Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations

Paolo Gorgi

Vrije Universiteit Amsterdam, and Tinbergen Institute, Amsterdam, The Netherlands

[Received August 2019. Final revision July 2020]

Summary. The paper introduces a general class of heavy-tailed auto-regressions for modelling integer-valued time series with outliers. The specification proposed is based on a heavy-tailed mixture of negative binomial distributions that features an observation-driven dynamic equation for the conditional expectation. The existence of a stationary and ergodic solution for the class of auto-regressive processes is shown under general conditions. The estimation of the model can be easily performed by maximum likelihood given the closed form of the likelihood function. The strong consistency and the asymptotic normality of the estimator are formally derived. Two examples of specifications illustrate the flexibility of the approach and the relevance of the theoretical results. In particular, a linear dynamic equation and a score-driven equation for the conditional expectation are studied. The score-driven specification is shown to be particularly appealing as it delivers a robust filtering method that attenuates the effect of outliers. Empirical applications to the series of narcotics trafficking reports in Sydney and the euro–pound sterling exchange rate illustrate the effectiveness of the method in handling extreme observations.

Keywords: Heavy-tailed distributions; Integer-valued time series; Observation-driven models; Robust filtering

1. Introduction

Time series data with integer-valued observations are often encountered in empirical applications. Classical continuous response models, such as auto-regressive moving average models, are not suited for the modelling of such series. Over the last few decades, researchers have developed methods that can properly account for the discreteness of the data. A standard approach is to consider observation-driven models that feature time variation in the intensity parameter of the Poisson distribution (Fokianos et al., 2009; Ferland et al., 2006; Davis et al., 2003). A limitation of the Poisson distribution is that it imposes equidispersion, i.e. the mean equal to the variance. Equidispersion is typically restrictive in empirical applications and therefore overdispersed distributions, such as the negative binomial distribution, are often considered (Zhu, 2011; Davis and Wu, 2009; McKenzie, 1986). Another extension that has been considered in the literature is the use of zero-inflated distributions, which are suited for time series with large numbers of 0s (Zhu, 2012). We refer the reader to Davis et al. (2016) for an overview of recent developments.

Extreme observations, or outliers, are often present when analysing time series data. The study of time series with outliers has a long history that dates back to Fox (1972). Ignoring...
extreme observations in the data set leads to statistical models that offer a poor description of the series of interest. Additionally, statistical inference can also be problematic in the presence of outliers. Some recent studies have focused on the analysis of integer-valued time series with outliers. Barczy et al. (2010, 2012) and Chen et al. (2019) have discussed the estimation of integer-valued auto-regressive models and binomial auto-regressive models that are contaminated with innovational and additive outliers. Hall (2001, 2003) derived the extremal index of integer-valued moving averages. Scotto et al. (2018) studied the properties of a first-order max-integer-valued auto-regressive model. Zhu et al. (2015, 2016) considered influence analysis of outliers for integer-valued generalized auto-regressive conditional heteroscedastic (INGARCH) models. Silva et al. (2019) discussed outlier detection in a Bayesian framework.

A standard modelling approach for data sets with outliers is to describe extreme observations by means of heavy-tailed distributions. There is a vast literature on modelling continuous-valued time series with heavy-tailed distributions. Student’s t-distribution is often used for this and robust specifications for the dynamic component of the model are employed to attenuate the effect of outliers (Creal et al., 2011; Harvey and Luati, 2014). In contrast, the current literature on integer-valued time is less well developed. One of the only contributions is given by the first-order integer-valued auto-regressive model of Qian et al. (2020). In the current paper, we introduce a general class of observation-driven models for integer-valued time series data with extreme observations. The approach is based on a heavy-tailed mixture of negative binomial distributions, known as the beta–negative binomial (BNB) distribution. The class of models features a dynamic location parameter and a BNB conditional distribution, which describes extreme observations as tail events. We derive conditions for stationarity, ergodicity and finiteness of moments for the proposed class of stochastic processes. Additionally, we show that inference can be easily performed by maximum likelihood (ML), given that the likelihood function is available in closed form. The strong consistency and the asymptotic normality of the ML estimator are proved under general conditions. We consider and study two specifications of the dynamic component of the model. The first is a simple linear auto-regression for the conditional mean. Instead, the second is based on the generalized auto-regressive score (GAS) framework of Creal et al. (2013) and Harvey (2013). This second specification delivers a robust filter that attenuates the effect of extreme observations on the conditional expectation of the BNB process. Finally, we present two empirical applications to the time series of police reports on narcotics trafficking in Sydney, Australia, and to the number of tick changes of the euro–pound sterling exchange rate. The results illustrate the ability of the proposed approach in modelling time series data with extreme observations.

The paper is structured as follows. Section 2 provides a brief review of the BNB distribution. Section 3 introduces the class of BNB auto-regressive processes and discusses their stochastic properties. Section 4 derives the asymptotic properties of the ML estimator. Section 5 introduces the linear specification of the model and discusses its properties. Section 6 introduces the score-driven specification. Section 7 presents the empirical applications. Section 8 concludes.

2. Preliminaries

We start by reviewing some properties of the BNB distribution, which will be useful in the rest of the paper. The BNB distribution arises as a beta mixture of negative binomial distributions. In particular, let $Y$ conditional on $P$ have a negative binomial distribution, $Y|P \sim \text{NB}(r, P)$, with dispersion parameter $r$ and success probability $P$. Assume further that $P$ has a beta distribution, $P \sim \text{beta}(\alpha, \beta)$, with shape parameters $\alpha$ and $\beta$. Then, the marginal distribution of $Y$ is BNB with the following probability mass function (PMF):
\[
\mathbb{P}(Y = y) = \frac{\Gamma(y + r)}{\Gamma(y + 1)\Gamma(r)} \frac{B(\alpha + r, \beta + y)}{B(\alpha, \beta)}, \quad \text{for } y \in \mathbb{N},
\]
where \(\Gamma(\cdot)\) denotes the gamma function and \(B(\cdot, \cdot)\) the beta function. The parameter \(\alpha\) is the tail parameter of the BNB, which determines the heaviness of the right-hand tail. The smaller \(\alpha\) the heavier the tail is. Throughout the paper, we parameterize the BNB distribution in terms of its mean. More specifically, we consider \(\beta = (\alpha - 1)\lambda / r\). In this way, the parameter \(\lambda\) represents the mean of the BNB distribution when the mean is finite, which is the case when \(\alpha > 1\). We say that \(Y \sim \text{BNB}(\lambda, r, \alpha)\) has a BNB distribution with mean \(\lambda > 0\), dispersion parameter \(r > 0\) and tail parameter \(\alpha > 1\) if
\[
\mathbb{P}(Y = y) = \frac{\Gamma(y + r)}{\Gamma(y + 1)\Gamma(r)} \frac{B(\alpha + r, (\alpha - 1)\lambda / r + y)}{B(\alpha, (\alpha - 1)\lambda / r)}, \quad y \in \mathbb{N}. \tag{1}
\]
The BNB distribution enables us to account for extreme observations, which can be seen as tail events. Furthermore, we note that the BNB distribution can approximate arbitrarily well the negative binomial distribution as well as the Poisson distribution. As the tail parameter diverges, \(\alpha \to \infty\), the BNB distribution \(\text{BNB}(\lambda, r, \alpha)\) converges to a negative binomial distribution with dispersion parameter \(r\) and success probability \(\lambda / (r + \lambda)\). Further, as the dispersion parameter diverges, \(r \to \infty\), the BNB converges to a Poisson distribution with mean \(\lambda\). For a more detailed review of the BNB distribution, we refer the reader to Wang (2011).

3. Beta–negative binomial auto-regressive models

Consider a time series of counts \(\{y_t\}_{t \in \mathbb{Z}}\) with the conditional distribution
\[
y_t | \mathcal{F}_{t-1} \sim \text{BNB}(\lambda_t, r, \alpha), \quad t \in \mathbb{Z}, \tag{2}
\]
where \(r > 0\), \(\alpha > 1\) and \(\mathcal{F}_t\) denotes the \(\sigma\)-field that is generated by \(\{y_t, y_{t-1}, \ldots\}\). The conditional mean process \(\lambda_t = \mathbb{E}(y_t | \mathcal{F}_{t-1})\) is specified by the following stochastic recurrence equation (SRE)
\[
\lambda_{t+1} = g_\theta(y_t, \lambda_t), \quad t \in \mathbb{Z}, \tag{3}
\]
where \(g_\theta(\cdot, \cdot)\) is a parametric updating function that maps from \(\mathbb{N} \times \mathbb{R}^+\) into \(\mathbb{R}^+\), and \(\theta \in \mathbb{R}^n\) is a parameter vector. We denote by \(\kappa \in \mathbb{R}^{n+2}\) the entire parameter vector of the model \(\kappa = (\xi^T, \theta^T)^T\), where \(\xi = (r, \alpha)^T\). Note that in the above formulation, for simplicity of exposition, \(\lambda_t\) is assumed to be \(\mathcal{F}_{t-1}\) measurable. Below, we show that under a drift condition and a semicontraction condition on \(g_\theta\) the model’s equations admit a stationary and ergodic solution and \(\lambda_t\) is \(\mathcal{F}_{t-1}\) measurable.

The BNB auto-regressive model that is specified in equations (2) and (3) can describe extreme observations in time series data by means of the heavy-tailed BNB conditional PMF. A small value of the tail parameter \(\alpha\) indicates that extreme observations are more likely to occur. The \(m\)th conditional moment of the BNB auto-regressive process \(\mathbb{E}(y^m_t | \mathcal{F}_{t-1})\) is finite if and only if \(\alpha > m\). However, as we shall discuss below, finiteness of unconditional moments requires further conditions on the updating function \(g_\theta\). In Sections 5 and 6, some examples of specifications of the updating function are presented. We shall see that a robust updating function may be desirable to reduce the effect of outliers on the conditional mean \(\lambda_t\). We also note that the specification of the conditional mean \(\lambda_t\) in equation (3) can be extended to a higher order by including more lags of \(\lambda_t\) and \(y_t\). The conditional mean of the BNB auto-regression of order \(\forall (p, q)\) can be expressed as
\[
\lambda_{t+1} = g_\theta(y_t, \ldots, y_{t-p+1}, \lambda_t, \ldots, \lambda_{t-q-1}).
\]
However, throughout the paper, for simplicity of exposition and ease of notation, we shall focus on the first-order specification ($p = q = 1$) that is given in equation (3).

In the rest of the section, we study the stochastic properties of the BNB auto-regression that is described by equations (2) and (3). We derive conditions on the updating function $g_\theta$ that ensure that the process is strictly stationary and ergodic and that guarantee the existence of the first two unconditional moments of $y_t$. The first result that we obtain is the stationarity and ergodicity of the process and existence of the first moment under a drift condition and a semicontraction condition. The proof of theorem 1 is based on the approach of Doukhan and Neumann (2019) for semicontractive GARCH-type processes.

**Theorem 1** (stationarity and ergodicity). Consider the BNB auto-regressive process, given by equations (2) and (3), and assume that

(a) the drift condition

$$g_\theta(y, \lambda) \leq c_0 + c_1 y + c_2 \lambda$$

holds where $c_0$, $c_1$ and $c_2$ are positive constants such that $c_1 + c_2 < 1$, and

(b) the semicontraction condition

$$|g_\theta(y, \lambda) - g_\theta(y, \lambda^*)| \leq b|\lambda - \lambda^*|$$

holds for any $y \in \mathbb{N}$, where $b$ is a positive constant such that $b < 1$.

Then, $\{ (y_t, \lambda_t) \}_{t \in \mathbb{Z}}$ admits a stationary and ergodic solution with a finite first moment $\mathbb{E}(y_t) < \infty$. Furthermore, $\lambda_t$ is $\mathcal{F}_{t-1}$ measurable.

The proof of theorem 1 is given in Appendix A.1. We highlight that the conditions of theorem 1 are weaker than the contraction condition that is typically considered in the literature of integer-valued processes, which is given by

$$|g_\theta(y, \lambda) - g_\theta(y, \lambda^*)| \leq c_1 |y - y^*| + c_2 |\lambda - \lambda^*|,$$

with $c_1 + c_2 < 1$ (see, for instance, Davis and Liu (2016) for an application to exponential families). More specifically, the above contraction condition immediately implies that conditions (4) and (5) are satisfied.

The drift condition (4) in theorem 1 requires that the sum of the coefficients $c_1$ and $c_2$ is smaller than 1 to ensure stationarity, ergodicity and finiteness of the first unconditional moment of $y_t$. However, a stricter condition is needed to obtain the finiteness of the second moment. The next result imposes sufficient conditions on $c_1$ and $c_2$ to obtain $\mathbb{E}(y_t^2) < \infty$, and hence the weak stationarity of the BNB auto-regressive process.

**Theorem 2** (weak stationarity). Let the conditions of theorem 1 hold. Furthermore, assume that $\alpha > 2$ and the drift condition (4) holds with

$$\frac{(r - 1)(\alpha - 1)}{r(\alpha - 2)} c_1^2 + c_2^2 + 2c_1 c_2 < 1.$$

Then, $y_t$ has a finite second moment $\mathbb{E}(y_t^2) < \infty$. Hence $\{y_t\}_{t \in \mathbb{Z}}$ is weakly stationary.

The proof of theorem 2 is given in Appendix A.1. We note that, besides a stricter drift condition, theorem 2 also requires that the tail parameter $\alpha$ is greater than 2. This is needed because the conditional second moment of $y_t$ is finite if and only if $\alpha > 2$.

The results in theorems 1 and 2 are of key importance to derive the asymptotic properties of the ML estimator. In particular, as we shall see in Section 4, the strict stationarity conditions
in theorem 1 are sufficient for the consistency of the ML estimator. However, the additional conditions in theorem 2 are needed for the asymptotic normality to hold. In Sections 5 and 6, theorems 1 and 2 will be employed to establish the stochastic properties of a linear and a score-driven BNB auto-regression.

4. Maximum likelihood estimation

In this section, we discuss the estimation of the BNB auto-regression by ML. We derive conditions to ensure consistency and asymptotic normality of the ML estimator. We assume that a subset of a realized path from the BNB auto-regressive process in equations (2) and (3) with true parameter value $\kappa = \kappa_0$ is observed $\{y_t\}_{t=1}^T$. Here $T$ denotes the sample size. The likelihood function is available in closed form through a prediction error decomposition. In particular, the average log-likelihood function is

$$
\hat{L}_T(\kappa) = \frac{1}{T} \sum_{t=1}^T \hat{\ell}_t(\kappa) = \frac{1}{T} \sum_{t=1}^T \log \left\{ p(y_t|\hat{\lambda}_t(\theta), r, \alpha) \right\},
$$

where $\hat{\lambda}_t$ denotes the conditional PMF, given by

$$
p\{y_t|\hat{\lambda}_t(\theta), r, \alpha\} = \frac{\Gamma(y_t + r)}{\Gamma(y_t + 1)\Gamma(r)} \frac{B\{\alpha + r, (\alpha - 1)\hat{\lambda}_t(\theta)/r + y_t\}}{B\{\alpha, (\alpha - 1)\hat{\lambda}_t(\theta)/r\}}.
$$

The filtered time-varying parameter $\hat{\lambda}_t(\theta)$ is obtained recursively by using the observed data $\{y_t\}_{t=1}^T$:

$$
\hat{\lambda}_{t+1}(\theta) = g_\theta\{y_t, \hat{\lambda}_t(\theta)\},
$$

where the recursion is initialized at a fixed point $\hat{\lambda}_1(\theta) \in \mathbb{R}^+$. We note that initializing the recursion in expression (6) is needed since the observed data start from time $t = 1$. This is quite standard in the literature of observation-driven models. Finally, the ML estimator $\hat{\kappa}_T$ is defined as the maximizer of the likelihood function

$$
\hat{\kappa}_T = \arg \sup_{\kappa \in K} \hat{L}_T(\kappa),
$$

where $K = \Xi \times \Theta$ with $\Xi \subset (0, \infty) \times (2, \infty)$ and $\Theta \subset \mathbb{R}^n$ being compact parameter sets.

4.1. Consistency

To establish the consistency of the ML estimator, we first derive the stochastic limit properties of the filtered parameter $\hat{\lambda}_t$ defined in expression (6). Note that $\hat{\lambda}_t(\cdot)$ is a stochastic function that maps from $\Theta$ into $\mathbb{R}^+$. The stability of the filtered parameter $\hat{\lambda}_t$ is often referred to in the literature as invertibility (Straumann and Mikosch, 2006; Blasques et al., 2018). Because of the initialization, $\hat{\lambda}_t$ evaluated at the true parameter value, $\hat{\lambda}_t(\theta_0)$, does not correspond to the true conditional mean $\lambda_t$. In what follows, we show that $\{\hat{\lambda}_t\}_{t \in \mathbb{N}}$ converges exponentially almost surely and uniformly in $\Theta$ to a stationary and ergodic sequence of functions $\{\lambda_t\}_{t \in \mathbb{N}}$ such that $\hat{\lambda}_t(\theta_0) = \lambda_t$ with probability 1. (A sequence of random variables $\{\tilde{x}_t\}_{t \in \mathbb{N}}$ converges exponentially almost surely to another sequence $\{\tilde{y}_t\}_{t \in \mathbb{N}}$ if there is a constant $c > 1$ such that $c^t|\tilde{x}_t - \tilde{y}_t| \to 0$ almost surely as $t \to \infty$.) We start by imposing a continuity condition on the updating function $g_\theta$, which ensures that $\theta \mapsto \hat{\lambda}_t(\theta)$ is continuous in the compact set $\Theta$ with probability 1.

Assumption 1. The function $(\theta, \lambda) \mapsto g_\theta(y, \lambda)$ is continuous in $\Theta \times \mathbb{R}^+$ for any $y \in \mathbb{N}$ and the set $\Theta$ is compact.
Next, we assume that the drift condition and the semicontraction condition hold for all $\theta$ in the parameter set $\Theta$, which contains the true parameter value $\theta_0$. We note that this assumption is not restrictive since, in general, $\Theta$ can be defined as a compact ball around the true parameter vector $\theta_0$.

**Assumption 2.** The drift condition (4) and the semicontraction condition (5) are satisfied for any $\theta \in \Theta$. Furthermore $\theta_0 \in \Theta$.

The next result ensures the uniform convergence over $\Theta$ of the filtered parameter $\hat{\lambda}_t(\theta)$ to a stationary and ergodic limit $\tilde{\lambda}_t(\theta)$. Here $\| \cdot \|_\Theta$ denotes the supremum norm. Given a function $f: \Theta \mapsto \mathbb{R}$, the supremum norm is $\| f \|_\Theta = \sup_{\theta \in \Theta} | f(\theta) |$.

**Proposition 1 (invertibility).** Let assumptions 1 and 2 hold. Then, the filter $\{\hat{\lambda}_t(\theta)\}_{t \in \mathbb{N}}$ converges exponentially almost surely and uniformly over $\Theta$ to a unique stationary and ergodic sequence $\{\tilde{\lambda}_t(\theta)\}_{t \in \mathbb{Z}}$, $\| \hat{\lambda}_t - \tilde{\lambda}_t \|_\Theta \to 0$, exponentially almost surely, $t \to \infty$.

for any initialization $\hat{\lambda}_1(\theta) \in \mathbb{R}^+$. Furthermore, $\tilde{\lambda}_t$ is $\mathcal{F}_{t-1}$ measurable with a bounded moment uniformly over $\Theta$, $\mathbb{E}\| \tilde{\lambda}_t \|_\Theta < \infty$, and such that $\tilde{\lambda}_t(\theta_0) = \lambda_t$.

Proposition 1 plays a crucial role to ensure that the likelihood function $\hat{l}_t(\kappa)$, which depends on the approximate filter $\hat{\lambda}_t$, converges to a stationary and ergodic limit $l_t(\kappa) = \log p\{y_t | \hat{\lambda}_t(\theta), r, \alpha\}$, which depends on limit filter $\tilde{\lambda}_t$. In this way, $l_t(\kappa_0)$ corresponds to the true conditional log-PMF of the BNB auto-regressive model.

Next, building on the invertibility result, we impose some additional conditions to obtain the strong consistency of the ML estimator. The next assumption imposes a lower bound on the updating function.

**Assumption 3 (lower bound).** There is a constant $\tilde{c} > 0$ such that $g_\theta(y_t, \lambda) > \tilde{c}$ for any $(\theta, y, \lambda) \in \Theta \times \mathbb{N} \times \mathbb{R}^+$. Finally, we consider an identification condition on the parametric updating function $g_\theta$ to ensure that any parameter value $\theta \in \Theta$, $\theta \neq \theta_0$, generates a path $\hat{\lambda}_t(\theta)$ that is observationally different from the true $\lambda_t$. Furthermore, we impose that $\lambda_t$ is not a degenerate random variable to guarantee the identification of the parameter $r$ of the BNB PMF. In particular, the BNB PMF $y \mapsto p(y | \lambda_t, r, \alpha)$ is equivalent for $r = \bar{r}$ and $r = (\alpha - 1)\lambda_t / \bar{r}, \bar{r} > 0$. Therefore, if $\lambda_t$ is a constant, the parameter $r$ is not identified. We refer to the proof of lemma 8 in the on-line supplement for a more detailed discussion. We also note that the assumption that $\lambda_t$ is not constant is typically implied by the identification condition on $g_\theta$; see, for example, the specifications of $\lambda_t$ in Sections 5 and 6.

**Assumption 4 (identifiability).** For any $\theta \in \Theta$, $g_\theta(y_t, \lambda_t) = g_{\theta_0}(y_t, \lambda_t)$ almost surely if and only if $\theta = \theta_0$. Furthermore, $\lambda_t$ is non-degenerate.

Under the assumptions that were stated above, we obtain the strong consistency of the ML estimator.

**Theorem 3 (consistency).** Let assumptions 1–4 hold and let $\kappa_0 \in K$; then the ML estimator that is defined in equation (7) is strongly consistent, i.e.

$$\hat{\kappa}_T \to \kappa_0, \quad \text{almost surely, as } T \to \infty.$$  

The proof is given in Appendix A.1.
4.2. Asymptotic normality

We now focus on deriving the asymptotic normality of the ML estimator. First, we require that the data-generating process has a finite second moment. A finite second moment is needed to ensure the existence of some moments for the derivatives of the log-likelihood, which are used to apply a central limit theorem to the score of the log-likelihood function.

**Assumption 5** (weak stationarity). The assumptions of theorem 2 are satisfied for $\lambda = \lambda_0$.

The next assumption is needed to ensure that the Fisher information matrix is positive definite.

**Assumption 6** (positive definite Fisher information). The random variables of the vector $\partial g_\theta (y_t, \lambda_t) / \partial \theta |_{\theta = \theta_0}$ are linearly independent.

Finally, the next assumption requires some regularity conditions on the updating function $g_\theta$. In particular, we impose that the updating function is twice continuously differentiable and has some of its derivatives bounded by some linear combinations of their arguments. Note that $\| \cdot \|$ denotes the $L_1$-norm when applied to a vector and the operator norm induced by the $L_1$-norm when applied to a matrix. Furthermore, we consider the following shorthand notation for the derivatives of the updating function: $g_\theta^d (y, \lambda) = \partial^d g_\theta (y, \lambda) / \partial d$. If $d$ is a two-dimensional vector, then $g_\theta^d (y, \lambda)$ denotes the second-order partial derivative with respect to the elements of the vector. For instance, if $d = (\theta, \theta)$, then $g_\theta^{\theta \theta} (y, \lambda) = \partial^2 g_\theta (y, \lambda) / \partial \theta \partial \theta^T$.

**Assumption 7.** The function $\theta, \lambda \mapsto g_\theta (y, \lambda)$ is twice continuously differentiable with respect to both $\theta$ and $\lambda$. Furthermore, for any $\theta \in \Theta$ the following inequalities hold

$$
\| g_\theta^d (y, \lambda) \| \leq b_0 + b_1 y + b_2 \lambda
$$

and

$$
\| g_\theta^d (y, \lambda) - g_\theta^d (y, \lambda^*) \| \leq b_3 | \lambda - \lambda^* |,
$$

for $d \in \{ \lambda, \theta, (\lambda, \lambda), (\theta, \lambda), (\theta, \theta) \}$ and some positive constants $b_0, b_1, b_2$ and $b_3$.

The next result delivers the asymptotic normality of the ML estimator. The proof is given in Appendix A.1.

**Theorem 4** (asymptotic normality). Let assumptions 1–7 hold and $\lambda_0 \in \text{int}(K)$; then the ML estimator defined in equation (7) has the asymptotic distribution

$$
\sqrt{T} (\hat{\lambda}_T - \lambda_0) \xrightarrow{d} N(0, F_0^{-1}), \quad \text{as } T \to \infty,
$$

where $F_0 = -\mathbb{E} \{ \partial^2 l_t (\lambda_0) / \partial \lambda \partial \lambda^T \}$.

5. Beta–negative binomial–integer-valued generalized auto-regressive conditional heteroscedastic model

5.1. The model

An intuitive and simple way to specify the conditional mean $\lambda_t$ is to consider a linear auto-regressive process driven by past observations. Count processes with a linear specification of the conditional mean are often referred to in the literature as INGARCH models (Ferland et al., 2006). We define the BNB–INGARCH model through the equations

$$
\begin{align*}
  y_t | \mathcal{F}_{t-1} & \sim \text{BNB}(\lambda_t, r, \alpha), \\
  \lambda_{t+1} & = \omega + \phi \lambda_t + \tau y_t,
\end{align*}
$$

where $\omega > 0, \phi \geq 0$ and $\tau > 0$ are static parameters to be estimated. These parameters are restricted to be positive to guarantee that $\lambda_t$ is strictly positive with probability 1. A linear auto-regressive
specification for the conditional expectation $\lambda_t$ is often considered for Poisson and negative binomial auto-regressions; see Ferland et al. (2006), Fokianos et al. (2009) and Zhu (2011). We refer to these models as Po–INGARCH and NB–INGARCH respectively. We can immediately see that BNB–INGARCH can approximate arbitrarily well both Po–INGARCH and NB–INGARCH since the BNB distribution converges to the negative binomial distribution as $\alpha \to \infty$ and to the Poisson distribution as, additionally, $r \to \infty$.

The next result relies on theorems 1 and 2 to derive conditions for strict and weak stationarity of the BNB–INGARCH process (8).

**Theorem 5.** Let the BNB–INGARCH process (8) satisfy $\tau + \phi < 1$. Then, the process admits a strictly stationary and ergodic solution with a finite first moment $\mathbb{E}(y_t) < \infty$. Additionally, let $\alpha > 2$ and

$$\frac{(r + 1)(\alpha - 1)}{r(\alpha - 2)} \tau^2 + \phi^2 + 2\tau \phi < 1. \quad (9)$$

Then, the stationary solution has a finite second moment $\mathbb{E}(y_t^2) < \infty$. Hence, it is weakly stationary.

The proof is given in the on-line supplementary material. Finally, we derive the strong consistency and asymptotic normality of the ML estimator of the BNB–INGARCH model by appealing to theorems 3 and 4.

**Theorem 6.** Let the observed series $\{y_t\}_{t=1}^T$ be generated by the BNB–INGARCH process (8) with parameter value $\kappa_0 = (\tau_0, \alpha_0, \omega_0, \phi_0, \tau_0)^{T}$. Furthermore, let $\kappa_0 \in K$, where $K$ is a compact parameter set such that $\phi + \tau < 1$, $\omega > 0$, $\phi \geq 0$ and $\tau > 0$ for any $\kappa \in K$. Then, the ML estimator (7) is strongly consistent:

$$\hat{\kappa}_T \to \kappa_0, \quad \text{almost surely, as } T \to \infty.$$  

Assume further that $\kappa_0$ satisfies condition (9), $\alpha_0 > 2$, and $\kappa_0 \in \text{int}(K)$. Then, the ML estimator is asymptotically normally distributed:

$$\sqrt{T}(\hat{\kappa}_T - \kappa_0) \overset{d}{\to} N(0, F_0^{-1}), \quad \text{as } T \to \infty,$$

where $F_0 = -\mathbb{E}\{\partial^2 l_i(\kappa_0)/\partial \kappa \partial \kappa^T\}$.

The proof of theorem 6, which is given in the on-line supplement, is obtained by checking that assumptions 1–7 are satisfied. The small sample properties of the ML estimator are studied in a Monte Carlo experiment in the next section.

We also note that a BNB–INGARCH process of order $(p, q)$ can be obtained by adding lags of $\lambda_t$ and $y_t$ in the specification of the conditional mean:

$$\lambda_{t+1} = \omega + \sum_{i=1}^{q} \phi_i \lambda_{t+1-i} + \sum_{i=1}^{p} \tau_i y_{t+1-i},$$

where $\omega > 0$, $\phi_i \geq 0$, for $i = 1, \ldots, q$, and $\tau_i > 0$, for $i = 1, \ldots, p$. Furthermore, the linear specification of $\lambda_t$ that is discussed in this section can be extended to include some form of non-linearity. Following Wang et al. (2014), we can consider a self-excited threshold auto-regressive process for $\lambda_t$ of the form

$$\lambda_{t+1} = \begin{cases} 
\omega_1 + \phi_1 \lambda_t + \tau_1 y_t, & \text{if } y_t \leq k, \\
\omega_2 + \phi_2 \lambda_t + \tau_2 y_t, & \text{if } y_t > k,
\end{cases}$$
where $\omega_i > 0$, $\phi_i \geq 0$, $\tau_i > 0$, for $i = 1, 2$, and $k \in \mathbb{N}$. From the theory that was developed in Section 3, we can derive conditions for stationarity and ergodicity and for the finiteness of the second moment of $y_t$. In particular, from theorem 1, we obtain that $\phi_m + \tau_2 < 1$ is sufficient for the stationarity and ergodicity of the BNB self-excited threshold auto-regressive process, where $\phi_m = \max\{\phi_1, \phi_2\}$. Moreover, from theorem 2, we obtain that the additional condition

$$\frac{(r + 1)(\alpha - 1)}{r(\alpha - 2)} \tau_2^2 + \phi_m^2 + 2\tau_2\phi_m < 1$$

is sufficient to ensure that $\mathbb{E}(y_t^2) < \infty$.

5.2. Monte Carlo simulation study

We investigate the small sample properties of the ML estimator by means of a Monte Carlo simulation experiment. We generate samples from the BNB–INGARCH model (8) for various sample sizes ($T = 250, 500, 1000, 2000$). The parameters are then estimated by using the ML estimator (7). Fig. 1 reports the kernel density of the ML estimates. Additional simulation results for different parameter values are given in the on-line supplementary material. The parameters $r$ and $\alpha$ are reparameterized in terms of their inverse. This is done because, especially in small samples, a given realized path from the BNB–INGARCH process may not present outliers and the estimate of $\alpha$ may become arbitrarily large since the likelihood function is flat for large $\alpha$.

The results confirm that the estimates are consistent since their distribution collapses towards the true parameter values as the sample size increases. Furthermore, the distributions of the ML estimators of $\delta$, $\delta = \omega/(1 - \phi - \tau)$, and $\tau$ seem symmetric for all sample sizes. This suggests that the asymptotic normal distribution is an accurate approximation. The results show a more skewed distribution for the estimates of $\phi$, $r$ and $\alpha$, especially for small $T$. We also highlight the presence of a small bump in the distribution of $r$ around the point $r = (\alpha_0 - 1)\delta_0/r_0$. This is not surprising since, from the identification of $r$, we expect the distribution of $r$ to have a second lower mode in small samples.

Finally, we note that the parameters $\omega$, $\phi$ and $\tau$, which determine the dynamic properties of $\lambda_t$, can also be estimated by alternative methods such as quasi-ML (QML). The Poisson QML estimator is often considered in the literature and it delivers consistent parameter estimates (Ahmad and Francq, 2016). We provide additional simulation results, which are reported in the on-line supplement, to compare the accuracy of ML with Poisson QML. The results show that ML performs substantially better than Poisson QML in terms of mean-squared error, especially for small values of the tail parameter $\alpha$. This finding is quite intuitive since we expect ML based on the BNB PMF to outperform competing estimation methods when outliers are present in the data.

6. A score-driven beta–negative binomial auto-regression

The BNB–INGARCH model that was presented in Section 5 accounts for extreme observations by means of the heavy tail of the conditional BNB distribution. However, the SRE for $\lambda_t$ is not robust against outliers: an extreme value of $y_t$ can have an arbitrary large effect on $\lambda_t+1$. In practice, it is often desirable to have a robust SRE that attenuates the effect of extreme observations on $\lambda_t$. We propose a robust specification that is based on the GAS framework of Creal et al. (2013) and Harvey (2013). The conditional mean $\lambda_t$ is specified as an auto-regressive process with innovation given by the score of the predictive likelihood. Score-driven models are widely used in the literature to specify robust updating functions in models with heavy-tailed distributions (Harvey and Luati, 2014; Opschoot et al., 2017).
Fig. 1. Kernel density estimator for the distribution of the ML estimates obtained from 1000 Monte Carlo replications for sample sizes 250, 500, 1000, and 2000 (true parameter values; the bandwidth of the kernel is selected by using Silverman's rule of thumb; the parameter δ represents the unconditional mean of yt, i.e., δ = ω = 1 - φ - ρ; the true parameter vector is (δ₀, φ₀, ρ₀, r₀, ν₀) = (10, 0.68, 0.2, 10, 5)). (a) δ; (b) φ; (c) τ; (d) r; (e) α.
Fig. 2. Effect of $y_t$ on the score innovation $s_t$ for various values of the tail parameter $\alpha$: , $\alpha = 5$; , $\alpha = 10$; , $\alpha = 30$; , $\alpha \to \infty$

The BNB auto-regression with a score-driven SRE, which we label BNB–GAS, is specified by the equations

$$y_t|\mathcal{F}_{t-1} \sim \text{BNB}(\lambda_t, r, \alpha),$$
$$\log(\lambda_{t+1}) = \omega + \phi \log(\lambda_t) + \tau s_t,$$

where $\omega \in \mathbb{R}$, $\phi \geq 0$ and $\tau > 0$. The score innovation $s_t = \partial \log \{p(y_t|\lambda_t, r, \alpha)\}/\partial \log(\lambda_t)$ is given by

$$s_t = \gamma \lambda_t \{\psi(\gamma \lambda_t + y_t) + \psi(\gamma \lambda_t + \alpha) - \psi(\gamma \lambda_t + y_t + \alpha + r) - \psi(\gamma \lambda_t)\},$$

where $\gamma = (\alpha - 1)/r$ and $\psi(\cdot)$ denotes the digamma function, i.e. $\psi(x) = \partial \log \{\Gamma(x)\}/\partial x$. The exponential link function in expression (10) is considered to ensure that $\lambda_t$ is strictly positive with probability 1. We refer to the score $s_t$ as the innovation of the process since $\mathbb{E}(s_t|\mathcal{F}_{t-1}) = 0$ with probability 1.

The peculiarity of BNB–GAS is that the functional form of the score innovation reduces the effect of outliers. Fig. 2 illustrates the sensitivity of $s_t$ to $y_t$ for various values of the tail parameter $\alpha$. We can see that the effect of large values of $y_t$ on $s_t$ is attenuated and the degree of attenuation depends on the tail parameter $\alpha$. The smaller the parameter $\alpha$ the more robust the score innovation $s_t$ is. This behaviour of $s_t$ is quite intuitive since a small $\alpha$ introduces heavy tails in the conditional PMF of $y_t$ and therefore it generates outliers in the observed time series; see also Harvey and Luati (2014) for a similar interpretation in the context of Student’s $t$-distributions. Furthermore, as is shown in the proof of theorem 7 below, the score innovation is bounded by a constant $|s_t| \leq \alpha + r + 1$. Therefore, $s_t$ is robust since it does not go to $\infty$ as $y_t \to \infty$.

BNB–GAS can approximate arbitrarily well some existing models that have been proposed in the literature. As $\alpha \to \infty$, BNB–GAS becomes a score-driven model with negative binomial distribution; see Gorgi (2018) for an application of the GAS framework with negative binomials. As additionally $r \to \infty$, the model becomes a Poisson auto-regressive model, which belongs to the class of models of Davis et al. (2003).

We now focus on the stochastic properties of BNB–GAS. The next theorem gives sufficient conditions for the existence of a stationary and ergodic solution for the BNB–GAS process.
Theorem 7. Consider the BNB–GAS process given in expressions (10) and (11) and let \(0 \leq \phi < 1\) and

\[
\{\tau(3\alpha + 2r + 3) + \phi\} \exp\{\tau(2\alpha + r + 2)\} < 1.
\]

Then, the process admits a stationary and ergodic solution with finite first moment \(\mathbb{E}(y_t) < \infty\).

The proof of theorem 7, which is given in the on-line supplement, is obtained by an application of theorem 1. The inequality in expression (12) is a parameter restriction that makes the semicontraction condition (5) hold. Given the complex functional form of \(s_t\), it is not straightforward to obtain a sharper upper bound for the Lipschitz coefficient \(b\) in equation (5). We note that, if inequality (12) is not satisfied, it does not mean that the process is non-stationary since the semicontraction condition in theorem 1 is sufficient but not necessary for stationarity. Condition (12) may be restrictive in practice. For instance, it is not satisfied for the parameter values that are estimated in the empirical applications in Section 7. However, this condition highlights that the stationarity region is not degenerate.

Remark 1. Under the conditions of theorem 7, the BNB–GAS process admits a stationary and ergodic solution with \(\mathbb{E}(y_t^m) < \infty\) if and only if \(\alpha > m\). This result holds true because \(\lambda_t\) takes values on a compact set with probability 1. This means that \(\mathbb{E}(\lambda_t^m) < \infty\) for any \(m \in \mathbb{N}\). Therefore, \(\mathbb{E}(y_t^m) < \infty\) if and only if \(\mathbb{E}(y_t^m|\mathcal{F}_{t-1}) < \infty\) with probability 1, namely, \(\alpha > m\).

Finally, we note that the BNB–GAS model (10) can be generalized by considering a different rescaling of the score innovation \(s_t\). The conditional Fisher information is often used for this; see Creal et al. (2013) for a discussion. A simple rescaling for BNB–GAS can be obtained by multiplying the score innovation by \(\lambda_t^{-k}\), \(k \in [0, 1]\), to make the magnitude of the innovation depend on the level of \(\lambda_t\). In the limit case, \(\alpha, r \to \infty\), the resulting class of models corresponds to the class of Poisson processes of Davis et al. (2003).

7. Empirical applications

7.1. Police reports on narcotics trafficking

In this section, we present an empirical application to the monthly number of police reports on narcotics trafficking in Sydney, Australia. The time series is from January 1995 to December 2016 and it is part of the New South Wales data set of police reports, which is available from http://www.bocsar.nsw.gov.au/. The data set can be found in the section ‘Datasets’ under the name ‘Recorded crime by offence’ and geographic breakdown by ‘LGA’. The code and data to reproduce the empirical results are available from

https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets.

Fig. 3 displays the plot and the empirical auto-correlation functions of the series. The sample mean and variance of the series are 8.4 and 79.5 respectively. We can see that the data set presents some extreme observations. In particular, the number of narcotics trafficking reports is exceptionally high in August 2000, March 2008 and May 2015. Therefore, BNB auto-regressive models seem particularly suited to describe the auto-correlation structure and to account for the outliers in the data.

Besides the BNB–INGARCH (8) and the BNB–GAS (10), we consider two negative binomial specifications: one with a linear updating function and one based on the GAS framework, which we label NB–INGARCH and NB–GAS respectively. As discussed in Sections 5 and 6, these two models are limit cases, \(\alpha \to \infty\), of the BNB–INGARCH and the BNB–GAS model. We
Fig. 3. (a) Monthly number of police reports for narcotics trafficking in Sydney and (b) sample auto-correlation function and (c) partial auto-correlation function of the series refer the reader to the on-line supplement for the specification of the negative binomial models. BNB–INGARCH and NB–INGARCH have the same specification for the conditional mean $\lambda_t$. Therefore, under correct specification of BNB–INGARCH, the parameter estimates from NB–INGARCH may be seen as negative binomial QML estimates. Theoretical results on negative binomial QML were derived in Aknouche et al. (2018), where a two-stage QML estimator was considered. In contrast, BNB–GAS and NB–GAS do not have the same specification for $\lambda_t$ and only the conditional mean of BNB–GAS is robust against outliers. Model selection between BNB and negative binomial models and among different specifications of $\lambda_t$ can be implemented by using information criteria, such as the Akaike information criterion AIC and Bayesian information criterion BIC, or by using a validation sample. We note that, although the BNB models can approximate arbitrarily well their corresponding negative binomial versions as $\alpha \to \infty$, a formal testing procedure for the null hypothesis $\alpha^{-1} = 0$ does not follow from the asymptotic results in Section 4. This is because $\alpha^{-1} = 0$ is a boundary point of the parameter space.

Table 1 reports the estimates of the models and several measures of in-sample fit and forecasting accuracy. The models BNB–static and NB–static are BNB and negative binomial models with constant conditional expectation, i.e. $\lambda_t$ is a static parameter to be estimated. The log-score criterion is used to measure the accuracy of PMF forecasts (Geweke and Amisano, 2011) and the mean absolute error MAE is used to measure the accuracy of point forecasts. The log-score criterion and MAE are obtained from the one-step-ahead predictions of the observations in the second half of the sample (the last 132 observations). A recursive estimation procedure is considered, where the models are re-estimated for each prediction using all previous observations. More specifically, the prediction of $y_t$ at a given time point $t = s + 1$ is obtained by estimating the models using the observations from time $t = 1$ to time $t = s$. In this way, the estimation win-
Table 1. ML estimates of the models for the series of narcotics trafficking reports in Sydney†

| Model                   | δ     | φ     | τ     | r     | α     | Log-likelihood | AIC     | BIC     | Log-score | MAE  |
|-------------------------|-------|-------|-------|-------|-------|----------------|---------|---------|-----------|------|
| BNB–INGARCH             | 8.549 | 0.481 | 0.267 | 6.301 | 4.819 | −807.66        | 1625.33 | 1643.21 | −416.58   | 6.39 |
|                         | (1.104) | (0.223) | (0.085) | (3.394) | (0.744) |                |         |         |           |      |
| BNB–GAS                 | 2.087 | 0.714 | 0.197 | 4.408 | 5.029 | −807.04        | 1624.09 | 1641.97 | −416.24   | 6.15 |
|                         | (0.107) | (0.169) | (0.056) | (1.923) | (0.849) |                |         |         |           |      |
| BNB–static              | 8.449 | —     | —     | 5.085 | 4.060 | −825.10        | 1656.20 | 1666.93 | −422.66   | 6.36 |
|                         | (0.567) |       |       | (6.582) | (0.582) |                |         |         |           |      |
| NB–INGARCH              | 8.675 | 0.307 | 0.308 | 1.561 | —     | −821.22        | 1650.45 | 1664.75 | −426.26   | 6.40 |
|                         | (0.907) | (0.254) | (0.099) | (0.159) | —     |                |         |         |           |      |
| NB–GAS                  | 2.102 | 0.699 | 0.140 | 1.540 | —     | −822.59        | 1653.19 | 1667.49 | −426.12   | 6.42 |
|                         | (0.087) | (0.232) | (0.051) | (0.156) | —     |                |         |         |           |      |
| NB–static               | 8.433 | —     | —     | 1.418 | —     | −832.73        | 1669.47 | 1676.62 | −425.49   | 6.37 |
|                         | (0.472) |       |       | (0.141) | —     |                |         |         |           |      |

†The last five columns contain the log-likelihood, AIC, BIC, log-score criterion and mean absolute error MAE. Standard errors (in parentheses) are obtained from the inverse of the observed Fisher information. A more detailed explanation on how the standard errors and the criteria are computed is given in the on-line supplement. The parameter \( \delta \) is \( \delta = \omega / (1 - \phi - \tau) \) for the INGARCH models and \( \delta = \omega / (1 - \phi) \) for the GAS models, and it is equal to the constant conditional expectation for the static models. Values with the best performance are highlighted in italics.

The results show that the BNB specifications give a better in-sample fit since they have lower values for AIC and BIC. The relevance of the BNB distribution can also be elicited from the relatively low estimates of the tail parameter \( \alpha \), which is estimated to be around 5 with a standard error of about 0.8 for both BNB–INGARCH and BNB–GAS. This means that the estimated conditional distribution of the data is heavy tailed with only four or five finite moments. BNB models deliver more accurate PMF forecasts since they have a larger log-score criterion compared with negative binomial models. This finding indicates that BNB models can deliver more accurate predictions for the probability of occurrence of certain events when the estimated tail parameter \( \alpha \) is relatively small. The more accurate point forecasts are given by BNB–GAS and all the other models have a similar performance in terms of MAE. This suggests that the use of a BNB conditional distribution is not relevant for point forecasts. However, the robust specification of \( \lambda \) of BNB–GAS is beneficial in this application. An illustration of how BNB–GAS attenuates the effect of outliers is given in the on-line supplementary material.

Diagnostic analyses, which are reported in the supplement, show that the BNB specifications are well suited to model the series. In particular, the Pearson residuals show no significant auto-correlation and parametric resampling confirms that the fitted model can properly capture the auto-correlation structure of the time series. Furthermore, the probability integral transformation histogram indicates that the BNB conditional PMF can accurately describe the distribution of the data.

7.2. Euro–pound sterling exchange rate

We present a second empirical application where we analyse the euro to British pound exchange rate EUR/GBP on December 12th, 2019. On this day, general elections were held in the UK. We consider the number of tick changes by minute of EUR/GBP from 9.00 a.m. to 9.00 p.m. (Greenwich mean time). As is well known, prices of financial assets change over a discrete
grid since the minimum price variation is determined by the tick size (Koopman et al., 2017). The series is constructed by using high frequency prices from the foreign exchange market. In particular, the closing price of each minute for EUR/GBP is considered and the number of tick changes is computed by dividing the absolute price variation by the tick size, which is $10^{-5}$. More specifically, denoting with $p_t$ the closing price at minute $t$, the number of tick changes of the exchange rate at minute $t$ is $n_t = |p_t - p_{t-1}| \times 10^5$. The sample size of the series is 720 observations, which corresponds to the number of minutes in the 12 h of the sample. The data were obtained from http://www.histdata.com/. The web site provides high frequency transaction data for several currencies that are traded in the foreign exchange market. The EUR/GBP-data can be found in the section ‘Download free Forex data’ under ‘M1 (1 minute bar) data’, selecting the desired format of the data file.

Fig. 4 shows the plot and the empirical auto-correlation functions of the series. The sample mean is 13.2 and the sample variance is 224.2. The series displays several extreme observations and the auto-correlation functions indicate that there is significant auto-correlation in the data. These features suggest that BNB auto-regressions are well suited for the modelling of the series and to describe the extreme observations by means of the heavy tail of the BNB distribution. We note that studying and predicting the probability of tail events of prices of financial assets are of practical interest to measure and forecast financial risk (Embrechts et al., 2013).

Table 2 reports the estimation results as considered in the previous empirical application. The log-score criterion and MAE are obtained from the prediction errors of the observations in the second half of the sample (the last 360 observations). The results show that the BNB models have a better in-sample fit in terms of AIC and BIC. Furthermore, PMF forecasts from BNB models are more accurate as shown by the log-score criterion. However, the point forecast
### Table 2. ML estimates of the models for the EUR/GBP-series†

| Model         | δ   | φ   | τ   | r   | α   | Log-likelihood | AIC  | BIC  | Log-score | MAE  |
|---------------|-----|-----|-----|-----|-----|----------------|------|------|-----------|------|
| BNB–INGARCH   | 13.263 | 0.837 | 0.094 | 1.918 | 5.390 | −2558.01       | 5126.01 | 5148.91 | −1308.54 | 9.66 |
|               | (1.230) | (0.069) | (0.032) | (0.238) | (1.172) |                 |      |      |           |      |
| BNB–GAS       | 2.559  | 0.915 | 0.089 | 1.688 | 7.079 | −2562.51       | 5135.01 | 5157.91 | −1311.18 | 9.75 |
|               | (0.076) | (0.056) | (0.031) | (0.177) | (2.030) |                 |      |      |           |      |
| BNB–static    | 13.218 | —    | —    | 1.899 | 4.496 | −2580.59       | 5167.17 | 5180.91 | −1320.73 | 9.95 |
|               | (0.567) | (0.265) | (0.937) | (0.265) | (0.937) |                 |      |      |           |      |
| NB–INGARCH    | 13.389 | 0.843 | 0.090 | 1.284 | —    | −2573.36       | 5154.71 | 5173.03 | −1317.77 | 9.66 |
|               | (1.102) | (0.050) | (0.025) | (0.072) | —    |                 |      |      |           |      |
| NB–GAS        | 2.561  | 0.919 | 0.076 | 1.287 | —    | −2572.44       | 5152.88 | 5171.20 | −1317.29 | 9.63 |
|               | (0.072) | (0.040) | (0.019) | (0.072) | —    |                 |      |      |           |      |
| NB–static     | 13.228 | —    | —    | 1.202 | —    | −2596.75       | 5197.49 | 5206.65 | −1330.06 | 9.96 |
|               | (0.470) | (0.066) |             |     | —    |                 |      |      |           |      |

†The last five columns contain the log-likelihood, AIC, BIC, log-score criterion and MAE. Standard errors (in parentheses) are obtained from the inverse of the observed Fisher information. A more detailed explanation on how the standard errors and the criteria are computed is given in the on-line supplement. The parameter δ is \( \delta = \omega / (1 - \phi - \tau) \) for the INGARCH models and \( \delta = \omega / (1 - \phi) \) for the GAS models, and it is equal to the constant conditional expectation for the static models. Values with the best performance are highlighted in italics.

Accuracy is similar across the models and the BNB models do not dominate the negative binomial models. This further confirms that the key advantage of the BNB models is in describing extreme observations as tail events and hence delivering more accurate PMF predictions. In contrast, we note that the accuracy of point predictions mostly depend on the specification of \( \lambda_t \). BNB–INGARCH and NB–INGARCH have the same specification for \( \lambda_t \) and therefore a similar point forecast accuracy is to be expected. BNB–INGARCH performs better than BNB–GAS in sample as well as out of sample. Therefore, the robust conditional mean of BNB–GAS does not provide benefits in this application. Overall, the dynamic models with a time-varying conditional mean \( \lambda_t \) perform better than the static models.

Also for this application, diagnostic analyses, which are reported in the on-line supplement, show that the BNB models are a good fit for this data set in terms of both modelling the auto-correlation of the series and describing the distribution of the data.

### 8. Conclusion

This paper introduces a general framework for modelling integer-valued time series with outliers. The paper proposes a class of observation-driven models that are based on a mixture of negative binomial distributions, known as the BNB distribution. The stochastic properties of the models and the asymptotic theory of ML estimation are formally discussed. Various specifications of the conditional mean are considered and studied. Two empirical applications illustrate the practical relevance of the approach. Further research may focus on extending the proposed method to more flexible specifications. Relevant developments may include embedding the model with a zero-inflated BNB distribution to handle time series with a large number of 0s and introducing time variation in the parameters \( r \) and \( \alpha \) to account for smooth changes in the shape of the distribution. Another relevant line of research may concern the development of a formal testing procedure to select between BNB and negative binomial models, which would enable detection of the presence of heavy tails in integer-valued time series data.
Acknowledgements
The author thanks two referees and the Associate Editor for their useful comments, which helped to improve the paper.

Appendix A

A.1. Proofs

A.1.1. Proof of theorem 1
The existence of a stationary and ergodic solution follows directly from an application of theorem 2.2 of Doukhan and Neumann (2019) by checking that conditions A1–A3 in Doukhan and Neumann (2019) are satisfied. Furthermore, the $\mathcal{F}_{t-1}$-measurability of $\lambda_t$ is also implied by the proof of theorem 2.2 of Doukhan and Neumann (2019) since it is shown that $\lambda_t$ can be expressed as a measurable function of $\{y_{t-1}, y_{t-2}, \ldots\}$. Below we show that conditions A1–A3 hold. It is straightforward to see that condition A1 is implied by the drift condition in equation (4) and assumption A2 is implied by the semicontraction condition (5). Condition A3 requires an upper bound on the total variation distance between two BNB distributions with different means. In particular, denoting with $\text{BNB}_{\lambda_1}$ the probability measure of a BNB distribution with mean $\lambda_1$, $\text{BNB}(\lambda_1, r, \alpha)$, and with $\text{BNB}_{\lambda_2}$ the probability measure of a BNB distribution with mean $\lambda_2$, $\text{BNB}(\lambda_2, r, \alpha)$, condition A3 requires the following upper bound to hold for some $\delta > 0$:

$$TV(\text{BNB}_{\lambda_1}, \text{BNB}_{\lambda_2}) \leq 1 - \exp(-\delta |\lambda_1 - \lambda_2|),$$

where the total variation distance is defined as

$$TV(\text{BNB}_{\lambda_1}, \text{BNB}_{\lambda_2}) = \sup_{A \in \mathcal{B}} |\text{BNB}_{\lambda_1}(A) - \text{BNB}_{\lambda_2}(A)|.$$

We obtain that the above upper bound on the total variation distance holds with $\delta = 1$ by lemma 7 in Appendix A.2.

Finally, we conclude the proof by showing that $\mathbb{E}(y_t) < \infty$. In what follows, we use the shorthand notation $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot |\mathcal{F}_t)$ to denote conditional expectations. The drift condition (4) implies that

$$\lambda_t \leq c_0 + c_1 y_{t-1} + c_2 \lambda_{t-1},$$

with probability 1. Therefore, taking the conditional expectation of $\lambda_t$, we obtain

$$\mathbb{E}_{t-2} (\lambda_t) \leq c_0 + c_1 \mathbb{E}_{t-2} (y_{t-1}) + c_2 \lambda_{t-1},$$

$$= \tilde{a}_{10} + \tilde{a}_{11} \lambda_{t-1},$$

where $\tilde{a}_{10} = c_0$ and $\tilde{a}_{11} = c_1 + c_2$. Reiterating conditional expectations, we have that

$$\mathbb{E}_{t-k} (\lambda_t) \leq \sum_{i=0}^{k-2} \tilde{a}_{i1} \tilde{a}_{10} + \tilde{a}_{i1}^{k-1} \lambda_{t-k+1}.$$

Since $\tilde{a}_{11} < 1$ by the drift condition (5), we can take the limit $k \to \infty$ and obtain $\mathbb{E}(\lambda_t) \leq (1 - \tilde{a}_{11})^{-1} \tilde{a}_{10}$, which implies that $\mathbb{E}(\lambda_t) < \infty$ and hence $\mathbb{E}(y_t) < \infty$ since $\alpha > 1$.

A.1.2. Proof of theorem 2
Given inequality (13), we derive the following upper bound for $\lambda_t^2$:

$$\lambda_t^2 \leq c_0^2 + c_1^2 y_{t-1}^2 + c_2^2 \lambda_{t-1}^2 + 2c_0c_1 y_{t-1} + 2c_0c_2 \lambda_{t-1} + 2c_1c_2 y_{t-1} \lambda_{t-1}.$$

Now, noting that the conditional expectation of $y_t^2$ can be expressed as

$$\mathbb{E}_{t-1} (y_t^2) = \frac{\alpha + r - 1}{\alpha - 2} \lambda_{t-1} + \frac{(r+1)(\alpha - 1)}{r(\alpha - 2)} \lambda_{t-1}^2,$$

we obtain that

$$\mathbb{E}_{t-2} (\lambda_t^2) \leq c_0^2 + c_1^2 \mathbb{E}_{t-2} (y_{t-1}^2) + c_2^2 \lambda_{t-1}^2 + 2c_0c_1 \mathbb{E}_{t-2} (y_{t-1}) + 2c_0c_2 \lambda_{t-1} + 2c_1c_2 \mathbb{E}_{t-2} (y_{t-1}) \lambda_{t-1}.$$
Calculating the conditional expectations on the right-hand side of the above equation yields
\[ \mathbb{E}_{t-2}(\lambda_t^2) \leq \tilde{a}_{20} + \tilde{a}_{21} \lambda_{t-1} + \tilde{a}_{22} \lambda_{t-1}^2, \] (15)
where
\[ \tilde{a}_{20} = \alpha^2, \]
\[ \tilde{a}_{21} = \frac{\alpha + r - 1}{\alpha - 2} - c_1^2 + 2c_0c_1 + 2c_0c_2 \]
and
\[ \tilde{a}_{22} = \frac{(r + 1)(\alpha - 1)}{\alpha - 2} - c_1^2 + c_2^2 + 2c_1c_2. \]

Therefore, combining equations (14) and (15), we can write the following bivariate system for \( x_t = (\lambda_t, \lambda_t^2)^T \):
\[ \mathbb{E}_{t-2}(x_t) \leq d_0 + D_1 x_{t-1}, \]
where
\[ d_0 = \begin{pmatrix} \tilde{a}_{10} \\ \tilde{a}_{20} \end{pmatrix}, \]
\[ D_1 = \begin{pmatrix} \tilde{a}_{11} & 0 \\ \tilde{a}_{21} & \tilde{a}_{22} \end{pmatrix}. \]

Reiterating conditional expectations, we obtain that
\[ \mathbb{E}_{t-k}(x_t) \leq \sum_{j=0}^{k-2} D_1^j d_0 + D_1^{k-1} x_{t-k+1}. \]

Since \( D_1 \) is a lower triangular matrix, the eigenvalues of \( D_1 \) are the diagonal entries and by assumption they are smaller than 1, \( \tilde{a}_{11} \leq \tilde{a}_{22} < 1 \). Therefore, taking the limit \( k \to \infty \) we obtain \( \mathbb{E}(x_t) \leq (I - D_1)^{-1} d_0 \), which implies that \( \mathbb{E}(\lambda_t^2) < \infty \). The final result follows noting that \( \mathbb{E}(\lambda_t^2) < \infty \) entails \( \mathbb{E}(y_t^2) < \infty \) since \( \alpha > 2 \).

### A.1.3. Proof of proposition 1

The proof of the invertibility result is obtained by an application of theorem 3.1 of Bougerol (1993). We follow Straumann and Mikosch (2006) (proposition 3.12) and apply Bougerol’s theorem in the space of continuous functions \( C(\Theta, \mathbb{R}) \) equipped with the uniform norm \( \| \cdot \|_\Theta \). We note that the \( \mathcal{F}_{t-1} \)-measurability of \( \lambda_t \) and \( \lambda_t(\theta_0) = \lambda_t \) also follow from proposition 3.12 of Straumann and Mikosch (2006). Our SRE is of the form
\[ \hat{\lambda}_{t+1}(\theta) = g_\theta \{ y_t, \hat{\lambda}_t(\theta) \}, \quad t \in \mathbb{N}. \]

First, we note that our SRE satisfies the stationarity and continuity requirements to apply Bougerol’s theorem in \( C(\Theta, \mathbb{R}) \). In particular, the function \( \theta \mapsto \hat{\lambda}_t(\theta) \) is continuous for any \( t \in \mathbb{N} \) by assumption 1, and the sequence \( \{ g_\theta(y_t, \lambda) \} \) is stationary and ergodic for any \( \lambda \in \mathbb{R}^+ \) by theorem 1. Next, we show that conditions C1 and C2 in theorem 3.1 of Bougerol (1993) are satisfied. As concerns condition C1, we obtain that
\[ \mathbb{E}\{ \log^+ \| g_\theta(y_t, \lambda) \|_\Theta \} \leq c_0 + c_1 \mathbb{E}(y_t) + c_2 \lambda < \infty \]

since the drift condition (4) holds over \( \Theta \) by assumption 2 and \( \mathbb{E}(y_t) < \infty \) by theorem 1. Finally, condition C2 is immediately satisfied as
\[ \frac{|g_\theta(y_t, \lambda) - g_\theta(y_t, \lambda^*)|}{|\lambda - \lambda^*|} \leq b < 1 \]
with probability 1 for any \( \theta \in \Theta \) by assumption 2.
Finally, we conclude the proof by showing that \( \mathbb{E}\|\tilde{\lambda}_t\|_\Theta < \infty \). By assumption 2, we obtain that with probability 1
\[
\|\tilde{\lambda}_t\|_\Theta \leq c_0 + c_1 \gamma_{t-1} + c_2 \|\tilde{\lambda}_{t-1}\|_\Theta \\
\leq \sum_{i=0}^{\infty} c_2 (c_0 + c_1 \gamma_{t-1-i}) + c_2 \|\tilde{\lambda}_0\|_\Theta.
\]
Therefore, since \( c_2 < 1 \), for sufficiently large \( t \) we obtain that
\[
\|\tilde{\lambda}_t\|_\Theta \leq 1 + \sum_{i=0}^{\infty} c_2 (c_0 + c_1 \gamma_{t-1-i}),
\]
and by stationarity we have that the above inequality holds for any \( t \). Therefore, the desired result follows by noting that \( \mathbb{E}(y_t) < \infty \) by theorem 1.

A.1.4. Proof of theorem 3
In what follows we show

(a) uniform convergence of the likelihood, i.e. \( \|L_T - L\|_K \to 0 \), almost surely, as \( T \to \infty \) (here the limit likelihood function \( L \) is defined as \( L(\kappa) = \mathbb{E}\{l_t(\kappa)\} \)) and

(b) identifiability of \( \kappa_0 \), i.e. \( L(\kappa_0) > L(\kappa) \forall \kappa \in K, \kappa \neq \kappa_0 \).

Then, given conditions (a) and (b) and the compactness of \( K \), the consistency \( \hat{\kappa}_T \to \kappa_0 \), almost surely, follows immediately by standard arguments that go back to Wald (1949).

(a) An application of the triangle inequality yields
\[
|\hat{L}_T(\kappa) - L(\kappa)| \leq |\hat{L}_T(\kappa) - L_T(\kappa)| + |L_T(\kappa) - L(\kappa)|,
\]
where \( L_T(\kappa) = (1/T) \sum_{t=1}^{T} l_t(\kappa) \). Therefore, the desired result follows if we can show that both terms on the right-hand side of inequality (16) are vanishing almost surely and uniformly as \( T \) diverges.

As concerns the first term in inequality (16), assumption 3 implies that \( \lambda_{t}(\theta) > \bar{c} \) with probability 1 for any \( \theta \in \Theta \). An application of the mean value theorem together with lemma 10 in Appendix A.2 yields
\[
|\hat{L}_t(\kappa) - L_t(\kappa)| \leq \sup_{\kappa \in K} \sup_{\lambda \in [0, \infty)} \partial \log p(y_t | \lambda, r, \alpha) \left| \hat{\lambda}_t(\theta) - \lambda_t(\theta) \right| \\
\leq c_3 |\hat{\lambda}_t(\theta) - \lambda_t(\theta)|.
\]
As a result, since \( \|\hat{\lambda}_t - \hat{\lambda}_0\|_\Theta \to 0 \), almost surely, by proposition 1, we obtain \( \|\hat{L}_t - L_t\|_K \to 0 \), almost surely, and therefore
\[
\|\hat{L}_T - L_T\|_K \leq \frac{1}{T} \sum_{t=1}^{T} \|\hat{L}_t - L_t\|_K \to 0, \quad \text{almost surely.}
\]
As concerns the second term in inequality (16), given the continuity of the log-likelihood and the compactness of \( K \), we obtain that \( \|L_T - L\|_K \to 0 \), almost surely, follows by an application of the ergodic theorem of Rao (1962), provided that the log-likelihood has a uniformly bounded moment, i.e. \( \mathbb{E}\|l_t\|_K < \infty \). We note that the log-likelihood cannot be bounded as follows:
\[
|l_t(\kappa)| \leq |\log \{B(r, \alpha)\}| + |\log \{\Gamma(y_t + r)\} - \log \{\Gamma(y_t + 1)\}| \\
+ |\log \{\{(\alpha - 1)\hat{\lambda}_t(\theta)/r + y_t + \alpha + r\}\} - \log \{\{(\alpha - 1)\lambda_t(\theta)/r + y_t\}\}| \\
+ |\log \{\{(\alpha - 1)\hat{\lambda}_t(\theta)/r + \alpha\}\} - \log \{\{(\alpha - 1)\lambda_t(\theta)/r\}\}|.
\]
Therefore, \( \mathbb{E}\|l_t\|_K < \infty \) follows by an application of lemma 9 below since \( \inf_{\theta \in \Theta} \lambda_t(\theta) \geq \bar{c} \), almost surely, by assumption 3, \( \mathbb{E}(y_t) < \infty \) by theorem 1 and \( \mathbb{E}\|\hat{\lambda}_t\|_\Theta < \infty \) by proposition 1. This concludes the proof of part (a).
(b) First, we show that \( l_\kappa (\kappa) = l_\kappa (\kappa_0) \) almost surely if and only if \( \kappa = \kappa_0 \). From lemma 8 below and the non-degeneracy of \( \lambda_t = \bar{\lambda}_t (\theta_0) \) in assumption 4, we obtain that

\[
p\{ y | \lambda_t (\theta), r, \alpha \} = p\{ y | \lambda_t, r_0, \alpha_0 \} \quad \forall \ y \in \mathbb{N}
\]

almost surely, if and only if \( \{ \lambda_t (\theta), r, \alpha \} = \{ \lambda_t, r_0, \alpha_0 \} \) almost surely.

Therefore, the desired result follows by showing that \( \lambda_t (\theta) = \lambda_t \) almost surely, if and only if \( \theta = \theta_0 \).

We prove the result by contradiction. Assume that \( \theta \neq \theta_0 \) and \( \lambda_t (\theta) = \lambda_t \) almost surely; then, given the stationarity of \( \lambda_t \), we have that \( \lambda_t (\theta) = \lambda_t \) almost surely for any \( t \). Therefore, it must be true that \( g_\theta (y_t, \lambda_t) = g_{\theta_0} (y_t, \lambda_t) \) almost surely. However, this contradicts assumption 4. We conclude that \( \lambda_t (\theta) \neq \lambda_t \) if and only if \( \theta = \theta_0 \) and therefore \( l_t (\kappa) = l_t (\kappa_0) \) almost surely, if and only if \( \kappa = \kappa_0 \).

Finally, we show that \( l_t (\kappa) = l_t (\kappa_0) \) almost surely, if and only if \( \kappa = \kappa_0 \) entails the identifiability condition (b). It is well known that \( \log(x) \leq x - 1 \) for any \( x \in \mathbb{R}^+ \) with the equality only in the case \( x = 1 \). This implies that almost surely

\[
l_t (\kappa) - l_t (\kappa_0) \leq \frac{p\{ y_t | \tilde{\lambda}_t (\theta), r, \alpha \}}{p\{ y_t | \lambda_t, r_0, \alpha_0 \}} - 1.
\]  

(17)

Moreover, we have that the inequality in expression (17) holds as a strict inequality with positive probability because \( p\{ y_t | \lambda_t (\theta), r, \alpha \} \neq p\{ y_t | \lambda_t, r_0, \alpha_0 \} \) with positive probability for any \( \kappa \in K, \kappa \neq \kappa_0 \).

As a result

\[
E [E \{ l_t (\kappa) - l_t (\kappa_0) | F_{t-1} \}] \leq E \left( \frac{p\{ y_t | \tilde{\lambda}_t (\theta), r, \alpha \}}{p\{ y_t | \lambda_t, r_0, \alpha_0 \}} \right) - 1 = 0 \quad \forall \ \kappa \neq \kappa_0,
\]

where the right-hand side of the inequality is equal to 0 as \( p\{ y_t | \lambda_t, r_0, \alpha_0 \} \) is the true conditional PMF. Finally, the desired result \( L(\kappa_0) > L(\kappa) \) follows by the law of total expectation:

\[
L(\kappa) - L(\kappa_0) = E [E \{ l_t (\kappa) - l_t (\kappa_0) | F_{t-1} \}] < 0 \quad \forall \ \kappa \neq \kappa_0.
\]

This concludes the proof of part (b).

A.1.5. Proof of theorem 4

For the asymptotic normality proof, we follow a similar argument to that in the proof in section 7 of Straumann and Mikosch (2006). First, we derive the asymptotic distribution of the ML estimator \( \tilde{\kappa}_T \) based on the limit likelihood \( L_T \), which is defined as

\[
\tilde{\kappa}_T = \arg \sup_{\kappa \in K} L_T (\kappa).
\]

Then, we show that \( \hat{\kappa}_T \) and \( \tilde{\kappa}_T \) have the same asymptotic distribution.

From the uniform convergence in lemmas 1 and 2 in Appendix A.2, we have that \( \bar{\lambda}_t \) is twice continuously differentiable in \( \theta \) with first and second derivatives given by \( \bar{\lambda}_t \) and \( \bar{\lambda}_t \), respectively. This immediately implies that the limit likelihood \( L_T \) is twice continuously differentiable in the compact set \( K \). Therefore, a Taylor series expansion around \( \kappa_0 \) yields

\[
L_T (\tilde{\kappa}_T) = L_T (\kappa_0) + L_T ''(\eta_T) (\tilde{\kappa}_T - \kappa_0),
\]

where \( \eta_T \) is a point between \( \tilde{\kappa}_T \) and \( \kappa_0 \). By definition, \( \tilde{\kappa}_T \) is the maximizer of \( L_T (\kappa) \). Therefore, we have that \( L_T (\tilde{\kappa}_T) = 0 \) for sufficiently large \( T \) since \( \tilde{\kappa}_T \to \kappa_0 \), almost surely, and \( \kappa_0 \in \text{int}(K) \). As a result, the following equation holds:

\[
\sqrt{T} L_T '' (\kappa_0) = - L_T '' (\eta_T) \sqrt{T} (\tilde{\kappa}_T - \kappa_0).
\]

By lemma 3 together with an application of the ergodic theorem of Rao (1962), we obtain that \( - L_T '' (\eta_T) \to F_0 \), almost surely, where \( F_0 = - E \{ l_T '' (\kappa_0) \} \). Furthermore, lemma 5 ensures that \( F_0 \) is positive definite and lemma 4 shows that \( \sqrt{T} L_T '' (\kappa_0) \to d N(0, F_0) \). Therefore, we obtain that

\[
\sqrt{T} (\tilde{\kappa}_T - \kappa_0) = F_0^{-1} \sqrt{T} L_T '' (\kappa_0) + o_p(1),
\]

where \( o_p(1) \) represents a term that converges to zero in probability.
which implies that \( \sqrt{T} (\hat{\kappa}_T - \kappa_0) \to_d N(0, F_0^{-1}) \) as \( T \to \infty \).

We conclude the proof by showing that \( \sqrt{T} (\hat{\kappa}_T - \kappa_0) \to_d N(0, F_0^{-1}) \) entails that \( \sqrt{T} (\hat{\kappa}_T - \kappa_0) \to_d N(0, F_0^{-1}) \).

A Taylor series expansion yields

\[
L'_T(\hat{\kappa}_T) = L'_T(\hat{\kappa}_T) + L''_T(\hat{\eta}_T)(\hat{\kappa}_T - \hat{\kappa}_T),
\]

where \( \hat{\eta}_T \) is a point between \( \hat{\kappa}_T \) and \( \hat{\kappa}_T \). Furthermore, we note that \( L'_T (\hat{\kappa}_T) = 0 \) and \( L'_T (\hat{\kappa}_T) = 0 \) for sufficiently large \( T \) since the estimators are strongly consistent and \( \kappa_0 \in \text{int}(K) \). Therefore, we have that

\[
\sqrt{T} \{ L'_T (\hat{\kappa}_T) - L'_T (\hat{\kappa}_T) \} = L''_T(\hat{\eta}_T) T (\hat{\kappa}_T - \hat{\kappa}_T).
\]

The left-hand side of this equation goes to 0 almost surely as \( T \to \infty \) by an application of lemma 6. Furthermore, lemma 3 ensures that \( L''_T(\hat{\eta}_T) \to -F_0 \), almost surely. Therefore, we obtain that \( \sqrt{T} (\hat{\kappa}_T - \hat{\kappa}_T) \to 0 \), almost surely. This concludes the proof of theorem 4 since

\[
\sqrt{T} (\hat{\kappa}_T - \kappa_0) = \sqrt{T} (\hat{\kappa}_T - \kappa_0) + \sqrt{T} (\hat{\kappa}_T - \hat{\kappa}_T) = \sqrt{T} (\hat{\kappa}_T - \kappa_0) + o_p(1).
\]

### A.2. Lemmas

Below, we report the statements of the lemmas that were used in Appendix A.1. The proofs of the lemmas are given in the on-line supplementary material.

**Lemma 1.** Let the assumptions of theorem 4 hold. Then, \( \| \hat{\lambda}_T - \hat{\theta}_T \|_\Theta \to 0 \), exponentially almost surely, as \( t \to \infty \), where \( \hat{\theta}_T \) is the stationary and ergodic derivative process of \( \hat{\lambda}_T \). Furthermore, \( \hat{\theta}_T \) has a uniformly bounded second moment, i.e. \( \mathbb{E} \| \hat{\theta}_T \|_\Theta < \infty \).

**Lemma 2.** Let the assumptions of theorem 4 hold. Then, \( \| \hat{\lambda}_T - \hat{\theta}_T \|_\Theta \to 0 \), exponentially almost surely, as \( t \to \infty \), where \( \hat{\theta}_T \) is the stationary and ergodic second-derivative process of \( \hat{\lambda}_T \). Furthermore, \( \hat{\theta}_T \) has a uniformly bounded first moment, i.e. \( \mathbb{E} \| \hat{\theta}_T \|_\Theta < \infty \).

**Lemma 3.** Let the assumptions of theorem 4 hold. Then, the second derivative of the likelihood function has a uniformly bounded moment, i.e. \( \mathbb{E} \| \hat{\theta}_T \|_K < \infty \).

**Lemma 4.** Let the assumptions of theorem 4 hold. Then, \( \sqrt{T} L'_T (\kappa_0) \to_d N(0, F_0) \) as \( T \to \infty \).

**Lemma 5.** Let the assumptions of theorem 4 hold. Then, the Fisher information matrix is positive definite, i.e. \( F_0 = -\mathbb{E} \{ \hat{\theta}_T^T (\kappa_0) \} = \mathbb{E} \{ \hat{\theta}_T (\kappa_0) \hat{\theta}_T^T (\kappa_0) \} > 0 \).

**Lemma 6.** Let the assumptions of theorem 4 hold. Then, \( \sqrt{T} \| \hat{\theta}_T - L'_T \|_K \to 0 \), almost surely, as \( T \to \infty \).

#### A.2.1. Lemmas on the gamma function and the beta–negative binomial probability mass function

**Lemma 7.** The total variation distance between BNB_{\lambda_1} and BNB_{\lambda_2} satisfies the inequality

\[
\text{TV}(\text{BNB}_{\lambda_1}, \text{BNB}_{\lambda_2}) \leq 1 - \exp(-|\lambda_1 - \lambda_2|)
\]

for any \( \lambda_1, \lambda_2 \in \mathbb{R}^+ \) and any \( (r, \alpha) \in \mathbb{R}^+ \times (1, \infty) \).

**Lemma 8.** For any \( r, r_0 \in \mathbb{R}^+ \), any \( \alpha, \alpha_0 \in (1, \infty) \), and any positive random variables \( \lambda \) and \( \lambda_0 \) such that \( \lambda_0 \) is not degenerate,

\[
p(y|\lambda, r, \alpha) = p(y|\lambda_0, r_0, \alpha_0) \quad \text{almost surely} \quad \forall \ y \in \mathbb{N}
\]

if and only if \( (\lambda, r, \alpha) = (\lambda_0, r_0, \alpha_0) \) almost surely.

**Lemma 9.** Let \( x \) be a random variable such that \( x \geq \bar{x} \) with probability 1 for some constant \( \bar{x} > 0 \). Furthermore, assume that \( \mathbb{E}(x) < \infty \). Then, \( \mathbb{E}(\log \{ \Gamma(x + c) \} - \log \{ \Gamma(x) \}) < \infty \) for any \( c \in \mathbb{R}^+ \).

**Lemma 10.** The following inequalities are satisfied for any \( \lambda \in [\bar{\lambda}, \infty) \), \( y \in \mathbb{N} \) and \( \xi = (r, \alpha)^T \in \Xi \), where \( \Xi \) is a compact set, \( \Xi \subset (0, \infty) \times (2, \infty) \):
(a) \| \partial \log \{ p(y|\lambda, r, \alpha) \} / \partial \lambda \| \leq c_\lambda,
(b) \| \partial \log \{ p(y|\lambda, r, \alpha) \} / \partial r \| \leq c_r + a_r (y + \lambda),
(c) \| \partial \log \{ p(y|\lambda, r, \alpha) \} / \partial \alpha \| \leq c_\alpha + a_\alpha y,
(d) \| \partial^2 \log \{ p(y|\lambda, r, \alpha) \} / \partial \lambda^2 \| \leq c_{\lambda\lambda},
(e) \| \partial^2 \log \{ p(y|\lambda, r, \alpha) \} / \partial r^2 \| \leq c_r + a_r \lambda,
(f) \| \partial^2 \log \{ p(y|\lambda, r, \alpha) \} / \partial \alpha^2 \| \leq c_{\alpha\alpha} + a_{\alpha\alpha} y,
(g) \| \partial^2 \log \{ p(y|\lambda, r, \alpha) \} / \partial \lambda \partial r \| \leq c_{\lambda r},
(h) \| \partial^2 \log \{ p(y|\lambda, r, \alpha) \} / \partial \lambda \partial \alpha \| \leq c_{\lambda\alpha},
(i) \| \partial^2 \log \{ p(y|\lambda, r, \alpha) \} / \partial r \partial \alpha \| \leq c_{r\alpha} + a_{r\alpha} \lambda,

for some positive constants \( c_i, a_i, c_{ij} \) and \( a_{ij} \), with \( i, j \in \{ \lambda, r, \alpha \} \).

References

Ahmad, A. and Francq, C. (2016) Poisson QMLE of count time series models. *J. Time Ser. Anal.*, 37, 291–314.
Akrouche, A., Bendjeddou, S. and Touche, N. (2018) Negative binomial quasi-likelihood inference for general integer-valued time series models. *J. Time Ser. Anal.*, 39, 192–211.
Barczy, M., Ispány, M., Pap, G., Scotto, M. and Silva, M. E. (2010) Innovational outliers in INAR (1) models. *Commun Statist. Theory Meth.*, 39, 3343–3362.
Barczy, M., Ispány, M., Pap, G., Scotto, M. and Silva, M. E. (2012) Additive outliers in INAR (1) models. *Statist. Pap.*, 53, 935–949.
Blasques, F., Gorgi, P., Koopman, S. J. and Wintenberger, O. (2018) Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electron. J. Statist.*, 12, 1019–1052.
Bougerol, P. (1993) Kalman filtering with random coefficients and contractions. *SIAM J. Control Optimizn*, 31, 942–959.
Chen, H., Li, Q. and Zhu, F. (2019) Binomial AR (1) processes with innovational outliers. *Communs Statist. Theory Meth.*, to be published, doi 10.1080/03610926.2019.1635704.
Creal, D., Koopman, S. J. and Lucas, A. (2011) A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *J. Bus. Econ. Stat.*, 29, 552–563.
Creal, D., Koopman, S. J. and Lucas, A. (2013) Generalized autoregressive score models with applications. *J. Appl. Econometr.*, 28, 777–795.
Davis, R. A., Dunsmuir, W. T. and Streett, S. B. (2003) Observation-driven models for Poisson counts. *Biometrika*, 90, 777–790.
Davis, R. A., Holan, S. H., Lund, R. and Ravishanker, N. (2016) *Handbook of Discrete-valued Time Series*. Boca Raton: CRC Press.
Davis, R. A. and Liu, H. (2016) Theory and inference for a class of nonlinear models with application to time series of counts. *Statist. Sin.*, 46, 1673–1707.
Davis, R. A. and Wu, R. (2009) A negative binomial model for time series of counts. *Biometrika*, 96, 735–749.
Doukhan, P. and Neumann, M. H. (2019) Absolute regularity of semi-contractive GARCH-type processes. *J. Appl. Probab.*, 56, 91–115.
Embrechts, P., Klüppelberg, C. and Mikosch, T. (2013) *Modelling Extremal Events: for Insurance and Finance*. New York: Springer Science and Business Media.
Ferland, R., Latour, A. and Oraichi, D. (2006) Integer-valued GARCH process. *J. Time Ser. Anal.*, 27, 923–942.
Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009) Poisson autoregression. *J. Am. Statist. Ass.*, 104, 1430–1439.
Fox, A. J. (1972) Outliers in time series. *J. R. Statist. Soc. B*, 34, 350–363.
Geweke, J. and Amisano, G. (2011) Optimal prediction pools. *J. Econometr.*, 164, 130–141.
Gorgi, P. (2018) Integer-valued autoregressive models with survival probability driven by a stochastic recurrence equation. *J. Time Ser. Anal.*, 39, 150–171.
Hall, A. (2001) Extremes of integer-valued moving average models with regularly varying tails. *Extremes*, 4, 219–239.
Hall, A. (2003) Extremes of integer-valued moving average models with exponential type tails. *Extremes*, 6, 361–379.
Harvey, A. (2013) *Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series*. New York: Cambridge University Press.
Harvey, A. and Luati, A. (2014) Filtering with heavy tails. *J. Am. Statist. Ass.*, 109, 1112–1122.
Koopman, S. J., Lit, R. and Lucas, A. (2017) Intraday stochastic volatility in discrete price changes: the dynamic Skellam model. *J. Am. Statist. Ass.*, 112, 1490–1503.
McKenzie, E. (1986) Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv. Appl. Probab.*, 18, 679–705.
Opschoor, A., Janus, P., Lucas, A. and Van Dijk, D. (2017) New HEAVY models for fat-tailed realized covariances and returns. *J. Bus. Econ. Statist.*, 36, no. 4, 1–15.
Qian, L., Li, Q. and Zhu, F. (2020) Modelling heavy-tailedness in count time series. *Appl. Math. Modelling*, **82**, 766–784.

Rao, R. R. (1962) Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, **33**, 659–680.

Scotto, M. G., Weiβ, C. H., Möller, T. A. and Gouveia, S. (2018) The max-INAR (1) model for count processes. *TEST*, **27**, 850–870.

Silva, M. E., Pereira, I. and McCabe, B. (2019) Bayesian outlier detection in non-Gaussian autoregressive time series. *J. Time Ser. Anal.*, **40**, 631–648.

Straumann, D. and Mikosch, T. (2006) Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *Ann. Statist.*, **34**, 2449–2495.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.

Wang, C., Liu, H., Yao, J.-F., Davis, R. A. and Li, W. K. (2014) Self-excited threshold Poisson autoregression. *J. Am. Statist. Ass.*, **109**, 777–787.

Zhu, F. (2011) A negative binomial integer-valued GARCH model. *J. Time Ser. Anal.*, **32**, 54–67.

Zhu, F. (2012) Zero-inflated Poisson and negative binomial integer-valued GARCH models. *J. Statist. Planng Inf.*, **142**, 826–839.

Zhu, F., Liu, S. and Shi, L. (2016) Local influence analysis for Poisson autoregression with an application to stock transaction data. *Statist. Neerland.*, **70**, 4–25.

Zhu, F., Shi, L. and Liu, S. (2015) Influence diagnostics in log-linear integer-valued GARCH models. *Adv. Statist. Anal.*, **99**, 311–335.

**Supporting information**

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material of “BNB autoregressions for modeling integer-valued time series with extreme observations”’.

---

Qian, L., Li, Q. and Zhu, F. (2020) Modelling heavy-tailedness in count time series. *Appl. Math. Modelling*, **82**, 766–784.

Rao, R. R. (1962) Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, **33**, 659–680.

Scotto, M. G., Weiβ, C. H., Möller, T. A. and Gouveia, S. (2018) The max-INAR (1) model for count processes. *TEST*, **27**, 850–870.

Silva, M. E., Pereira, I. and McCabe, B. (2019) Bayesian outlier detection in non-Gaussian autoregressive time series. *J. Time Ser. Anal.*, **40**, 631–648.

Straumann, D. and Mikosch, T. (2006) Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *Ann. Statist.*, **34**, 2449–2495.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.

Wang, C., Liu, H., Yao, J.-F., Davis, R. A. and Li, W. K. (2014) Self-excited threshold Poisson autoregression. *J. Am. Statist. Ass.*, **109**, 777–787.

Zhu, F. (2011) A negative binomial integer-valued GARCH model. *J. Time Ser. Anal.*, **32**, 54–67.

Zhu, F. (2012) Zero-inflated Poisson and negative binomial integer-valued GARCH models. *J. Statist. Planng Inf.*, **142**, 826–839.

Zhu, F., Liu, S. and Shi, L. (2016) Local influence analysis for Poisson autoregression with an application to stock transaction data. *Statist. Neerland.*, **70**, 4–25.

Zhu, F., Shi, L. and Liu, S. (2015) Influence diagnostics in log-linear integer-valued GARCH models. *Adv. Statist. Anal.*, **99**, 311–335.

**Supporting information**

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material of “BNB autoregressions for modeling integer-valued time series with extreme observations”’.

---