Bayesian generalizations of the integer-valued autoregressive model

Helton Graziadei
USP - São Paulo

Hedibert F. Lopes
Insper - São Paulo

Paulo C. Marques F.
Insper - São Paulo

October 2019

Abstract

We develop two Bayesian generalizations of the Poisson integer-valued autoregressive model. The AdINAR(1) model accounts for overdispersed data by means of an innovation process whose marginal distributions are finite mixtures, while the DP-INAR(1) model is a hierarchical extension involving a Dirichlet process, which is capable of modeling a latent pattern of heterogeneity in the distribution of the innovation rates. The probabilistic forecasting capabilities of both models are put to test in the analysis of crime data in Pittsburgh, with favorable results.

1 Introduction

Integer-valued time series models are essential inferential tools in areas such as epidemiology, econometrics, environmental studies, and public policy [1]. An important pioneering development was the proposal of the integer-valued autoregressive model (INAR(1) model hereafter) by McKenzie [2] and Al-Osh and Alzaid [3]. In a nutshell, the INAR(1) model defines a time-homogeneous Markovian process for which the counts at one epoch are a random fraction of the counts at the previous epoch plus the counts associated with an independent innovation process with marginal Poisson distributions. In recent decades, a large body of research has been dedicated to modifications and generalizations of this original INAR(1) model [4, 5, 6, 7, 8].

In this paper we develop two Bayesian generalizations of the INAR(1) model. In our first proposal, the adaptive integer-valued autoregressive model (AdINAR(1) model hereafter) we make use of a finite mixture to define the marginal distributions of the innovation process. The motivation for this AdINAR(1) model is the possibility to account for overdispersed time series. Our second contribution is a hierarchical extension of the INAR(1) model, implemented with the help of a Dirichlet process [9] placed at the top of the model hierarchy. In this DP-INAR(1) model, the innovation rates may vary through time, and in their modeling we benefit from the clustering properties of the

Dirichlet process. Our main goal in the paper is to assess the forecasting capabilities of the two new models when compared to the original INAR(1) model, regarding out-of-sample predictions.

The paper is organized as follows. In Section 2, we present a slightly generalized version of the INAR(1) model, on which the two models defined in the following sections are based. The AdINAR(1) and the DP-INAR(1) models are developed in Sections 3 and 4, respectively. Necessary facts about the Dirichlet process are presented in the Appendix. For both models we devise data augmentation schemes that result in full conditional distributions in simple analytical forms, enabling the exploration of the posterior distributions through Gibbs sampling [10, 11]. In Section 5, we put the models to work in the forecasting of crime data in Pittsburgh, USA. In this application, the AdINAR(1) and DP-INAR(1) model outperform the original INAR(1) model in the majority of the patrol areas. Computer simulations are coded in C++ inside the R environment [12], using the Rcpp library [13]. Computer code and data are available as supplementary materials.

2 Generalized INAR(1) model

We begin by generalizing the original INAR(1) model of McKenzie [2] and Al-Osh and Alzaid [3] as follows.

Let $\{Y_t\}_{t\geq 1}$ be an integer-valued time series, and let the *innovations* $\{Z_t\}_{t\geq 2}$, given positive parameters $\{\lambda_t\}_{t\geq 2}$, be a sequence of conditionally independent random variables. Given a parameter $\alpha \in [0, 1]$, let $\{B_i(t) : i \geq 0, t \geq 2\}$ be a family of conditionally independent and identically distributed Bernoulli(α) random variables. Furthermore, given all the parameters, assume that the innovations $\{Z_t\}_{t\geq 2}$ and the family $\{B_i(t) : i \geq 0, t \geq 2\}$ are conditionally independent. The generalized INAR(1) model is defined by the functional relation

$$Y_t = \alpha \circ Y_{t-1} + Z_t,$$

for $t \geq 2$, in which $\circ$ denotes the binomial thinning operator, defined by $\alpha \circ Y_{t-1} = \sum_{i=1}^{Y_{t-1}} B_i(t)$, if $Y_{t-1} > 0$, and $\alpha \circ Y_{t-1} = 0$, if $Y_{t-1} = 0$. In the homogeneous case, when $\lambda_2 = \dots \lambda_T =: \lambda$, and Z_t has Poisson(λ) distribution, given λ, we recover the original INAR(1) model.

This model can be interpreted as specifying a birth-and-death process, in which, at epoch t, the number of cases Y_t is equal to the new cases Z_t plus the cases that survived from the previous epoch; the role of the binomial thinning operator being to remove a random number of the Y_{t-1} cases present at the previous epoch $t - 1$.

Let $y = (y_1, \dots, y_T)$ denote the values of an observed time series. For simplicity, we assume that $Y_1 = y_1$ with probability one. Since the process $\{Y_t\}_{t\geq 1}$ is Markovian, the joint distribution of $Y_1, \dots, Y_T$, given parameters α and $\lambda = (\lambda_2, \dots, \lambda_T)$, can be factored as

$$\Pr\{Y_1 = y_1, \dots, Y_T = y_T \mid \alpha, \lambda\} = \prod_{t=2}^{T} \Pr\{Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha, \lambda_t\}.$$

Since, with probability one, $\alpha \circ Y_{t-1} \leq Y_{t-1}$ and $Z_t \geq 0$, by the law of total probability and the definition of the generalized INAR(1) model we have that

$$\begin{aligned}
\Pr\{Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha, \lambda_t\} &= \Pr\{\alpha \circ Y_{t-1} + Z_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha, \lambda_t\} \\
&= \Pr\left\{ \sum_{i=1}^{Y_{t-1}} B_i(t) + Z_t = y_t \;\middle|\; Y_{t-1} = y_{t-1}, \alpha, \lambda_t \right\} \\
&= \sum_{m_t=0}^{\min\{y_t, y_{t-1}\}} \Pr\left\{ \sum_{i=1}^{y_{t-1}} B_i(t) = m_t, Z_t = y_t - m_t \;\middle|\; \alpha, \lambda_t \right\} \\
&= \sum_{m_t=0}^{\min\{y_t, y_{t-1}\}} \Pr\left\{ \sum_{i=1}^{y_{t-1}} B_i(t) = m_t \;\middle|\; \alpha \right\} \Pr\{Z_t = y_t - m_t \mid \lambda_t\}.
\end{aligned}$$

Hence, the generalized INAR(1) model likelihood function is given by

$$L_y(\alpha, \lambda) = \prod_{t=2}^{T} \sum_{m_t=0}^{\min\{y_{t-1}, y_t\}} \binom{y_{t-1}}{m_t} \alpha^{m_t} (1-\alpha)^{y_{t-1}-m_t} \times \Pr\{Z_t = y_t - m_t \mid \lambda_t\}.$$

3 AdINAR(1) model

The AdINAR(1) model is defined assuming that $\lambda_2 = \ldots \lambda_T =: \lambda$, and that, given λ and two additional parameters $0 < \theta \leq 1$ and $0 \leq w \leq 1$, the innovations Z_t are conditionally independent and identically distributed as the two component mixture $w \times \text{Geometric}(\theta) + (1-w) \times \text{Poisson}(\lambda)$. Therefore, using the results in Section 2, the AdINAR(1) model likelihood function is given by

$$L_y(\alpha, \theta, \lambda, w) = \\ \prod_{t=2}^{T} \sum_{m_t=0}^{\min\{y_{t-1}, y_t\}} \binom{y_{t-1}}{m_t} \alpha^{m_t} (1-\alpha)^{y_{t-1}-m_t} \left(w \times \theta (1-\theta)^{y_t - m_t} + (1-w) \times \frac{e^{-\lambda} \lambda^{y_t - m_t}}{(y_t - m_t)!} \right).$$

The introduction of certain latent (unobservable) random variables allows us to specify the AdINAR(1) model in terms of a set of conditional distributions. This alternative representation leads to a factorization of the model joint distribution which is of key importance to our Monte Carlo simulations.

3.1 Data augmentation

In the AdINAR(1) model, suppose that, in addition to the values of the counts $Y_1, \ldots, Y_T$, we could observe the values of the *maturations* $M_t = \alpha \circ Y_{t-1}$, as well as the values of a set of *mixture component indicators* $U_t \in \{0, 1\}$, for $t = 2, \ldots, T$. The M_t's would inform us the number of cases that matured from the previous epoch, breaking down Y_t into two parcels (maturations plus innovations), while the U_t's would tell us from which component of the mixture the value of the Z_t's were generated in one realization of the process.

We postulate that

$$(U_t \mid w) \sim \text{Bernoulli}(w)$$

and

$$(M_t \mid \alpha, Y_{t-1} = y_{t-1}) \sim \text{Binomial}(y_{t-1}, \alpha).$$

Furthermore, we assume that

$$\Pr\{Y_t = y_t \mid M_t = m_t, U_t = u_t, \theta, \lambda\} = \begin{cases} \theta(1-\theta)^{y_t - m_t}\, \mathbb{I}_{\{m_t, m_{t+1}, \dots\}}(y_t) & \text{if } \; u_t = 1 \\ \dfrac{e^{-\lambda}\lambda^{y_t - m_t}}{(y_t - m_t)!}\, \mathbb{I}_{\{m_t, m_{t+1}, \dots\}}(y_t) & \text{if } \; u_t = 0 \end{cases}$$

in which $\mathbb{I}_A$ denotes the indicator function of the set A, defined by $\mathbb{I}_A(x) = 1$, if $x \in A$, and $\mathbb{I}_A(x) = 0$, if $x \notin A$.

Using the law of total probability and the product rule, we have that

$$\begin{aligned}
&\Pr\{Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha, \theta, \lambda, w\} \\
&\quad = \sum_{m_t=0}^{y_{t-1}} \sum_{u_t \in \{0,1\}} \Pr\{Y_t = y_t, M_t = m_t, U_t = u_t \mid Y_{t-1} = y_{t-1}, \alpha, \theta, \lambda, w\} \\
&\quad = \sum_{m_t=0}^{y_{t-1}} \sum_{u_t \in \{0,1\}} \Bigg(\Pr\{M_t = m_t \mid Y_{t-1} = y_{t-1}, \alpha\} \\
&\qquad\qquad \times \Pr\{Y_t = y_t \mid M_t = m_t, U_t = u_t, \theta, \lambda\} \Pr\{U_t = u_t \mid w\} \Bigg).
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{I}_{\{m_t, m_{t+1}, \dots\}}(y_t) \times \mathbb{I}_{\{0,1,\dots,y_{t-1}\}}(m_t) &= \mathbb{I}_{\{0,1,\dots,y_t\}}(m_t) \times \mathbb{I}_{\{0,1,\dots,y_{t-1}\}}(m_t) \\
&= \mathbb{I}_{\{0,1,\dots,\min\{y_t, y_{t-1}\}\}}(m_t),
\end{aligned}$$

we come to the conclusion that this data augmented model [14, 15] induces the AdINAR(1) model likelihood function.

In the following, we take advantage of this alternative representation of the AdINAR(1) model by the data augmentation scheme to derive simple closed forms for the model parameters and latent variables full conditional distributions, after the forms of the prior distributions have been specified.

3.2 Full conditionals

For convenience, we adopt a simplified notation in the following derivations, using the same letters p and π to denote different probability functions or densities, with distinctions made clear from context.

Let the prior distibutions be

$$\begin{aligned} \alpha &\sim \text{Beta}(a_\alpha, b_\alpha), \qquad & \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda), \\ \theta &\sim \text{Beta}(a_\theta, b_\theta), \qquad & w &\sim \text{Beta}(a_w, b_w). \end{aligned}$$

Define $m = (m_2, \dots, m_T)$ and $u = (u_2, \dots, u_T)$. The joint distribution of the data augmented AdINAR(1) model factors as

$$p(y, m, u, \alpha, \theta, \lambda, w) = \left(\prod_{t=2}^{T} p(y_t \mid m_t, u_t, \theta, \lambda)\, p(m_t \mid y_{t-1}, \alpha) \right) \pi(\alpha)\, \pi(\theta)\, \pi(\lambda)\, \pi(w).$$

Using the symbol $\propto$ to denote proportionality up to a suitable normalization factor, and the label "all others" to designate the observed counts y, and all the other latent variables and model parameters, with the exception of the one under consideration, the full conditional distributions are derived by inspection of this factorization.

$$(\alpha \mid \text{all others}) \sim \text{Beta}\left(a_\alpha + \sum_{t=2}^{T} m_t, b_\alpha + \sum_{t=2}^{T} (y_{t-1} - m_t) \right)$$

$$(\theta \mid \text{all others}) \sim \text{Beta}\left(a_\theta + \sum_{t=2}^{T} u_t, b_\theta + \sum_{t=2}^{T} (y_t - m_t) \mathbb{I}_{\{1\}}(u_t) \right)$$

$$(\lambda \mid \text{all others}) \sim \text{Gamma}\left(a_\lambda + \sum_{t=2}^{T} (y_t - m_t) \mathbb{I}_{\{0\}}(u_t), b_\lambda + (T-1) - \sum_{t=2}^{T} u_t \right)$$

$$(w \mid \text{all others}) \sim \text{Beta}\left(a_w + \sum_{t=2}^{T} u_t, b_w + (T-1) - \sum_{t=2}^{T} u_t \right)$$

$$\begin{aligned} \Pr\{U_t = 1 \mid \text{all others}\} &\propto w\ \theta (1-\theta)^{y_t - m_t}; \\ \Pr\{U_t = 0 \mid \text{all others}\} &\propto (1-w)\, \frac{e^{-\lambda} \lambda^{y_t - m_T}}{(y_t - m_t)!}, \end{aligned}$$

for $t = 2, \dots, T$.

$$\Pr\{M_t = m_t \mid \text{all others}\}$$
$$\propto \begin{cases} \dfrac{1}{(y_{t-1} - m_t)!\ m_t!} \left(\dfrac{\alpha}{(1-\theta)(1-\alpha)} \right)^{m_t} \mathbb{I}_{\{0,1,\dots,\min\{y_t, y_{t-1}\}\}}(m_t) & \text{if } u_t = 1 \\ \dfrac{1}{(y_t - m_t)!\ (y_{t-1} - m_t)!\ m_t!} \left(\dfrac{\alpha}{\lambda\ (1-\alpha)} \right)^{m_t} \mathbb{I}_{\{0,1,\dots,\min\{y_t, y_{t-1}\}\}}(m_t) & \text{if } u_t = 0 \end{cases}$$

for $t = 2, \dots, T$.

Using these full conditional distributions, we can code a Gibbs sampler [10, 11] to explore the posterior distribution.

3.3 Forecasting

The Gibbs sampler described above yields a sample $\{\alpha^{(n)}, \theta^{(n)}, \lambda^{(n)}, w^{(n)}\}_{n=1}^{N}$ from the posterior distribution. Uncertainty about future counts is represented by the h-steps-ahead posterior predictive distribution

$$Y_{T+h} \mid Y_1 = y_1, \dots, Y_T = y_T,$$

for some target $h \geq 1$. In particular, a pointwise forecast is obtained as a suitable summary of this posterior predictive distribution. In particular, a pointwise forecast is obtained as a suitable summary of this posterior predictive distribution.

To get a Monte Carlo approximation of the h-steps-ahead posterior predictive distribution, we use the AdINAR(1) model definition to propagate the process to the future sequentially, generating synthetic counts

$$\begin{aligned} y_{T+1}^{(n)} &= \alpha^{(n)} \circ y_T + z_{T+1}^{(n)}, \\ &\vdots \\ y_{T+h}^{(n)} &= \alpha^{(n)} \circ y_{T+h-1}^{(n)} + z_{T+h}^{(n)}, \end{aligned}$$

for $n = 1, \dots, N$, in which the synthetic innovations $z_{T+1}^{(n)}, \dots, z_{T+h}^{(n)}$ are drawn independently from a Geometric$(\theta^{(n)})$ distribution, with probability $w^{(n)}$, or from a Poisson$(\lambda^{(n)})$ distribution, with probability $1 - w^{(n)}$.

From the sample $\{y_{T+h}^{(n)}\}_{n=1}^{N}$, we approximate the h-steps-ahead posterior probability function by the respective empirical averages

$$p(y_{T+h} \mid y_1, \dots, y_T) \approx \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}_{\{y_{T+h}\}}(y_{T+h}^{(n)}),$$

for $y_{T+h} \geq 0$.

As a pointwise forecast $\hat{y}_{T+h}$, since we are dealing with discrete observations, we compute a generalized median of the h-steps-ahead posterior predictive distribution, defined by

$$\hat{y}_{T+h} = \arg \min_{y_{T+h} \geq 0} \left| 0.5 - \sum_{r=0}^{y_{T+h}} p(r \mid y_1, \dots, y_T) \right| .$$

We use a form of predictive cross-validation to evaluate the forecasting performance of the model. For an observed time series $y_1, \dots, y_T$, we pick some $T^* < T$, and treat the counts $y_{T^*}, \dots, y_T$ as a holdout (test) sample. For $t \geq T^*$, we train the model conditioning only on the values $y_1, \dots, y_{t-1}$ and making an h-steps-ahead out-of-sample prediction $\hat{y}_{t+h}$. To score the forecast performance, we average the median deviations $|\hat{y}_{t+h} - y_{t+h}|$ over all out-of-sample predictions. This cross-validation procedure is depicted in Figure 1.

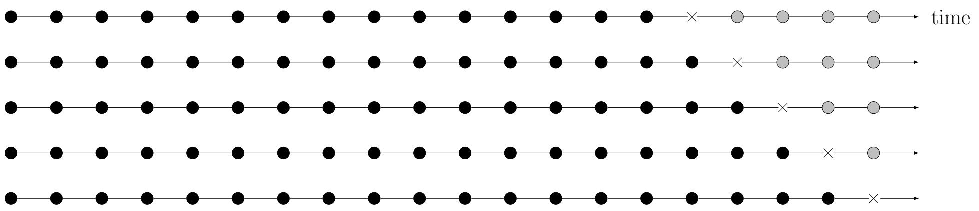

Figure 1: Predictive cross-validation scheme for one-step-ahead predictions. For each line, the black dots indicate the training set. Predictions are made for the target epoch marked with an ×.

4 DP-INAR(1) model

The DP-INAR(1) model completes the generalized INAR(1) model defined in Section 2, placing a Dirichlet process $\mathbb{G}$ at the top of the hierarchy. We recollect the necessary Dirichlet process facts and notations in the Appendix.

Formally, the innovations Z_t are modeled, given λ_t, as conditionally independent and identically distributed, with distribution Poisson(λ_t), and the innovation rates $\lambda_2, \dots, \lambda_T$, given $\mathbb{G} \sim \text{DP}(\tau\, G_0)$, are conditionally independent and identically distributed, with $\Pr\{\lambda_t \in B \mid \mathbb{G} = G\} = G(B)$, for every Borel set B. The prior distributions for α and τ are Beta(a_α, b_α) and Gamma(a_τ, b_τ), respectively. The base probability measure G_0 is a Gamma(a_0, b_0) distribution.

4.1 Data augmentation

The DP-INAR(1) model can also be data augmented, postulating that the maturations are distributed as

$$M_t \mid \alpha, Y_{t-1} = y_{t-1} \sim \text{Binomial}(y_{t-1}, \alpha),$$

and

$$\Pr\{Y_t = y_t \mid M_t = m_t, \lambda_t\} = \frac{e^{-\lambda_t} \lambda_t^{y_t - m_t}}{(y_t - m_t)!}\, \mathbb{I}_{\{m_t, m_{t+1}, \dots\}}(y_t).$$

Figure 2 displays a graphical representation of the data augmented DP-INAR(1) model. In the graph, absence of an arrow connecting two random objects means that they are conditionally independent given their parents (see [16] for a witful discussion of graphical models).

4.2 Full conditionals

Define $m = (m_2, \dots, m_T)$, and let $\mu_{\mathbb{G}}$ denote the distribution of $\mathbb{G}$. Marginalizing $\mathbb{G}$ on the graph, we have

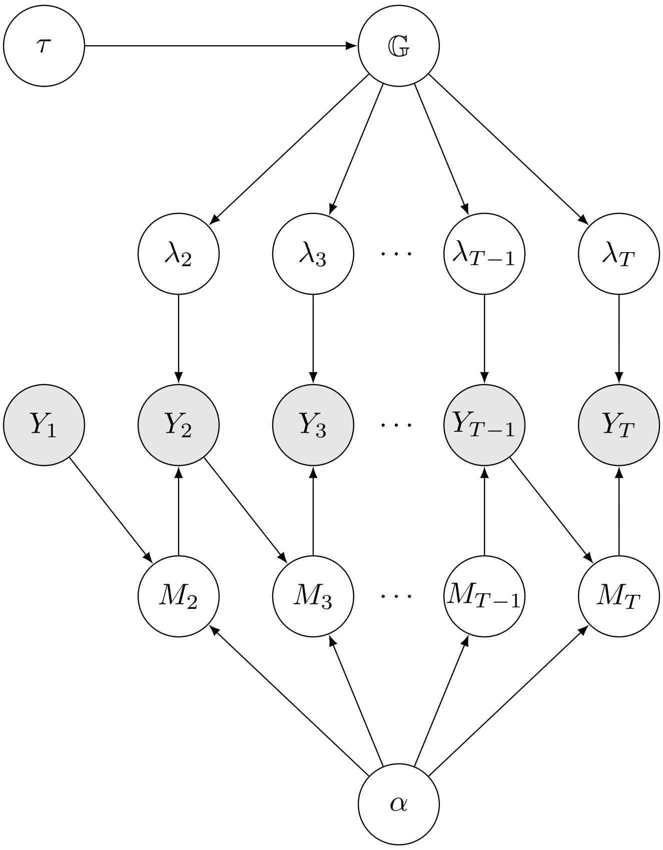

Figure 2: The data augmented DP-INAR(1) model.

$$\begin{aligned} p(y, m, \alpha, \lambda) &= \int p(y, m, \alpha, \lambda \mid G)\, d\mu_{\mathbb{G}}(G) \\ &= \left\{ \prod_{t=2}^{T} p(y_t \mid m_t, \lambda_t)\; p(m_t \mid y_{t-1}, \alpha) \right\} \times \pi(\alpha) \times \int \prod_{t=2}^{T} p(\lambda_t \mid G)\, d\mu_{\mathbb{G}}(G). \end{aligned}$$

Since the random vector $(\lambda_2, \dots, \lambda_T)$ has an exchangeable distribution, using this symmetry and the product rule, we can always make $p(\lambda_2, \dots, \lambda_T)$ depend on a certain λ_t only through $p(\lambda_t \mid \lambda_{\setminus t})$, in which $\lambda_{\setminus t}$ denotes the vector λ with the component λ_t removed. Hence,

$$p(\lambda_t \mid \text{all others}) \propto p(y, m, \alpha, \lambda) \propto p(\lambda_t \mid \lambda_{\setminus t})\; p(y_t \mid m_t, \lambda_t) \propto e^{-\lambda_t} \lambda_t^{y_t - m_t}\; p(\lambda_t \mid \lambda_{\setminus t}).$$

Therefore, the Pólya-Blackwell-MacQueen urn process yields the full conditional distribution of λ_t as the mixture

$$\begin{aligned} \lambda_t \mid \text{all others} \sim \frac{\tau \cdot b_0^{a_0} \cdot \Gamma(y_t - m_t + a_0)}{\Gamma(a_0)(b_0 + 1)^{y_t - m_t + a_0}} &\times \text{Gamma}(y_t - m_t + a_0, b_0 + 1) \\ + \sum_{r \neq t} \lambda_r^{y_t - m_t} e^{-\lambda_r} \delta_{\{\lambda_r\}}&, \end{aligned}$$

in which $\delta_{\{\lambda_r\}}$ denotes a point mass at λ_r. In the former expression we suppressed the normalization constant which makes all mixture weights add up to one.

The derivations of the full conditionals for α and m_t are straightforward.

$$\alpha \mid \text{all others} \sim \text{Beta}\left(a_\alpha + \sum_{t=2}^{T} m_t, b_\alpha + \sum_{t=2}^{T} (y_{t-1} - m_t) \right).$$

$$p(m_t \mid \text{all others}) \propto \frac{1}{m_t!(y_t - m_t)!(y_{t-1} - m_t)!} \left(\frac{\alpha}{\lambda_t (1-\alpha)} \right)^{m_t} \mathbb{I}_{\{0,1,\dots,\min\{y_{t-1},y_t\}\}}(m_t).$$

West [17] shows how to derive the full conditional distribution of the concentration parameter τ in simple closed form, after the introduction of an auxiliary random variable U. Using this technique, we have the full conditionals

$$U \mid \text{all others} \sim \text{Beta}(\tau + 1, T - 1);$$

$$\begin{aligned} \tau \mid \text{all others} \sim & \frac{\Gamma(a_\tau + k)}{(b_\tau - \log u)^{a_\tau + k - 1}} \times \text{Gamma}(a_\tau + k, b_\tau - \log u) \\ & + \frac{(T-1) \cdot \Gamma(a_\tau + k - 1)}{(b_\tau - \log u)^{a_\tau + k - 1}} \times \text{Gamma}(a_\tau + k - 1, b_\tau - \log u), \end{aligned}$$

in which we suppressed the normalization constant which makes the two mixture weights add up to one.

These full conditional distributions allow us to explore the model posterior distribution by coding a Gibbs sampler [10, 11]. Experimentation with this Gibbs sampler shows that, as pointed out by Escobar and West [18] in a similar context, we can improve mixing by resampling simultaneously the values of all λ_t's inside the same cluster at the end of each iteration. Formally, let $(\lambda_1^*, \dots, \lambda_k^*)$ be the k unique values among $(\lambda_2, \dots, \lambda_T)$ and define the number of occupants of cluster j by $n_j = \sum_{t=2}^{T} \mathbb{I}_{\{\lambda_j^*\}}(\lambda_t)$. It follows that

$$\lambda_j^* \mid \text{all others} \sim \text{Gamma}\left(a_0 + \sum_{t=2}^{T} (y_t - m_t) \cdot \mathbb{I}_{\{\lambda_j^*\}}(\lambda_t), b_0 + n_j \right).$$

for $j = 1, \dots, k$. After the λ_j^*'s are sampled from this distribution, we update the values of all λ_t's inside each cluster by the corresponding λ_j^*.

4.3 Choice of prior parameters

Extending the original scheme proposed by Dorazio [19], we choose the parameters a_τ and b_τ of the τ prior by minimizing the Kullback-Leibler divergence between the prior distribution of the number of clusters K and a uniform discrete distribution on a suitable range. Using the results in the Appendix, the marginal probability function of K can be computed as

$$\pi(k) = \int_0^\infty \Pr\{K = k \mid \tau\} \, \pi(\tau) \, d\tau = \frac{b_\tau S(T-1, k)}{\Gamma(a_\tau)} I(a_\tau, b_\tau; k),$$

for $k = 1, \dots, T-1$, in which

$$I(a_\tau, b_\tau; k) = \int_0^\infty \frac{\tau^{k+a_\tau-1}\ e^{-b_\tau \tau}\ \Gamma(\tau)}{\Gamma(\tau + T - 1)}\, d\tau.$$

Using the information available about the phenomena under consideration to make a sensible choice for the integers k_{min} and k_{max}, and letting q be the probability function of a uniform discrete distribution on $\{k_{\mathrm{min}}, \dots, k_{\mathrm{max}}\}$, that is

$$q(k) = \frac{1}{(k_{\mathrm{max}} - k_{\mathrm{min}} + 1)}\, \mathbb{I}_{\{k_{\mathrm{min}},\dots,k_{\mathrm{max}}\}}(k),$$

we find, by numerical integration and optimization, the values of a_τ and b_τ that minimize the Kullback-Leibler divergence

$$\begin{aligned} \mathrm{KL}[\pi \parallel q] &= \sum_{k=k_{\mathrm{min}}}^{k_{\mathrm{max}}} q(k) \log\left(\frac{q(k)}{\pi(k)}\right) \\ &= (\text{constant}) + \log\Gamma(a_\tau) - a_\tau \log b_\tau - \frac{1}{(k_{\mathrm{max}} - k_{\mathrm{min}} + 1)} \sum_{k=k_{\mathrm{min}}}^{k_{\mathrm{max}}} \log I(a_\tau, b_\tau; k). \end{aligned}$$

We choose the parameters a_0 and b_0 of the base probability density g_0 in a similar fashion, minimizing the Kullback-Leibler divergence between g_0 and a uniform distribution on a suitable range $[0, \lambda_{\mathrm{max}}]$, in which λ_{max} is chosen by taking into consideration the available information on the studied phenomena. Letting h be a uniform density on $[0, \lambda_{\mathrm{max}}]$, that is

$$h(\lambda) = \left(\frac{1}{\lambda_{\mathrm{max}}}\right) \mathbb{I}_{[0,\lambda_{\mathrm{max}}]}(\lambda),$$

we find, by numerical optimization, the values of a_0 and b_0 that minimize the Kullback-Leibler divergence

$$\begin{aligned} \mathrm{KL}[g_0 \parallel h] &= \int_0^{\lambda_{\mathrm{max}}} \left(\frac{1}{\lambda_{\mathrm{max}}}\right) \log\left(\frac{1/\lambda_{\mathrm{max}}}{g_0(\lambda)}\right) d\lambda \\ &= -\log\lambda_{\mathrm{max}} - a_0 \log b_0 + \log\Gamma(a_0) - (a_0 - 1)(\log\lambda_{\mathrm{max}} - 1) + \frac{b_0 \lambda_{\mathrm{max}}}{2}. \end{aligned}$$

4.4 Forecasting

Let $\{\alpha^{(n)}, \lambda^{(n)}\}_{n=1}^N$ be a sample from the posterior distribution obtained by Gibbs sampling. Using the law of total probability, the product rule, and simplifying the conditional independences in the model, we can write the posterior predictive probability function as

$$\begin{aligned} p(y_{T+h} \mid y_1, \dots, y_T) = \int & p(y_{T+h} \mid y_T, \alpha, \lambda_{T+1}, \dots, \lambda_{T+h}) \\ & \times \prod_{i=1}^h p(\lambda_{T+i} \mid \lambda_2, \dots, \lambda_{T+i-1}) \\ & \times p(\alpha, \lambda_2, \dots, \lambda_T \mid y_1, \dots, y_T)\, d\alpha\, d\lambda_2 \dots d\lambda_{T+h}. \end{aligned}$$

A nice property of the DP-INAR(1) model is that we can derive a simple analytical expression for the first factor in the integrand above.

Proposition 4.1. *The probability function of Y_{t+h}, given $Y_t = y_t$, α, and $(\lambda_{t+1}, \dots, \lambda_{t+h})$, can be writen as the convolution of a Bin(y_t, α^h) distribution and a Poisson(μ_h) distribution,*

$$p(y_{t+h} \mid y_t, \alpha, \lambda_{t+1}, \dots, \lambda_{t+h}) = \sum_{m=0}^{\min\{y_t, y_{t+h}\}} \binom{y_t}{m} (\alpha^h)^m (1-\alpha^h)^{y_t - m} \left(\frac{\mu_h^{y_{t+h}-m} e^{-\mu_h}}{(y_{t+h}-m)!} \right),$$

in which

$$\mu_h = \sum_{i=1}^{h} \alpha^{h-i} \lambda_{t+i}.$$

Proof. We prove the result by induction. For $h = 1$, using a simplified notation, the conditional moment generating function is given by

$$M_{Y_{t+1}|Y_t}(s) = \mathrm{E}\big[e^{sY_{t+1}} \mid Y_t\big] = \mathrm{E}\big[e^{s(\alpha \circ Y_t + Z_{t+1})} \mid Y_t\big] = \mathrm{E}\Big[e^{s(\sum_{i=1}^{Y_t} B_i(t) + Z_{t+1})} \mid Y_t\Big],$$

But since $\{Z_t\}_{t \geq 2}$ is a sequence of conditionally independent random variables, which is also conditionally independent of $\{B_i(t) : i \geq 0, t \geq 2\}$, we have that

$$M_{Y_{t+1}|Y_t}(s) = \mathrm{E}\Big[e^{s\sum_{i=1}^{Y_t} B_i(t)} \mid Y_t\Big]\, \mathrm{E}\big[e^{sZ_{t+1}}\big] = (\alpha e^s + (1-\alpha))^{Y_t} \exp(\lambda_{t+1}(e^s - 1)),$$

which is the product of the generating functions of a Binomial(Y_t, α) random variable and a Poisson(λ_{t+1}) random variable. Now, suppose the result holds for an arbitrary $h \geq 2$. Then,

$$\begin{aligned} M_{Y_{t+h+1}|Y_t}(s) &= \mathrm{E}\big[e^{sY_{t+h+1}} \mid Y_t\big] = \mathrm{E}\big[\mathrm{E}\big[e^{sY_{t+h+1}} \mid Y_{t+h}\big] \mid Y_t\big] \\ &= \mathrm{E}\big[e^{uY_{t+h}} \mid Y_t\big] \exp(\lambda_{t+h+1}(e^s - 1)), \end{aligned}$$

in which we defined $e^u = \alpha e^s + (1-\alpha)$. Consequently, from the induction hypothesis, we have that

$$\begin{aligned} M_{Y_{t+h+1}|Y_t}(s) &= (\alpha^h e^u + (1-\alpha^h))^{Y_t} \exp(\mu_h(e^u - 1)) \exp(\lambda_{t+h+1}(e^s - 1)) \\ &= (\alpha^h(\alpha e^s + (1-\alpha)) + (1-\alpha^h))^{Y_t} \exp(\mu_h((\alpha e^s + (1-\alpha)) - 1)) \\ &\quad \times \exp(\lambda_{t+h+1}(e^s - 1)) \\ &= (\alpha^{h+1} e^s + (1-\alpha^{h+1}))^{Y_t} \exp(\mu_{h+1}(e^s - 1)), \end{aligned}$$

in which $\mu_{h+1} = \alpha\mu_h + \lambda_{t+h+1}$. Hence, the result holds for $h+1$, completing the proof. □

Using the Pólya-Blackwell-MacQueen urn process repeatedly, for $n = 1 \dots, N$, we draw a sample $\{\lambda_{T+1}^{(n)}, \dots, \lambda_{T+h}^{(n)}\}_{n=1}^N$ from $\prod_{i=1}^h p(\lambda_{T+i} \mid \lambda_2, \dots, \lambda_{T+i-1})$ sequentially as follows:

$$\begin{aligned} \lambda_{T+1}^{(n)} &\sim \frac{\tau}{\tau+T}\, G_0 + \frac{1}{\tau+T} \sum_{t=2}^{T} \delta_{\{\lambda_t^{(n)}\}}; \\ \lambda_{T+2}^{(n)} &\sim \frac{\tau}{\tau+T+1}\, G_0 + \frac{1}{\tau+T+1} \sum_{t=2}^{T+1} \delta_{\{\lambda_t^{(n)}\}}; \\ &\vdots \\ \lambda_{T+h}^{(n)} &\sim \frac{\tau}{\tau+T+h-1}\, G_0 + \frac{1}{\tau+T+h-1} \sum_{t=2}^{T+h-1} \delta_{\{\lambda_t^{(n)}\}}. \end{aligned}$$

Combining all these elements, we approximate the integral representation of the h-steps-ahead posterior predictive probability function by the Monte Carlo average

$$p(y_{T+h} \mid y_1, \dots, y_T) \approx \frac{1}{N} \sum_{n=1}^{N} p(y_{T+h} \mid y_T, \alpha^{(n)}, \lambda_{T+1}^{(n)}, \dots, \lambda_{T+h}^{(n)}),$$

for $y_{T+h} \geq 0$.

5 Pittsburgh crime data

In this section, we analyze monthly time series of burglary events in Pittsburgh, USA, from January 1990 to December 2001 [20]. In this dataset, each time series has a length of 144 months and corresponds to a certain patrol area, comprising a total of 36 time series.

In what follows, we use patrol area 58 to exemplify the training procedures for the AdINAR(1) and the DP-INAR(1) models. This patrol area presents substantial overdispersion in the monthly counts of burglary events, with mean 10.4 and variance 31.1.

In all runs of the Gibbs samplers, we discard the first 10^3 simulated values, which correspond to the burn-in period, and end up with a posterior sample of size 10^4.

For the AdINAR(1) model hyperparameters, we make the choices $a_\alpha = 1$, $b_\alpha = 1$, $a_\lambda = 1$, $b_\lambda = 0.1$, $a_\theta = 1$, $b_\theta = 1$, $a_w = 1$, and $b_w = 1$, which correspond to reasonably flat priors.

Figure 3 displays the marginal posterior distributions of the AdINAR(1) model parameters. The posterior distribution of the thinning parameter α is fairly concentrated, with posterior mean 0.31, showing that the autoregressive component is not negligible for this patrol area. The posterior mean of λ is 6.78, while the posterior mean of θ is 0.12. Also, the posterior distribution of w, with posterior mean 0.38, shows that the geometric component of the mixture has less weight for this patrol area.

For the DP-INAR(1) model, we specify the hyperparameters as follows. To determine a_τ and b_τ, the optimization procedure described in Section 4.3, with $k_{\min} = 1$ and $k_{\max} = 143$, yields $a_\tau = 0.519$ and $b_\tau = 0.003$. Note that these values of $k_{\min}$ and $k_{\max}$ correspond, within our scheme, to the most spread choice for the prior distribution of the number of clusters K. With regard to the base measure, Figure 4 displays the contour plot of the corresponding Kullback-Leibler divergence $\mathrm{KL}[g_0 \parallel h]$. The minimum is attained at $a_0 = 1.778$ and $b_0 = 0.096$. Finally, we choose a uniform prior for the thinning parameter α, making $a_\alpha = 1$ and $b_\alpha = 1$.

The DP-INAR(1) marginal posterior distributions of the parameters α, λ_3, λ_{18}, and λ_{96} are displayed in Figure 5. For this patrol area, the posterior mean of the thinning parameter α is 0.19. The posterior means of λ_3, λ_{18} and λ_{96} are equal to 6.50, 13.61 and 32.01, respectively, showing that different regimes of innovation rates were captured in the learning process. Figure 6 shows both the prior and posterior distributions of the number of clusters K. While the prior distribution is reasonably flat in the range 1 to 143, the posterior distribution is concentrated around 7, the posterior mode.

The Markov chains in Figures 7 and 8 indicate that proper mixing is achieved by the Gibbs samplers for both models.

With regard to the forecasting performance within this dataset, Table 1 presents the mean absolute deviations of the out-of-sample predictions for the INAR(1), AdINAR(1), and DP-INAR(1), in the 36 patrol areas. In this table, the mean absolute deviations are computed predicting the values of the last 44 months of each time series, using the predictive cross-validation procedure described in Section 3.3.

The results in Table 1 show that the AdINAR(1) and the DP-INAR(1) models outperform the INAR(1) model in 75% of the patrol areas. From the last two columns of the table, we see that the AdINAR(1) model and the DP-INAR(1) model produce substantial relative gains in the mean absolute deviations, with the exception of five areas in which the INAR(1) performs better, but with smaller relative gains.

6 Conclusions

Two Bayesian generalizations of the INAR(1) model are proposed. The AdINAR(1) model accounts for overdispersion in the time series using a Geometric-Poisson mixture as the marginal distribution of the innovation process, while the DP-INAR(1) model is capable of learning a latent pattern of heterogeneity in the distribution of the innovation rates by means of a Dirichlet process placed at the top of the model hierarchy. For both models, we devise data augmentation schemes from which we derive full conditional distributions in simple analytical forms. Simulations of the posterior distributions through Gibbs sampling, and a predictive cross-validation procedure, give evidence of good forecasting performance in the analysis of times series of burglary events in Pittsburgh, USA. An open source R package implementing both models is available at: https://github.com/heltongraziadei/BINAR.

Acknowledgements

Helton Graziadei and Hedibert F. Lopes thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for financial support through grants numbers 2017/10096-6 and 2017/22914-5.

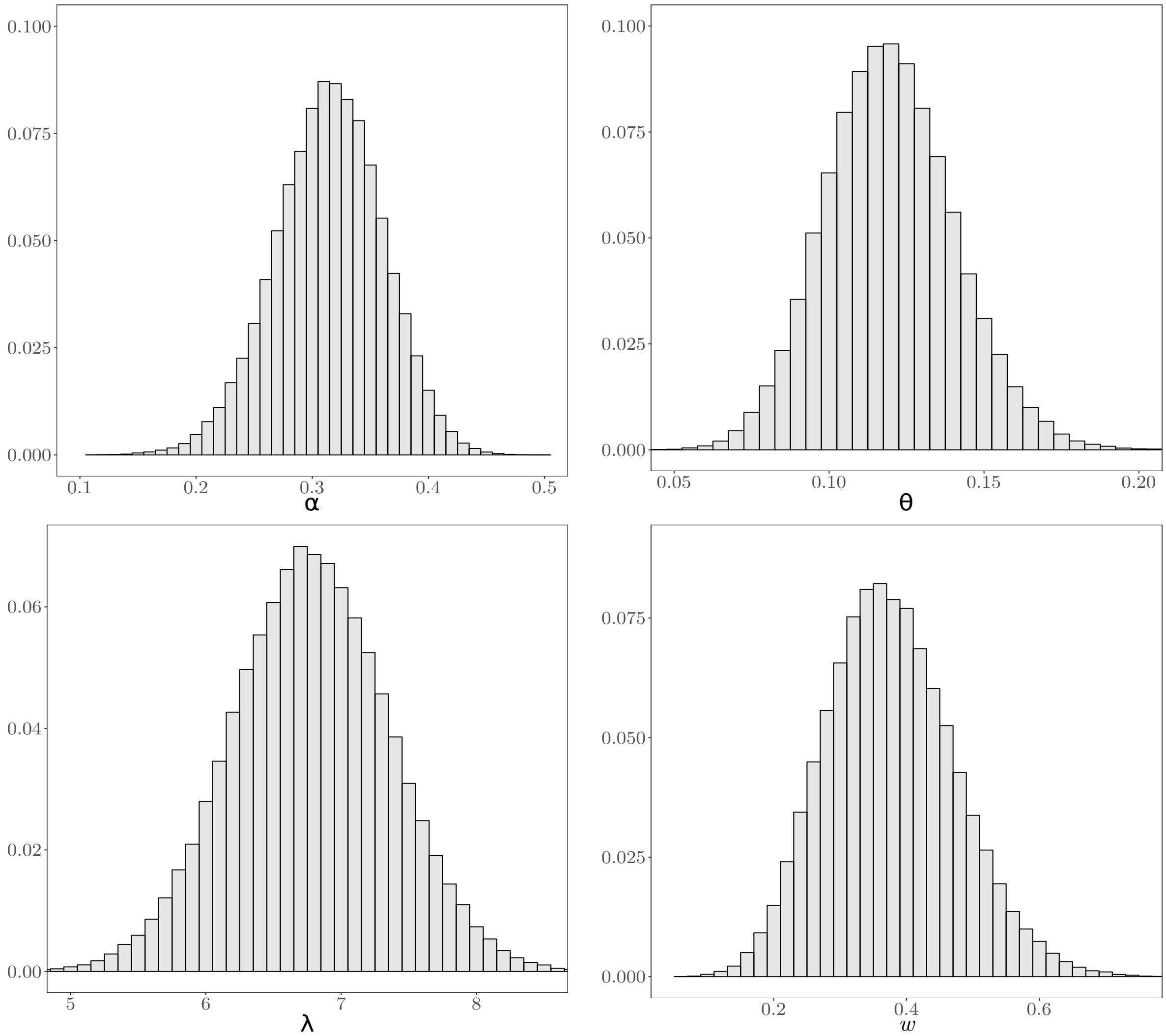

Figure 3: Marginal posterior distributions of the AdINAR(1) model parameters α, θ, λ, and w for patrol area 58.

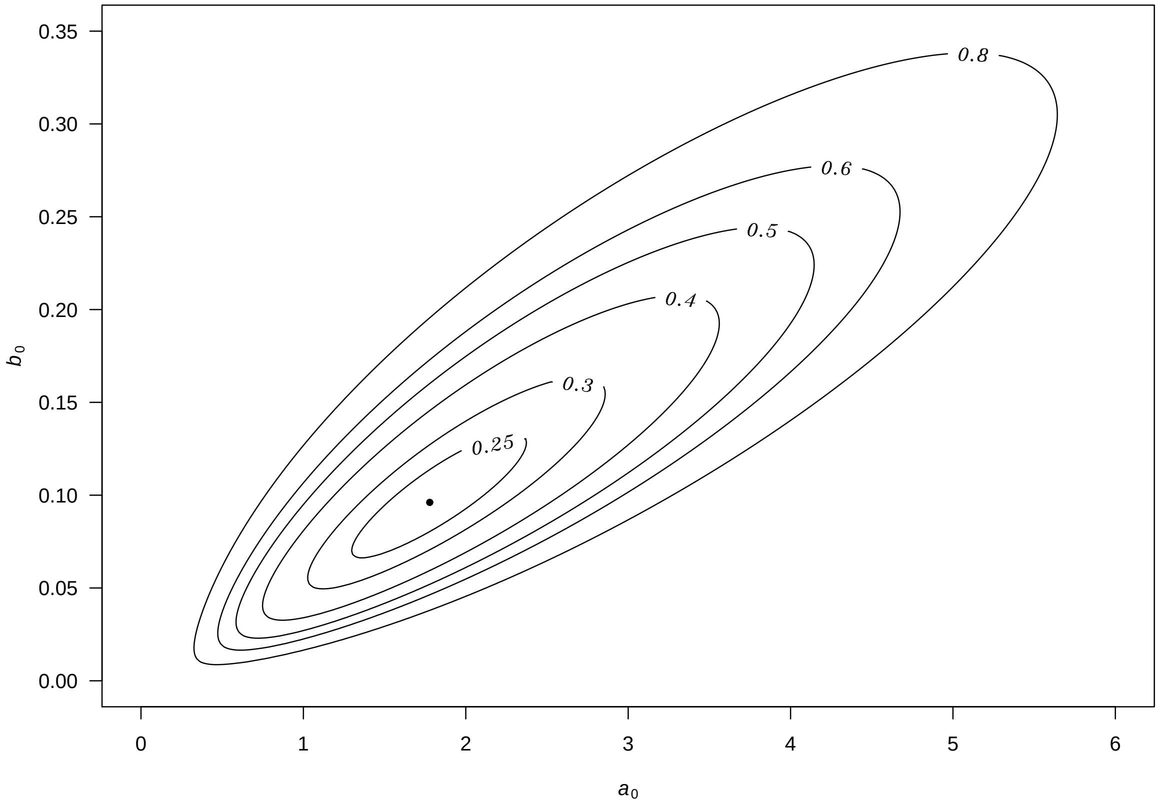

Figure 4: Contour plot of the Kullback-Leibler divergence associated with the optimization of the base measure hyperparameters for patrol area 58.

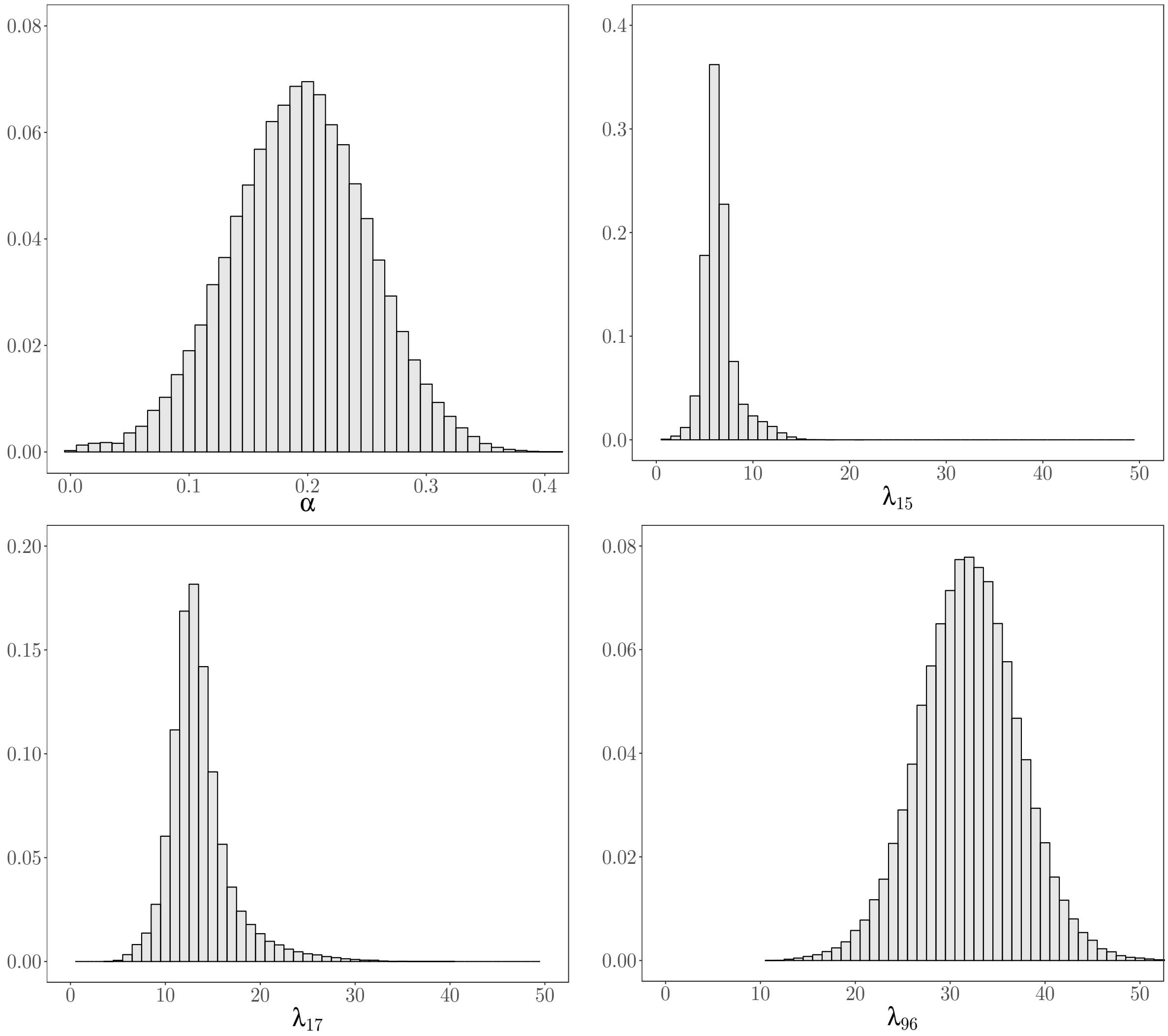

Figure 5: Marginal posterior distributions of the DP-INAR(1) model parameters α, λ_3, λ_{18}, and λ_{96} for patrol area 58.

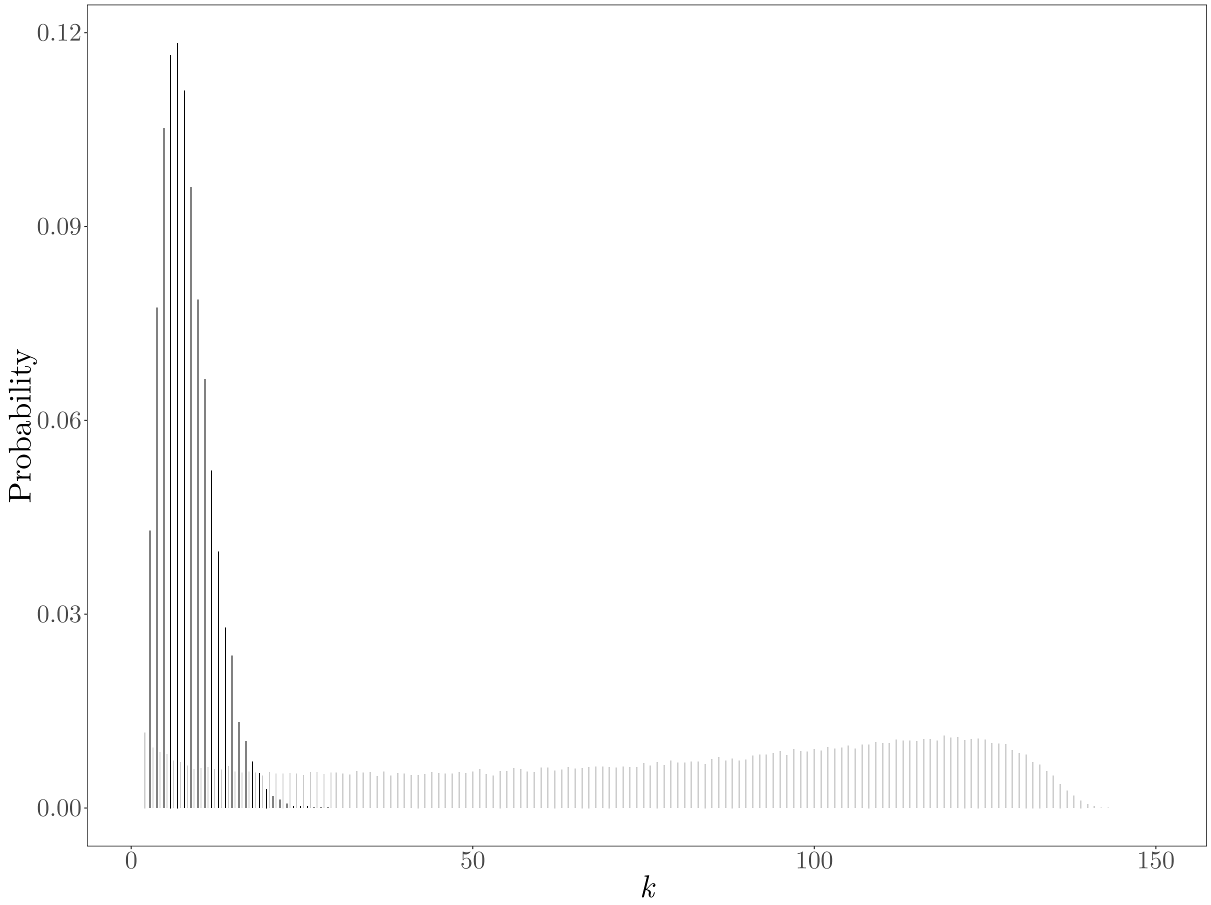

Figure 6: DP-INAR(1) model prior and posterior distributions for the number of clusters K, in gray and black respectively, for patrol area 58.

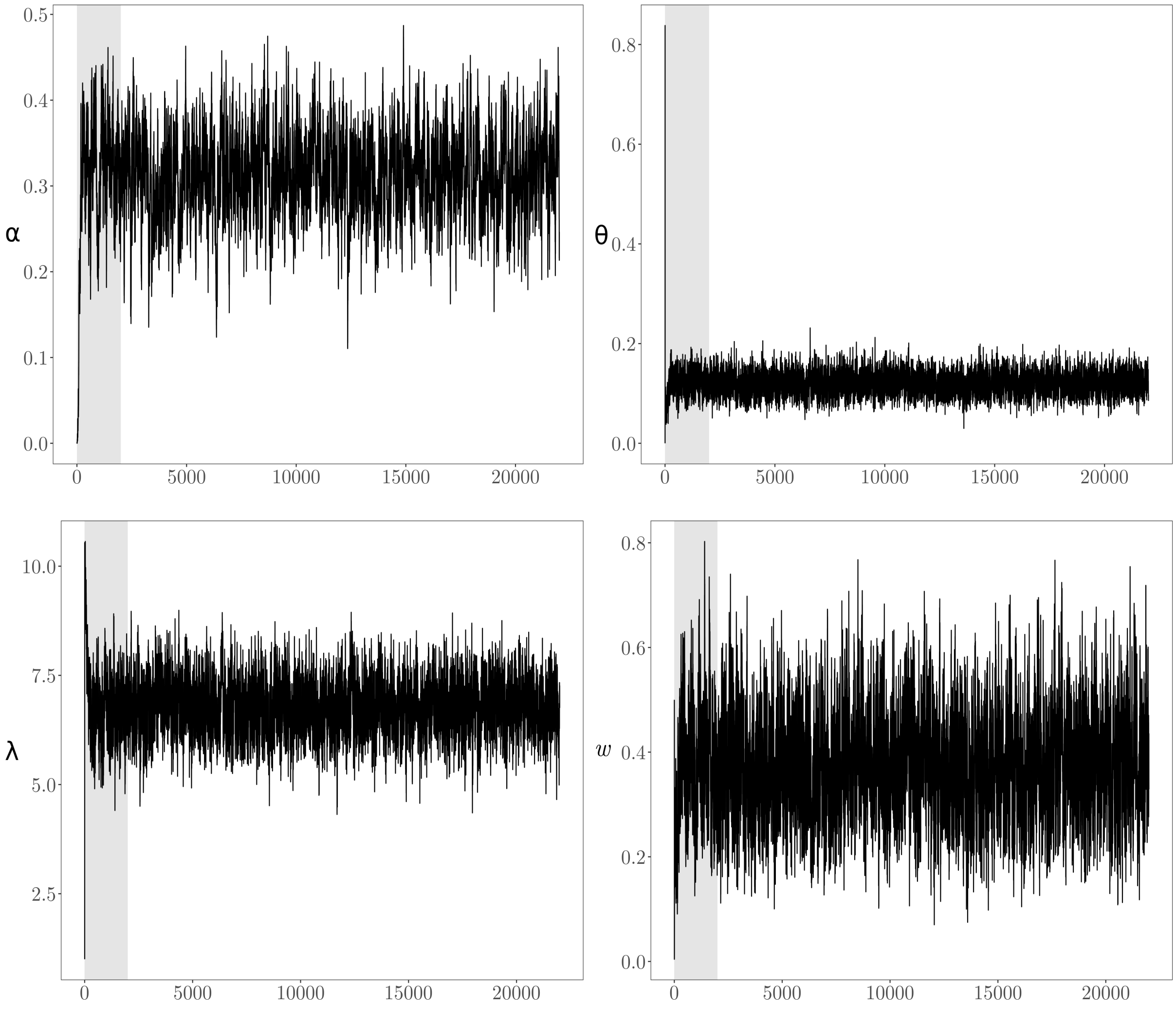

Figure 7: Markov chains associated with the AdINAR(1) model marginal posterior distributions of parameters α, θ, λ, and w for patrol area 58. The gray rectangles indicate the burn-in periods.

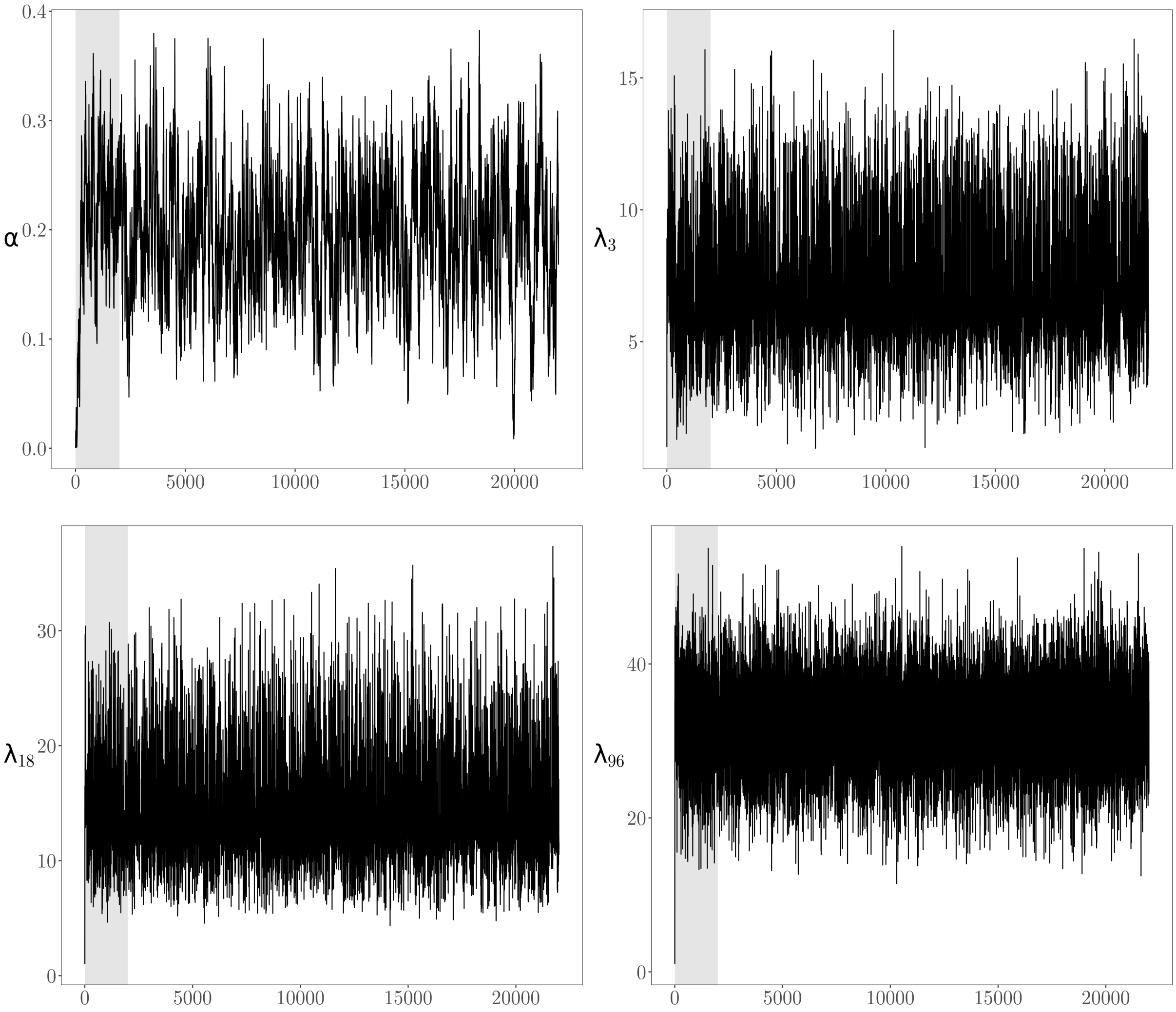

Figure 8: Markov chains associated with the DP-INAR(1) model marginal posterior distributions of parameters α, λ_3, λ_{18}, and λ_{96} for patrol area 58. The gray rectangles indicate the burn-in periods.

Table 1: Mean absolute deviations for the out-of-sample predictions of the INAR(1), AdINAR(1) and DP-INAR(1) models. The last two columns show the relative variations for the AdINAR(1) and DP-INAR(1) models with respect to the INAR(1) model (lower is better). For each patrol area, the best mean absolute deviation is written in bold face.

Patrol Area	INAR(1)	AdINAR(1)	DP-INAR(1)	$\Delta_{\text{AdINAR(1)}}$	$\Delta_{\text{DP-INAR(1)}}$
11	**1.209**	**1.209**	**1.209**	0.000	0.000
12	3.907	**3.349**	3.512	-0.143	-0.101
13	2.674	**2.628**	2.698	-0.017	0.009
14	2.581	**2.488**	2.535	-0.036	-0.018
15	2.791	**2.721**	**2.721**	-0.025	-0.025
16	2.093	**1.930**	2.000	-0.078	-0.044
17	2.279	**2.233**	2.256	-0.020	-0.010
21	1.186	**1.140**	**1.140**	-0.039	-0.039
22	2.279	**2.116**	**2.116**	-0.071	-0.071
23	3.302	3.256	**3.209**	-0.014	-0.028
24	1.651	**1.465**	1.535	-0.113	-0.070
25	1.302	1.395	**1.233**	0.071	-0.054
26	2.023	**1.209**	1.512	-0.402	-0.253
27	1.349	**1.186**	**1.186**	-0.121	-0.121
28	**0.814**	0.860	0.837	0.057	0.029
29	2.767	**2.744**	2.767	-0.008	0.000
31	3.488	3.698	**3.442**	0.060	-0.013
32	**3.442**	**3.442**	3.488	0.000	0.014
33	1.930	**1.721**	1.814	-0.108	-0.060
34	3.581	**3.535**	3.674	-0.013	0.026
41	2.372	**2.349**	2.395	-0.010	0.010
42	3.302	**3.209**	3.302	-0.028	0.000
43	2.163	**2.093**	2.186	-0.032	0.011
44	1.837	**1.721**	1.791	-0.063	-0.025
45	2.395	**2.326**	2.395	-0.029	0.000
46	2.744	2.744	**2.628**	0.000	-0.042
47	2.302	2.465	**2.256**	0.071	-0.020
51	**2.860**	3.093	2.930	0.081	0.024
52	**3.814**	3.977	3.930	0.043	0.030
53	**2.837**	2.930	2.884	0.033	0.016
54	2.884	2.558	**2.535**	-0.113	-0.121
55	**4.512**	5.419	4.884	0.201	0.082
56	2.093	**1.884**	1.930	-0.100	-0.078
57	**1.977**	2.047	**1.977**	0.035	0.000
58	2.977	**2.372**	2.512	-0.203	-0.156

References

[1] R. Davis, S. Holan, R. Lund, and N. Ravishanker, *Handbook of discrete-valued time series.* Chapman & Hall / CRC, 2015.

[2] E. McKenzie, "Some simple models for discrete variate time series," *Journal of the American Water Resources Association*, vol. 21, no. 4, pp. 645–650, 1985.

[3] M. Al-Osh and A. Alzaid, "First-order integer-valued autoregressive (INAR(1)) process: distributional and regression properties," *Statistica Neerlandica*, vol. 42, pp. 53–61, 1988.

[4] K. Freeland, *Statistical analysis of discrete time series with application to the analysis of workers' compensation data.* PhD thesis, University of British Columbia, Vancouver, 1998.

[5] C. H. Weiß, *An introduction to discrete-valued time series.* John Wiley & Sons, 2018.

[6] C. H. Weiß, "Thinning operations for modeling time series of counts -— a survey," *AStA Advances in Statistical Analysis*, vol. 92, no. 3, p. 319, 2008.

[7] I. Silva, M. E. Silva, I. Pereira, and N. Silva, "Replicated INAR(1) processes," *Methodology and Computing in applied Probability*, vol. 7, no. 4, pp. 517–542, 2005.

[8] N. Silva, I. Pereira, and M. E. Silva, "Forecasting in INAR(1) model," *REVSTAT–Statistical Journal*, vol. 7, no. 1, pp. 119–134, 2009.

[9] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[10] D. Gamerman and H. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* Chapman & Hall / CRC, 2006.

[11] C. Robert and G. Casella, *Monte Carlo statistical methods.* Springer Science & Business Media, 2013.

[12] R Core Team, *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.

[13] D. Eddelbuettel, *Seamless R and C++ integration with Rcpp.* Springer Publishing Company, Incorporated, 2013.

[14] M. Tanner and W. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.

[15] D. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[16] M. Jordan, “Graphical models,” *Statistical Science*, vol. 19, no. 1, pp. 140–155, 2004.

[17] M. West, *Hyperparameter estimation in Dirichlet process mixture models.* Duke University ISDS discussion paper #92-A03, 1992.

[18] M. Escobar and M. West, “Computing nonparametric hierarchical models,” in *Practical nonparametric and semiparametric Bayesian statistics* (D. Dey, P. Müller, and D. Sinha, eds.), ch. 1, pp. 1–22, Springer-Verlag, 1998.

[19] R. Dorazio, “On selecting a prior for the precision parameter of Dirichlet process mixture models,” *Journal of Statistical Planning and Inference*, vol. 139, no. 9, pp. 3384–3390, 2009.

[20] http://www.forecastingprinciples.com/index.php/crimedata.

[21] D. Blackwell and J. MacQueen, “Ferguson distributions via Pólya urn schemes,” *The Annals of Statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[22] C. Antoniak, “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

Appendix

Dirichlet process

Suppose that we represent our uncertainties about quantities assuming values in a sampling space $\mathscr{X}$, with sigma-field $\mathscr{B}$, by means of an underlying probability space $(\Omega, \mathscr{F}, \Pr)$.

A mapping $\mathbb{G} : \mathscr{B} \times \Omega \to [0,1]$ is a random probability measure if $\mathbb{G}(\,\cdot\,,\omega)$ is a probability measure over $(\mathscr{X}, \mathscr{B})$, for every $\omega \in \Omega$, and $\mathbb{G}(B) = \mathbb{G}(B,\cdot\,)$ is a random variable, for each $B \in \mathscr{B}$.

Ferguson [9] defined a random probability measure $\mathbb{G}$ descriptively as follows. Let β be a finite nonzero measure over $(\mathscr{X}, \mathscr{B})$ and postulate that for each $\mathscr{B}$-measurable partition $\{B_1, \dots, B_k\}$ of $\mathscr{X}$ the random vector $(\mathbb{G}(B_1), \dots, \mathbb{G}(B_k))$ has the ordinary Dirichlet distribution with parameters $(\beta(B_1), \dots, \beta(B_k))$. In this case, we say that $\mathbb{G}$ is a Dirichlet process with base measure β, and use the notation $\mathbb{G} \sim \mathrm{DP}(\beta)$. Ferguson proved that $\mathbb{G}$ is a properly defined random process in the sense of Kolmogorov's consistency theorem.

Defining the concentration parameter $\tau = \beta(\mathscr{X})$, and the base probability measure G_0 by $G_0(B) = \beta(B)/\beta(\mathscr{X})$, it follows from the usual properties of the Dirichlet distribution that $\mathrm{E}[\mathbb{G}(B)] = G_0(B)$ and $\mathrm{Var}[\mathbb{G}(B)] = G_0(B)(1 - G_0(B))/(\tau + 1)$, for every $B \in \mathscr{B}$. Therefore, $\mathbb{G}$ is centered on G_0, and τ controls the concentration of $\mathbb{G}$ around G_0. In terms of the concentration parameter and the base probability measure, we write $\mathbb{G} \sim \mathrm{DP}(\tau\, G_0)$.

Inference with the Dirichlet process is tractable. In particular, Ferguson proved that the Dirichlet process is closed under sampling: if $X_1, \dots, X_n$ are conditionally independent and identically distributed, given $\mathbb{G} \sim \mathrm{DP}(\tau\, G_0)$, such that $\Pr\{X_i \in B \mid \mathbb{G} = G\} = G(B)$, for every B in $\mathscr{B}$, then

$$\mathbb{G} \mid X_1 = x_1, \dots, X_n = x_n \sim \mathrm{DP}\left((\tau + n) \left(\frac{\tau}{\tau + n}\, G_0 + \frac{1}{\tau + n} \sum_{i=1}^{n} \mathbb{I}_B(x_i) \right) \right).$$

Notice that, using the law of total expectation, we have

$$\begin{aligned} \Pr\{X_{n+1} \in B \mid X_1, \dots, X_n\} &= \mathrm{E}[\Pr\{X_{n+1} \in B \mid \mathbb{G}, X_1, \dots, X_n\} \mid X_1, \dots, X_n] \\ &= \mathrm{E}[\Pr\{X_{n+1} \in B \mid \mathbb{G}\} \mid X_1, \dots, X_n] \\ &= \mathrm{E}[\mathbb{G}(B) \mid X_1, \dots, X_n], \end{aligned}$$

almost surely, for every B in $\mathscr{B}$, in which the second equality follows from the conditional independence of the X_i's. Hence, the posterior predictive distribution is

$$\Pr\{X_{n+1} \in B \mid X_1 = x_1, \dots, X_n = x_n\} = \frac{\tau}{\tau + n}\, G_0(B) + \frac{1}{\tau + n} \sum_{i=1}^{n} I_B(x_i).$$

This expression of the posterior predictive distribution unleashes important features of the Dirichlet process, thereby showing how it can be used as a modeling tool. In particular, it defines a data generating process known as the Pólya-Blackwell-MacQueen urn [21]. If we imagine the sequential generation of the X_i's, for $i = 1, \dots, n$, we see that a value is generated anew from G_0 with probability proportional to τ, or we repeat one the previously generated values with probability proportional to its multiplicity. This shows that, almost surely, realizations of a Dirichlet process $\mathbb{G}$ are discrete probability measures, maybe with denumerably infinite support, depending on the nature of G_0. Also, this data generating process associated with the Pólya-Blackwell-MacQueen urn implies that the X_i's are clustered, meaning that there is a positive probability that $X_i = X_j$, for $i \neq j$. Antoniak [22] derived the conditional distribution of the number of distinct X_i's, that is, the number of clusters K, given the concentration parameter τ, as

$$\Pr\{K = k \mid \tau\} = S(n,k)\, \tau^k \, \frac{\Gamma(\tau)}{\Gamma(\tau + n)} \, \mathbb{I}_{\{1,2,\dots,n\}}(k),$$

in which $S(n,k)$ denotes the unsigned Stirling number of the first kind.