{\psi,H}$ is a trivariate extreme value copula. Other simple choices used for the data set in Section 2.6 are the Frank copula for $H$ together with the positive stable or logarithmic series Laplace transform for $\psi$. Both the Gaussian copula and $C_{\psi,H}$ can be extended to AR($p$). Other alternatives for copulas for Markov order $p \geq 2$ are based on discrete D-vines (see Panagiotelis et al. 2012). For copula versions of Gaussian MA($q$) series, analogues of (2.8) can be constructed for dependent $U(0,1)$ sequences which can then be converted with the inverse probability transform $F_Y^{-1}$. For a $q$-dependent sequence, a $(q + 1)$-variate copula $K_{1(q+1)}$ is needed with margin $K_{1q}$ being an independence copula. Then $K_{1(q+1)}$ is the copula of $(\epsilon_{t-q}, \ldots, \epsilon_{t-1}, U_t)$, where $(\epsilon_t)$ is a sequence of independent $U(0,1)$ innovation random variables, and $U_t = K_{1(q+1)}^{-1}(\epsilon_t | \epsilon_{t-q}, \ldots, \epsilon_{t-1})$. Here $K_{q+1}|_{1q}$ is the conditional distribution of variable $q + 1$ given variables $1, \ldots, q$. 

© 2016 by Taylor & Francis Group, LLC
For count time series, $q$-dependence for some positive integer $q$ is generally not expected based on context, as there may be no reason for independence to occur with longer lags. Mixed Markov/$q$-dependent models with small $p,q$ are better models to consider if the dependence is not as simple as Markov dependence.

The copula version with $p = q = 1$ is defined below (it does extend to $p,q \geq 1$ by combining the above constructions for Markov order $p$ and $q$-dependence). Let $C_{12}$ and $K_{12}$ be two different bivariate copulas. Define

$$W_s = C_{2|1}^{-1}(\epsilon_s|W_{s-1}), \quad U_t = K_{2|1}^{-1}(\epsilon_t|W_{t-1}), \quad Y_t = F_Y^{-1}(U_t),$$

where $\{\epsilon_t\}$ is a sequence of innovation random variables and $\{W_s\}$ is a sequence of unobserved $U(0,1)$ random variables.

### 2.6 Statistical Inference and Model Comparisons

For NB margins, numerical maximum likelihood is possible for all of the Markov models in the preceding three sections. For GP margins, all are possible except the ones based on the generalized thinning operators denoted as $I_2$ and $I_3$. With numerical maximum likelihood for Markov models, the asymptotic likelihood inference theory is similar to that for independent observations, using the results in Billingsley (1961). The CLS method can estimate the mean of a stationary distribution but not additional parameters such as overdispersion.

For applications to count time series with small counts, generally low-order Markov models are adequate. The $q$-dependent models which are the analogies of Gaussian MA($q$) are usually less appealing within the context of count data. The analogue of Gaussian ARMA(1,1) has been used in the literature on count time series, but there is less experience with such models. It would be desirable to have the ARMA(1,1) equivalent of the above models in the three categories. The joint likelihood of $Y_1, \ldots, Y_n$ (time series of length $n$) involves either a high-dimensional sum or integral so that maximum likelihood estimation is not feasible. However for the ARMA(1,1) analogue in the CCID class or copula approach, the joint pmf of three consecutive observations is, respectively, a triple sum, and a triple sum plus an integral. Therefore, likelihood inference could proceed with composite likelihood based on sum of log-likelihoods of subsets of three consecutive observations, as in Davis and Yau (2011) and Ng et al. (2011). Because of space constraints, the data analysis below will only involve Markov models.

In the remainder of this section, the three classes of Markov count time series models are fitted and discussed for one data set, and some results are briefly summarized for another data set that had previously been analyzed.

For some count time data sets with low counts, serial dependence and explanation of trends from covariates, we consider monthly downloads to specialized software. A source for such series is http://econpapers.repec.org/software/. Such data are considered here because the next month’s total download is plausibly based on the current month’s total download through a generalized thinning operator. A similar data set with daily downloads for the program TeXpert is used in Weiß (2008).

The specific series consist of the monthly downloads of Signalpak (a package for octave, which is a Matlab clone) from September 2000 to March 2013. See Figure 2.1.
This series looks close to stationary. For longer periods of time, there is no reason to expect
the series to be stationary because software can go through periods or local trends of more
or less demand. We also consider a covariate for forecasting that is a surrogate for the
popularity of the octave software; the surrogate is the total monthly abstract views of
octave codes that are at the website referred to earlier. A summary of model fits is given in
Table 2.1.

The use of the covariate marginally improves the log-likelihood and root mean square
prediction error. Otherwise the Markov order 1 models fit about the same, and Markov
2 models did not add explanatory power. As might be expected, the generalized thinning
operators fit a little better than binomial thinning since those operators are more intuitive
for the context of these data. For thinning operators I_2 and I_3, the γ parameter was set
to be the largest possible so that they lead to more conditional heteroscedasticity than the
model with binomial thinning. Experience with I_2 and I_3 is that for short time series, if γ
is estimated, it is usually at one of the boundaries.

Table 2.1 lists maximum likelihood estimates and corresponding standard errors for one
of the better fitting models without/with the covariate. The estimates of the univariate
parameters are similar for the different models as well as the strength of serial dependence
at lag 1. Note that the addition of the covariate decreases the estimation of overdispersion
and lag 1 serial dependence.

For another count time series data set, we also comment on comparisons of fits of
models for the data set in Zhu and Joe (2006) on monthly number of claims of short-
term disability benefits, made by workers in the logging industry with cut injuries, to
the B.C. Workers’ Compensation Board for the period of 10 years from 1985 to 1994.
This data set of claims is one of few data sets in the literature where “survivor” inter-
pretation of binomial thinning is plausible. There is clear seasonality from the time series
plot and autocorrelation function plot, so we use the covariates (sin(2πt/12), cos(2πt/12)).
This led to root mean square prediction errors (RMSPE) that were from 2.78 to 2.86 with
no seasonal covariates and 2.64 to 2.73 with seasonal covariates. The Markov model with
a Gaussian copula fitted a little better than those based on thinning operators in terms of
### TABLE 2.1

The Covariate Is One Hundredth of the Number of Downloads of Octave Programs at the Website in the Preceding Month

<table>
<thead>
<tr>
<th>regr.</th>
<th>Dependence</th>
<th>Order</th>
<th>−Loglik</th>
<th>rmspe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>Gaussian copula</td>
<td>1</td>
<td>424.49</td>
<td>4.17</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian copula</td>
<td>1</td>
<td>424.36</td>
<td>4.17</td>
</tr>
<tr>
<td>NB</td>
<td>Gaussian copula</td>
<td>2</td>
<td>422.72</td>
<td>4.13</td>
</tr>
<tr>
<td>NB</td>
<td>Frank copula</td>
<td>1</td>
<td>428.65</td>
<td>4.24</td>
</tr>
<tr>
<td>NB</td>
<td>Gumbel copula</td>
<td>1</td>
<td>427.35</td>
<td>4.24</td>
</tr>
<tr>
<td>NB</td>
<td>refl.Gumbel copula</td>
<td>1</td>
<td>424.07</td>
<td>4.15</td>
</tr>
<tr>
<td>NB</td>
<td>conv.-closed oper.</td>
<td>1</td>
<td>425.24</td>
<td>4.18</td>
</tr>
<tr>
<td>NB</td>
<td>binomial thinning</td>
<td>1</td>
<td>426.49</td>
<td>4.18</td>
</tr>
<tr>
<td>NB</td>
<td><strong>I2</strong> thinning</td>
<td>1</td>
<td>425.36</td>
<td>4.18</td>
</tr>
<tr>
<td>NB</td>
<td><strong>I3</strong> thinning</td>
<td>1</td>
<td>425.83</td>
<td>4.18</td>
</tr>
<tr>
<td>One covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB1</td>
<td>Gaussian copula</td>
<td>1</td>
<td>416.07</td>
<td>4.02</td>
</tr>
<tr>
<td>NB2</td>
<td>Gaussian copula</td>
<td>1</td>
<td>416.03</td>
<td>4.02</td>
</tr>
<tr>
<td>GP1</td>
<td>Gaussian copula</td>
<td>1</td>
<td>415.99</td>
<td>4.02</td>
</tr>
<tr>
<td>NB1</td>
<td>Gaussian copula</td>
<td>2</td>
<td>415.79</td>
<td>4.01</td>
</tr>
<tr>
<td>NB2</td>
<td>Gaussian copula</td>
<td>2</td>
<td>415.76</td>
<td>4.02</td>
</tr>
<tr>
<td>NB1</td>
<td>Frank copula</td>
<td>1</td>
<td>416.98</td>
<td>4.05</td>
</tr>
<tr>
<td>NB2</td>
<td>Frank copula</td>
<td>1</td>
<td>416.90</td>
<td>4.05</td>
</tr>
<tr>
<td>NB1</td>
<td>Gumbel copula</td>
<td>1</td>
<td>416.00</td>
<td>4.01</td>
</tr>
<tr>
<td>NB2</td>
<td>Gumbel copula</td>
<td>1</td>
<td>416.25</td>
<td>4.01</td>
</tr>
<tr>
<td>NB1</td>
<td>refl.Gumbel copula</td>
<td>1</td>
<td>417.14</td>
<td>4.02</td>
</tr>
<tr>
<td>NB2</td>
<td>refl.Gumbel copula</td>
<td>1</td>
<td>417.04</td>
<td>4.02</td>
</tr>
<tr>
<td>NB1</td>
<td>conv.-closed oper.</td>
<td>1</td>
<td>415.99</td>
<td>4.02</td>
</tr>
<tr>
<td>NB</td>
<td>binomial thinning</td>
<td>1</td>
<td>416.10</td>
<td>4.02</td>
</tr>
<tr>
<td>NB</td>
<td><strong>I2</strong> thinning</td>
<td>1</td>
<td>415.63</td>
<td>4.02</td>
</tr>
<tr>
<td>NB</td>
<td><strong>I3</strong> thinning</td>
<td>1</td>
<td>415.86</td>
<td>4.02</td>
</tr>
<tr>
<td></td>
<td>Model (Markov order 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB with Gaussian copula and no covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_0 = 2.40 )</td>
<td></td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\xi} = 0.89 )</td>
<td></td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\rho} = 0.42 )</td>
<td></td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>NB1 with Gaussian copula</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_0 = 1.93 )</td>
<td></td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\beta}_1 = 0.09 )</td>
<td></td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\xi} = 0.48 )</td>
<td></td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \hat{\rho} = 0.22 )</td>
<td></td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** For the models with covariates based on the **I2** and **I3** thinning operators, the formulation of Section 2.3.5 is used. The root mean square prediction error (rmspe) is defined as \( \{(n - p)^{-1} \sum_{t=1+p}^{n} (y_t - \hat{E}(Y_t|Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}))^2 \}^{1/2} \), where \( n \) is the length of the series and \( p \) is the Markov order and \( \hat{E} \) means the conditional expectation with parameter equal to the maximum likelihood estimate (MLE) of the model. In the bottom, MLEs and corresponding standard errors are given for one of the better fitting models without and with the covariate.
log-likelihood, but the Frank copula model with NB2 regression margins was the best in terms of log-likelihood and RMSPE.

Similar to Table 2.1, there is a bit more variation among copula model fits than those based on thinning operators (the latter models tend to be close to the Gaussian copula model). An explanation is that different copula families have a wide variety of shapes of conditional expectation and variance functions.

Finally, we indicate an advantage of count time series models with known univariate marginal distributions. In this case, univariate and conditional probabilities can be easily obtained and also predictive intervals with and without the previous observation(s). A predictive interval without the previous time point is a regression inference and a predictive interval with the previous time point is an inference that combines regression and forecasting. The first type of predictive interval would not be easy to do with other classes of count time series models that are based on conditional specifications.

Acknowledgments
This work was supported by an NSERC Discovery grant.

References


