Bayesian nonparametric forecasting for INAR models

Bayesian nonparametric forecasting for INAR models

Luisa Bisaglia[b], Antonio Canale[a,*]

[a] *Dept. of Economics and Statistics, University of Turin and Collegio Carlo Alberto, Italy*
[b] *Dept. of Statistical Sciences, University of Padua, Italy*

Abstract

A nonparametric Bayesian method for producing coherent predictions of count time series with the nonnegative integer-valued autoregressive process is introduced. Predictions are based on estimates of h-step-ahead predictive mass functions, assuming a nonparametric distribution for the innovation process. That is, the distribution of errors are modeled by means of a Dirichlet process mixture of rounded Gaussians. This class of prior has large support on the space and probability mass functions and can generate almost any kind of count distribution, including over/under-dispersion and multimodality. An efficient Gibbs sampler is developed for posterior computation, and the method is used to analyze a dataset of visits to a web site.

Keywords: Count time series, INAR(1), Dirichlet process mixtures, Forecasting, Gibbs sampling algorithm

1. Introduction

There has been recent growing interest in studying non-negative integer-valued time series and, in particular, time series of counts. Examples are categorical time series, binary processes, birth-death models and counting series as, for instance, the monthly number of active customers of a mobile phone service provider, daily number of traded stocks in a firm, daily number of visitors to a website, monthly incidence of a disease, and so on.

In some cases, the discrete values of the time series are large numbers and may be analyzed by using continuous-valued models such as ARMA models with Gaussian errors. However, according to Chatfield (2000), a good model

*Dept. of Economics and Statistics, University of Turin, Corso Unione Sovietica 218bis, 10134 Torino; Tel. +39 011 6705724

for time series should be consistent with the properties of the data and be unable to predict values which violate known constraints. This means that, in the case of counting data, we have to consider a model that is forecast-coherent and a method of forecasting that produces integer values. In the light of this requirement the usual linear ARMA processes are of limited use for modeling and especially for forecasting purposes.

The most common approach to build an integer-valued autoregressive process is by means of the thinning operator. An interpretation behind this probabilistic operation is to consider a random count as the size of a population, which is randomly reduced. Using binomial thinning, Al-Osh and Alzaid (1987) and McKenzie (1988) introduce INteger-valued AutoRegressive processes (INAR) for models with one lag. Although the theoretical properties of INAR models have been extensively studied in the literature (see, for instance, Freeland and McCabe (2004a), and the references therein), relatively few contributions discuss the development of forecasting methods which are coherent in producing only integer forecasts of the count variable. In the context of INAR(1) processes with Poisson innovations, Freeland and McCabe (2004b) suggested a method to produce optimal coherent forecasts based on the integer-valued median of the predictive distribution. Extensions taking into account higher-order dependence structure and overdispersion can be found in Jung and Tremayne (2006) and Schweer and Weiß (2014), respectively.

Since previous solutions are somewhat problem-specific, McCabe and Martin (2005) examined a method for producing coherent forecasts of low count time series from the Bayesian point of view. The predictive probability mass function, defined only over the support of the discrete count variable, is a natural outcome of Bayes theorem, and both parameter uncertainty and that due to the specification of the model are thus directly incorporated into the predictive probability mass function. The authors assumed that the model generating Y_t is any one within a set of K models $M_1, \ldots, M_K$. Thus, they define the $h-$step-ahead predictive probability mass function as:

$$Pr(Y_{T+h} = y_{T+h} \mid \mathbf{y}) = \sum_{k=1}^{K} Pr(Y_{T+h} = y_{T+h} \mid \mathbf{y}, M_k) Pr(M_k \mid \mathbf{y}), \quad (1)$$

where $\mathbf{y} = (y_1, \ldots, y_T)$ is the vector of the observed time series, $Pr(Y_{T+h} = y_{T+h} \mid \mathbf{y}, M_k)$ is the kth model-specific $h-$step-ahead predictive probability mass function and $Pr(M_k \mid \mathbf{y})$ is the posterior probability of model M_k.

In particular, the authors focused on three alternative error distributions, Poisson, binomial and negative binomial. The complexity of (1) does not allow us to work with it directly. In particular, its evaluation requires a numerical approach which depends on the nature of the models in the model set.

The approach of McCabe and Martin (2005) relies on parametric assumptions for the distribution of innovations, which may lead to misspecification errors if the true distribution is not included among the set of all possible distributions. From this point of view, an interesting proposal under the frequentist approach came from McCabe et al. (2011) in which efficient probabilistic forecasts are produced by treating the arrival process non-parametrically and proving the asymptotic (non-parametric) efficiency of the estimated forecast distribution.

With regard to the non-parametric estimation of the innovation process, Drost et al. (2009) consider a semiparametric INAR(p) model where there are essentially no restrictions on the innovation distribution. The authors provide a (semiparametrically) efficient estimator of both the auto-regression parameters and the innovation distribution, but they do not consider the problem of forecasting.

Our purpose is to develop a forecasting procedure within the context of these models which preserves the integer structure of the data. To this end we discuss, under a Bayesian paradigm, a nonparametric specifications of the error term. Assuming a nonparametric prior with large support for the innovation distribution bypasses the need to specify a finite set of discrete distributions for the innovations, as in McCabe and Martin (2005). This approach leads to two main improvements: first, we overcome the specification of the predictive probability as a mixture of K predictive distributions; and, second, we do not rely on the usual strict assumptions of standard parametric models. We concentrate on INAR(1) models, leaving the generalization to INAR(p) models for future work for two main reasons. First, this is (to our knowledge) the first attempt at applying a Bayesian nonparametric approach to forecasting count time series and so we want to focus on the innovation structure of the INAR models. Second, the extension for $p > 1$ is not unique and there are several proposal in this sense. See for example Jung and Tremayne (2011) and the references therein.

Different proposals to estimate count probability distributions have been introduced in the Bayesian nonparametric literature. Poisson mixtures are a natural choice (for a review, see Karlis and Xekalaki, 2005) which unfor-

tunately lack of flexibility. Indeed the Poisson kernel has a single parameter corresponding to both mean and variance and so a mixture of Poissons is not able to produce probability functions that are underdispersed. A mixture of multinomials can be seen as a first alternative, but this requires a bounded support for the count variable. An alternative nonparametric Bayes approach, avoids the mixture specification and directly uses the Dirichlet process (DP) prior (Ferguson, 1973, 1974). Exploiting the almost sure discreteness of the DP, Carota and Parmigiani (2002) propose to place it directly as prior for the count distribution. Despite the flexibility of the latter approach, there are some major disadvantages for both small and large sample sizes. Indeed, the posterior expectation of the nonparametric probability mass function is a mixture of the DP base measure and the empirical probability mass function with non-smooth deviations between neighboring integers. Canale and Dunson (2011) recently discussed a flexible prior for count distribution, with appealing properties in terms of large support and posterior consistency, which we use here.

The paper is organized as follows. In Section 2 we introduce our model, reviewing the structure of the INAR process and of the nonparametric prior introduced by Canale and Dunson (2011). In Section 3 a simulation study is conducted, and in Section 4 the proposed method is applied to a dataset on the daily number of visitors to the web site of the "statistical calendar" (`http://cal.stat.unipd.it/eng`), a students' project of the department of Statistics of the University of Padua, Italy. Section 5 concludes.

2. Model specification

2.1. INAR(1) model

Introduced by Al-Osh and Alzaid (1987) and McKenzie (1988), INAR(1) is suitable for counting processes in which one element of the process at time t may be either the survival of an element of the process at previous times, or the outcome of an innovation process with a certain discrete distribution. INARs provide a useful class of integer-valued processes for modeling time series, and their representation makes use of the thinning operator, '$\circ$', defined as follows (Steutel and Van Harn, 1979):

Definition 1. *If Y is a non-negative integer-valued random variable, then,*

for any $\alpha \in [0,1]$,

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

where X_i is a sequence of iid Bernoulli random variables, independent of Y, with success probability α.

This operator is the multiplication counterpart in the integer-valued context. Some interesting properties of the thinning operator are discussed in Silva (2005).

To find a maximum of parallelism with the AR(1) model, Al-Osh and Alzaid (1987) define the INAR(1) $\{Y_t; t \in \mathbb{Z}\}$ by the recursion

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t, \tag{2}$$

where $\alpha \in [0, 1)$, and ϵ_t is a sequence of iid positive discrete random variables with finite first and second moments, $\mu_\epsilon > 0$ and σ_ϵ^2 respectively.

The components of process $\{Y_t\}$ are the surviving elements of process Y_{t-1} during period $(t-1, t]$, and the number of elements which entered the system in the same interval, ϵ_t. Each element of Y_{t-1} survives with probability α, and its survival has no effect on that of other elements, nor on ϵ_t. Observe that, given $Y = y$, the random variable $\alpha \circ Y$ follows the binomial distribution with parameters y and α.

The marginal distribution of model (2) may be expressed in terms of arrival process ϵ_t as

$$Y_t \stackrel{d}{=} \sum_{j=0}^{\infty} \alpha^j \circ \epsilon_{t-j}.$$

The unconditional moments (see, for example, Jung et al., 2005) are $\mathbb{E}(Y_t) = \mu_\epsilon/(1-\alpha)$ and $\mathbb{V}ar(Y_t) = (\alpha\mu_\epsilon + \sigma_\epsilon^2)/(1-\alpha^2)$. The autocorrelation function of the process is $\rho(k) = \alpha^k, \quad k = 1, 2, \ldots$, but only positive autocorrelation is allowed.

Usual distributional assumptions for the error term include Poisson, binomial, negative binomial and geometric. If the error terms are distributed as a Poisson distribution with mean parameter λ, then it can be shown that the marginal distribution of the observed counts is a Poisson distribution with (unconditional) mean and variance equal to $\lambda/(1-\alpha)$. In this case, therefore, the model cannot take into account over/under-dispersion in data.

The negative binomial allows for over-dispersion, but assumes a truncated support for innovations.

The distribution of Y_t, given y_{t-1}, α, and innovation distribution p, is

$$Pr(Y_t = y_t \mid y_{t-1}, \alpha, p) = \sum_{s=0}^{\min\{y_t, y_{t-1}\}} Pr(B^{\alpha}_{y_{t-1}} = s) \times p(y_t - s)$$

where $B^{\pi}_k \sim \text{Bin}(k, \pi)$. The likelihood function of $\theta = (\alpha, p)$, given a sample $\mathbf{y} = (y_1, \ldots, y_T)$ of size T, is

$$L(\theta \mid \mathbf{y}) = \prod_{t=2}^{T} \sum_{s=0}^{\min\{y_t, y_{t-1}\}} Pr(B^{\alpha}_{y-1} = s) \times p(y_t - s), \tag{3}$$

where $\theta \in \Theta$ and $\Theta = (0, 1) \times \mathcal{C}$, with $\mathcal{C}$ the space of probability mass functions on non-negative integers.

From a Bayesian perspective, we must specify a prior distribution for θ. In the following, we assume independent prior distributions for $\alpha \sim \pi_\alpha$ and $p \sim \Pi$, leading to prior $\pi(\theta) = \Pi \times \pi_\alpha$. With this formulation, the posterior h-step-ahead probability mass function for $j \in \mathbb{N}$ is

$$Pr(Y_{T+h} = j \mid \mathbf{y}) = \int_{\Theta} Pr(Y_{T+h} = j \mid \mathbf{y}, \theta) \mathrm{d}\pi(\theta \mid \mathbf{y}), \tag{4}$$

where $\pi(\theta \mid \mathbf{y})$ is the posterior distribution of the parameters given the data.

2.2. Rounded mixture priors

To define a nonparametric model for counts, Canale and Dunson (2011) rounded an underlying variable having an unknown density, given a Dirichlet process mixture of Gaussians prior, (Lo, 1984; Escobar and West, 1995). This rounded mixture of Gaussians (RMG) is not only highly flexible and with excellent performance in small samples, but also has appealing asymptotic properties in terms of large support and strong posterior consistency. The choice of modeling the distribution of errors ϵ_t with such a prior allows us to leverage on a model with large support, robust for any model misspecification while eliminating the need to average over several models, as proposed by McCabe and Martin (2005).

For each $j \in \mathbb{N}$, let

$$p(j) = p(\epsilon_t = j) = g(f)[j] = \int_{a_j}^{a_{j+1}} f(\epsilon^*) \mathrm{d}\epsilon^*, \tag{5}$$

where f is a continuous density, ϵ^* a latent continuous variable, and $a_0 = -\infty$ and $a_j = j - 1$ for $j \in \{1, 2, \dots\}$. We then model the underlying f via the nonparametric mixture model

$$f(\epsilon^*; P) = \int \phi(\epsilon^*; \mu, \sigma^2) \mathrm{d}P(\mu, \sigma^2), \quad P \sim DP(\eta P_0), \tag{6}$$

where $\phi(\cdot; \mu, \sigma^2)$ is a Gaussian density with mean μ and variance σ^2. To the mixing random measure P, a DP prior with precision parameter η and base measure P_0 is assigned. The DP is a probability distribution on the space of probability measures and has several characterizations. For sake of brevity and to directly understand its usefulness in equation (6), we describe the so called stick-breaking construction (Sethuraman, 1994), in which the random mixing measure P can be written as

$$P = \sum_{l=1}^{\infty} \pi_l \delta_{\theta_l}, \;\; \theta \overset{iid}{\sim} P_0,$$

and $\pi_1 = V_1$, $\pi_l = V_l \prod_{r<l}(1 - V_r)$ with $V_l \sim \text{beta}(1, \eta)$. The mechanism that generates the weights gives the name to this representation since it may be thought as breaking a stick of length one into infinitely many pieces with length proportional to the sequence of weight. In our model $\theta_l = (\mu_l, \sigma_l^2)$ and P_0 is chosen to be Normal-Gamma, i.e.

$$\sigma_l^{-2} \overset{iid}{\sim} \text{Ga}(a, b), \quad \mu_l \overset{iid}{\sim} N(\mu_0, \kappa\sigma_l^2).$$

Note that equation (5) defines mapping function $g(\cdot)$ between the spaces of continuous densities and of probability mass functions. A related rounding function, $r : \mathbb{R} \to \mathbb{N}$, is such that $r(\epsilon^*) = j$ if $a_j \leq \epsilon^* < a_{j+1}$. As default hyperparameter choice, when no prior information is available, we can let $a_0 = -\infty$ and $a_j = j$ for $j = 1, \dots,$ and let $a = b = 1/2$ and, with an empirical Bayes approach $\mu_0 = \bar{y}$ and $\kappa = s^2$, the sample mean and variance respectively. Equations (5)–(6) induce a prior $p \sim \Pi$ over the space of probability mass functions on the integers.

2.3. Computation by MCMC

Since the joint posterior distribution is not in closed form, we rely on Markov Chain Montecarlo (MCMC) simulation from the posterior distribution. In particular we resort to an iterative Gibbs sampler. In a first data

augmentation step we simulate the latent survivor and birth processes. Conditionally on the latent underlying continuous innovations ϵ_t^*, simulated under the constraints that $r(\epsilon_t^*) = \epsilon_t$, one can resort to the algorithm described in Canale and Dunson (2011). Specifically, we introduce latent $S_1, \ldots, S_T$ where $S_t = l$ if the t-th innovation is drawn from the l-th mixture component. With such an approach, conditionally on S_t, each ϵ_t^* is drawn from a single normal distribution and hence the updated of each cluster-specific set of parameters can be done easily. The conditional posterior distribution of α is not in closed form, so we rely on a Metropolis-Hastings substep. As proposal density we choose a $\text{Be}(1, (1-\alpha_{last})/\alpha_{last})$, where α_{last} is the last value of the Markov chain for the thinning parameter. With this choice, the expectation of the proposal density is centered on the last available value of the Markov chain. The Gibbs sampler is summarized below.

1. Data augmentation step given p and α :
   (a) simulate $\mathbf{B} = \{B_2, \ldots, B_T\}$ where each B_t has multinomial distribution with cell probability $P(B_t = j) \propto \binom{y_{t-1}}{j} \alpha^j (1-\alpha)^{y_{t-1}-j} \times p(y_t - j)$ for $j = 0, \ldots, y_t$;
   (b) for $t = 2, \ldots, T$, simulate $\epsilon_t^* \sim f$ where f is as in (5)–(6) under constraints $a_{y_t - B_t} \leq \epsilon_t^* \leq a_{y_t - B_t + 1}$.
2. Update p with the Gibbs sampler described in Canale and Dunson (2011), namely:
   (a) sample S_t from a multinomial with cell probabilities equal to
   $$\Pr(S_t = l | \epsilon_t^*) \propto \pi_l \phi(\epsilon_t^*; \mu_l, \sigma_l^2);$$
   (b) update the stick-breaking weights using
   $$V_l \sim \text{Be}\left(1 + n_l, \eta + \sum_{r > l+1} n_r\right),$$
   where n_l is the sample size of the l-th cluster;
   (c) sample (μ_l, σ_l^2) from
   $$N\left(\hat{\mu}_l, \hat{\kappa}_l \sigma_l^2\right) \text{InvGam}(a + n_l/2 + 1, b + \hat{b}_l),$$
   where
   $$\hat{\mu}_l = \hat{\kappa}_l \left(\kappa \mu_0 + n_l \bar{\epsilon_l^*}\right), \ \hat{\kappa}_l = (\kappa^{-1} + n_l)^{-1},$$
   $$\hat{b}_l = \frac{1}{2}\left\{\sum_{S_t = l} (\epsilon_t^* - \bar{\epsilon_l^*})^2 + n_l/(1 + \kappa n_l)(\bar{\epsilon_l^*} - \mu_0)^2\right\},$$

and $\bar{\epsilon}_l^*$ is the sample mean in the l-th cluster.

3. Update α with a Metropolis-Hastings step from its conditional posterior distribution $\pi_\alpha(\alpha|\mathbf{y}, \mathbf{B}, p) \propto \pi_\alpha(\alpha) \times L(\alpha|\mathbf{y}, \mathbf{B}, p)$, i.e.
   (a) generate $\alpha^* \sim \text{Be}(1, (1-\alpha_{last})/\alpha_{last})$, where α_{last} is the last available value of the Markov chain;
   (b) accept α^* as next value with probability equal to

$$\min\left\{1, \frac{\pi(\alpha^*|\mathbf{y}, \mathbf{B}, p)}{\pi(\alpha_{last}|\mathbf{y}, \mathbf{B}, p)} \times \frac{\text{Be}(\alpha_{last}; 1, (1-\alpha^*)/\alpha^*)}{\text{Be}(\alpha^*; 1, (1-\alpha_{last})/\alpha_{last})}\right\},$$

   otherwise keep α_{last} as next value.

3. Simulation study: results and discussion

To evaluate the performance of the proposed method in obtaining coherent predictions for h-step-ahead forecasts we conducted a Monte Carlo experiment. We simulated $R = 500$ independent realizations of size $n = 50$, $n = 100$, and $n = 250$ from a wide variety of scenarios under the INAR(1) model. We chose $\alpha = 0.25, 0.5, 0.9$ for the thinning parameter, $h = 1, 2, 3, 4$ for the number of steps, and various discrete distributions for the arrival process, namely

i. Poisson, with mean parameter $\lambda = 4$;
ii. Negative Binomial, with $k = 6,\ p = 0.4$;
iii. Binomial, with $k = 15,\ p = 0.8$;
iv. Conway-Maxwell-Poisson distribution, with $\lambda = 30$ and $\nu = 3$.

Each independent sample was generated with 200 additional burn-in values to obtain random starting values.

For each scenario we run our Gibbs sampling algorithm with default hyperparameters choice as described in Section 2.3. We assess the predictive performance of our method by means of scoring rules for predictive distribution (Gneiting and Raftery, 2007). Scoring rules assess the quality of a probabilistic forecasts, by assigning a numerical score considering the estimated predictive distribution and the actual observed value. In particular, since we are dealing with discrete values, we computed the quadratic loss functions as

$$S(\hat{p}_h, y_{T+h}) = 2\hat{p}_h(y_{t+h})^2 - \sum_j \hat{p}_h(j)^2 - 1, \tag{7}$$

(a) (b) (c) (d)

Figure 1: Quadratic scoring rules in function of h for $\alpha = 0.5$ (continuous line), $\alpha = 0.75$ (dashed line), and $\alpha = 0.9$ (dotted line), under Scenario i (a), Scenario ii (b), Scenario iii (c), and Scenario iv (d), for $n = 50$.

where $\hat{p}_h$ is the posterior mean of the h-step-ahead probability mass function obtained with our approach, and the sum in j is taken across the range of the observed data $\pm$ a buffer of 20. Table 1 reports the results. The expected increase of the quality of the predictive performance for increasing sample size, is only mild. The performance of the method is indeed similar for all the sample sizes. This means that the proposed approach can be utilized also for medium-short time series. To better perceive the performance of our procedure in function of the number of steps ahead and the value of α, Figure 1 reports a graphical representation of the results for $n = 50$ only. Poorer performance is obtained in the scenarios involving a thinning parameter of $\alpha = 0.9$. This is probably due to the fact that the value of the parameters is close to the upper boundary of the domain and thus very close to the situation of non stationarity of the time series. We were expecting a general decrease of the performance as h increases. In fact, only a mild decrease of the performance for increasing h is evident particularly for the cases in which $\alpha = 0.9$. For $\alpha = 0.5$ and $\alpha = 0.75$, the predictive performance remain stable. Note that qualitatively similar results were obtained also using a spherical scoring rules, which values are reported in the Supplementary Material, see Appendix A.

Table 2 shows a Monte Carlo approximation to the mean Bhattacharya coefficient distance (BC) and Kullback-Leibler divergence (KL) between the posterior mean probability mass function of the arrivals ($\hat{p}$) and the true distribution (p_0). The BC and KL are calculated as

$$\text{BC} = \sum_j -\log\left(\sqrt{p_0(j)\hat{p}(j)}\right), \ \text{KL} = \sum_j p_0(j)\log\left(p_0(j)/\hat{p}(j)\right),$$

where, again, the sums in j are taken across the range of the observed data $\pm$ a buffer of 20, $\hat{p}$ is the estimated h-step-ahead probability mass function and p_0 is the true predictive probability mass function. The scenarios involving the Poisson and the Conway-Maxwell-Poisson are those where our

Table 1: Quadratic score of h step-ahead predictive distribution

n	α	$p(\epsilon)$	$h=1$	$h=2$	$h=3$	$h=4$
50	0.25	Pois	-0.892	-0.883	-0.887	-0.886
		NB	-0.950	-0.944	-0.946	-0.942
		Bin	-0.907	-0.900	-0.902	-0.899
		CMP	-0.806	-0.810	-0.811	-0.806
	0.50	Pois	-0.912	-0.905	-0.908	-0.903
		NB	-0.955	-0.956	-0.956	-0.953
		Bin	-0.925	-0.924	-0.925	-0.928
		CMP	-0.856	-0.852	-0.862	-0.855
	0.90	Pois	-0.937	-0.944	-0.950	-0.950
		NB	-0.968	-0.970	-0.972	-0.973
		Bin	-0.944	-0.951	-0.956	-0.962
		CMP	-0.905	-0.916	-0.925	-0.928
100	0.25	Pois	-0.887	-0.880	-0.886	-0.888
		NB	-0.946	-0.943	-0.944	-0.941
		Bin	-0.895	-0.888	-0.893	-0.889
		CMP	-0.798	-0.809	-0.809	-0.801
	0.50	Pois	-0.907	-0.903	-0.906	-0.905
		NB	-0.951	-0.954	-0.954	-0.953
		Bin	-0.917	-0.922	-0.923	-0.924
		CMP	-0.846	-0.852	-0.861	-0.855
	0.90	Pois	-0.928	-0.940	-0.947	-0.949
		NB	-0.964	-0.969	-0.971	-0.973
		Bin	-0.953	-0.961	-0.964	-0.966
		CMP	-0.895	-0.912	-0.923	-0.928
250	0.25	Pois	-0.885	-0.880	-0.885	-0.887
		NB	-0.948	-0.944	-0.944	-0.941
		Bin	-0.891	-0.886	-0.890	-0.886
		CMP	-0.794	-0.805	-0.805	-0.796
	0.50	Pois	-0.906	-0.902	-0.905	-0.904
		NB	-0.950	-0.953	-0.953	-0.952
		Bin	-0.918	-0.922	-0.923	-0.923
		CMP	-0.845	-0.850	-0.859	-0.853
	0.90	Pois	-0.926	-0.938	-0.945	-0.948
		NB	-0.967	-0.973	-0.975	-0.975
		Bin	-0.956	-0.962	-0.964	-0.966
		CMP	-0.891	-0.909	-0.920	-0.924

Table 2: Kullback-Leibler divergence, Bhattacharya coefficient

α	$p(\epsilon)$	$n = 50$		$n = 100$		$n = 250$	
		KL	BC	KL	BC	KL	BC
0.25	Pois	1.25	0.51	0.59	0.28	0.39	0.19
	NB	1.73	0.68	0.71	0.32	0.52	0.23
	Bin	14.40	3.60	1.79	0.82	1.52	0.67
	CMP	1.87	0.67	0.73	0.35	0.50	0.24
0.50	Pois	1.43	0.60	0.75	0.35	0.47	0.22
	NB	2.22	0.88	0.82	0.38	0.03	0.01
	Bin	11.41	2.99	4.65	2.07	2.96	1.31
	CMP	1.61	0.63	0.95	0.44	0.68	0.31
0.90	Pois	5.07	2.27	2.61	1.15	1.94	0.82
	NB	8.18	3.08	7.29	2.41	6.88	2.03
	Bin	19.20	8.71	19.37	9.26	23.63	11.49
	CMP	4.25	1.99	2.46	1.14	1.48	0.65

nonparametric prior registers the better performance: both the Bhattacharya coefficient and Kullback-Leibler divergence are constantly smaller than those obtained under the binomial and negative binomial scenario. Also in estimating the innovation probability mass functions the worst performance are obtained when the true value of α equals 0.9. This is consistent with the results in terms of goodness of prediction. To conclude, the estimation of the innovation probability mass function tends to be better for increasing n. Indeed, our prior is posterior consistent, i.e. the posterior probability of a neighborhood of the true data generating process (the probability mass function of the innovations) goes to one as $n \to \infty$ (see Canale and Dunson, 2011, Section 2.3, for more details). This finite sample behavior confirm this asymptotic property.

4. Application to website traffic data

We apply the proposed method to a real time series of the daily count of visits to the web site of the "statistical calendar" (Durante et al., 2012), a students' project of the department of Statistics of the University of Padua, Italy. The data used go from the 1st of April, 2012 to the first week of September, 2012. The site is the output of a contest sponsored by the Italian Statistical Society on the the topic "Statistics and statisticians: ideas to foster and spread the statistical culture".

The reason for a h-step-ahead prediction lies in promoting the site, after its launch at the end of 2011. Knowing the whole predictive probability mass function allows the webmasters to compute both the median forecast for the following day and the probability of having less than k visits, when

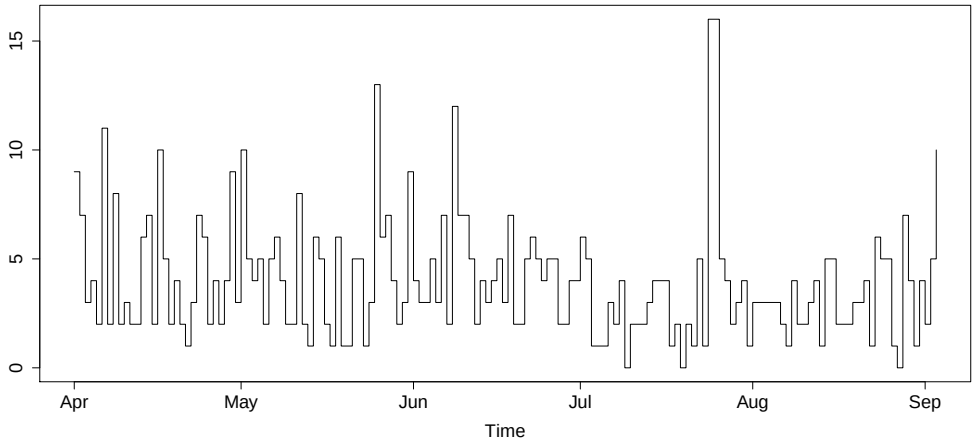

Figure 2: Time series of the daily number of visit to the "statistical calendar" website.

k is a lower bound, in order to have a fairly large number of visitors. The total number of visitors for today is typically made up of loyal visitors from yesterday and new visitors. Thus, the birth-death interpretation of the INAR model fits this particular dataset. It is important to underline, however, that the birth-and-death interpretation is just an intuitive interpretation and it is not critical if one is interested in prediction rather than in estimate and interpret the birth or death sub-components.

The dataset consists of 156 daily counts, from 0 to 16, but we remove the last 6 observations to perform out-of-sample forecasts. A plot of the series is shown in Figure 2. The series has median 3, mean 3.96 and mode 2, more than 98% of counts are less than 12, and variance is 7.81, meaning that the data are over-dispersed.

We run the Gibbs sampler for 17,000 iterations, discarding the first 2,000 for burn-in. The values of the predictive probability mass function $Pr(Y_{T+h} = j \mid \mathbf{y})$ for a wide variety of js are monitored to asses the convergence of the Markov chains. The trace plots show excellent mixing and Geweke (1992) diagnostics (as implemented in the R package coda) typically do not reject the hypothesis of the equality of the means of the first and last part of a Markov chains. Additional plots and comments on convergence and mixing are reported in the Supplementary Material, see Appendix A.

The predictive performance in terms of quadratic scores are -0.761,

-0.914, -0.816, -0.767, -0.918, and -1.101 for $h = 1, \ldots, 6$, respectively. As in the simulation study the predictive performance are stable for increasing h, and only a mild decrease of the performance can be noted. Figure 3 shows the posterior mean h-step-ahead predictive probability, $Pr(Y_{T+h} = j \mid \mathbf{y})$ for $j = 0, 1, \ldots, 20$, together with its 95% credible bands. These intervals are estimated as the 2.5th-97.5th percentiles of the MCMC samples collected after burn-in.

Figure 4 shows the posterior predictive probability of having a number of visits less than the median, i.e. $Pr(Y_{t+h} \leq 2)$ together with its 95% credible bands for $h = 1, \ldots, 6$. As before, the intervals are obtained considering the 2.5th-97.5th percentiles of the MCMC samples. The posterior mean of α turns out to be 0.23, with posterior probability of 95% between 0.07 and 0.34, which clearly indicates short-run dependence.

5. Conclusions

A Bayesian nonparametric method for producing coherent forecasts of count time series has been presented. Introducing a nonparametric distribution for the error term has several advantages. First of all, the lack of robustness implicitly present in specifying a particular family of distributions is overcome. In addition, the large support of our prior overcomes the usual parametric assumptions for counts and the approach of McCabe and Martin (2005) of defining a set of possible error distributions is bypassed. Bayesian reasoning is also appealing in that it allows us to use prior information, if available (e.g., the probability mass function of the errors is concentrated for small values, or α is centered on a given value *a priori*) and easy, reliable MCMC implementation is possible so that general h-step-ahead predictions can be made without tedious calculations.

Acknowledgments

Comments and suggestions by two anonymous referees are gratefully acknowledged. The authors thank Bruno Scarpa for generously providing the data analyzed in Section 4. This research was partially supported by the University of Padua CPDA097208/09 grant.

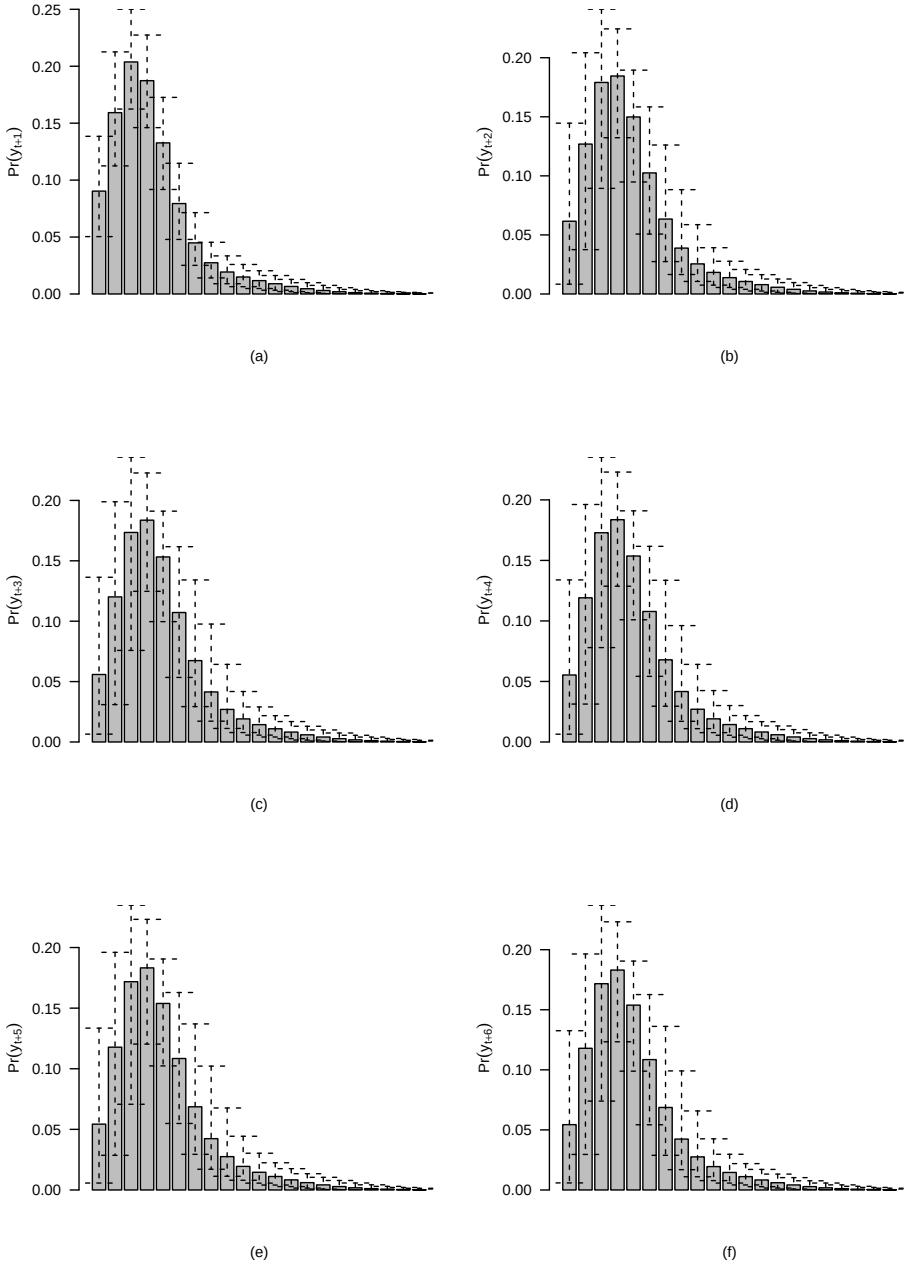

Figure 3: Posterior means of predictive probability mass function and 95% posterior credible bands for h-step-ahead and $h = 1$ (a), $h = 2$ (b), $h = 3$ (c), $h = 4$ (d), $h = 5$ (e), and $h = 6$ (f).

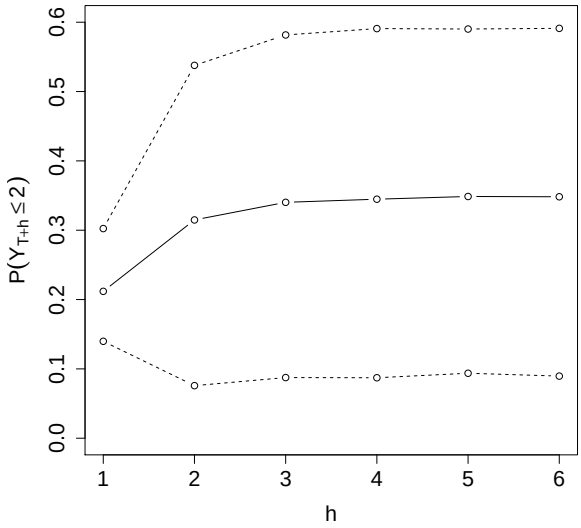

Figure 4: Posterior mean probability of having fewer than 2 counts (continuous lines) with 95% credible bands (dashed lines).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at [INSERT DOI HERE]

References

Al-Osh, M. A., Alzaid, A. A., 1987. First order integer-valued autoregressive INAR(1) process. Journal of Time Series Analysis 8 (3), 261–275.

Canale, A., Dunson, D. B., 2011. Bayesian kernel mixtures for counts. Journal of the American Statistical Association 106, 1528–1539.

Carota, C., Parmigiani, G., 2002. Semiparametric regression for count data. Biometrika 89 (2), 265–281.

Chatfield, C., 2000. Time-series forecasting. Chapman & Hall.

Drost, F., van den Akker R., B.J.M., W., 2009. Efficient estimation of auto-regression parameters and innovation distributions for semiparamet-

ric integer-valued AR(p) models. Journal of the Royal Statistical Society B 71, 467–485.

Durante, D., Vettori, S., Vidotto, D., 2012. The statistical calendar. `http://cal.stat.unipd.it/eng`.

Escobar, M. D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Ferguson, T. S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.

Ferguson, T. S., 1974. Prior distributions on spaces of probability measures. The Annals of Statistics 2, 615–629.

Freeland, R. K., McCabe, B. P. M., 2004a. Analysis of low count time series data by poisson autoregression. Journal of Time Series Analysis 25, 701–722.

Freeland, R. K., McCabe, B. P. M., 2004b. Forecasting discrete valued low count time series. International Journal of Forecasting 20, 427–434.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), Bayesian Statistics 4. Oxford: Oxford University Press.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Jung, R. C., G., R., Tremayne, A. R., 2005. Estimation in conditional first order autoregression with discrete support. Statistical Papers 46, 195–224.

Jung, R. C., Tremayne, A. R., 2006. Coherent forecasting in integer time series models. International Journal of Forecasting 22, 223–238.

Jung, R. C., Tremayne, A. R., 2011. Convolution-closed models for count time series with applications. Journal of Time Series Analysis 32, 268–280.

Karlis, D., Xekalaki, E., 2005. Mixed Poisson distributions. International Statistical Review 73 (1), 35–58.

Lo, A. Y., 1984. On a class of Bayesian nonparametric estimates: I. Density estimates. The Annals of Statistics 12, 351–357.

McCabe, B. P. M., Martin, G. M., 2005. Bayesian predictions of low count time series. International Journal of Forecasting 21, 315–330.

McCabe, B. P. M., Martin, G. M., Harris, D., 2011. Efficient probabilistic forecasts for counts. Journal of the Royal Statistical Society 73, 253–272.

McKenzie, E., 1988. Some ARMA models for dependent sequences of poisson counts. Advances in Applied Probability 20, 822–835.

Schweer, S., Weiß, C. H., 2014. Compound Poisson INAR(1) processes: Stochastic properties and testing for overdispersion. Computational Statistics and Data Analysis 77, 267 – 284.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.

Silva, I., 2005. Contribution to the analysis of discrete-valued time series. Ph.D. thesis, Universidade do Porto, Portugal.

Steutel, F. W., Van Harn, K., 1979. Discrete analogues of self-decomposability and stability. Annals of Probability 7, 893–899.