On Some Mixture Models for INAR(1) Processes

Helton Graziadei, Paulo Marques and Hedibert Lopes

IME-USP & Insper

February 2021

- Marques, Graziadei and Lopes (2020)
  Bayesian generalizations of the INAR model.
  Journal of Applied Statistics.
  https://doi.org/10.1080/02664763.2020.1812544.
- Graziadei, Lijoi, Lopes, Marques and Prunster (2020)
  Prior sensitivity analysis in a semi-parametric INAR model.
  Entropy, 22, 69
  https://doi.org/10.3390/e22010069

Outline

Introduction

- Time series of counts arise in a wide range of applications such as econometrics, public policy and environmental studies.
- Traditional time series models consider continuously valued processes. In count scenarios, continuous time series models are not suitable for analyzing discrete data.
- We WILL NOT pursue the well known class of generalized dynamic linear models.
- We assume here a special autoregressive structure for discrete variables [Alzaid and Al-Osh, 1987, McKenzie, 1985].
- We consider some mixture models on the innovation process as a means to improve forecasting accuracy.

INAR(1) process

Consider a Markov process $\{Y_t\}_{t\in\mathbb{N}}$ represented by the following functional form [McKenzie, 1985, Alzaid and Al-Osh, 1987]:

$$\underbrace{Y_t}_{\text{Count at time } t} = \underbrace{\alpha \circ Y_{t-1}}_{\text{Survivors from } t-1} + \underbrace{Z_t}_{\text{Innovation at time } t},$$

where

$$M_t = \alpha \circ Y_{t-1} = \sum_{i=1}^{Y_{t-1}} B_i(t),$$

is refereed to here as *maturation* at time t, and $\{B_i(t)\}$ is a collection of independent Bernoulli(α) random variables.

The original formulation assumes that Z_t follows a parametric model, usually a Poisson or a Geometric distribution.

Our contributions

1. Model Z_t via a Poisson-Geometric mixture to account for over-dispersion in time series of counts.

2. Develop a semi-parametric model based on the Dirichlet Process in order to learn the patters of heterogeneity in time series of counts.

3. Investigate the Pitman-Yor process to robustify inference for the number of clusters.

The AdINAR(1) Model

- The AdINAR(1) model is defined such that Z_t is a mixture of a Geometric and a Poisson distributions

$$z_t \mid \theta, \lambda, w \sim w \, Geometric(\theta) + (1 - w) \, Poisson(\lambda)$$

$t = 2, \ldots, T, w \in [0, 1]$.

- As w becomes large, the innovation is contaminated by the Geometric distribution in the mixture, increasing variability of the process.

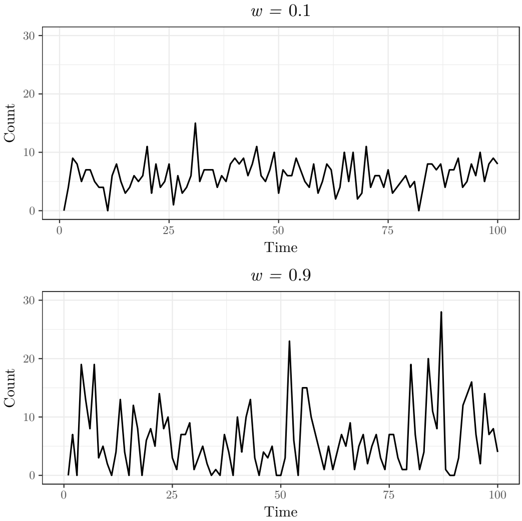


Figure: Typical simulated series for $w = 0.1$ and $w = 0.9$.

- The joint distribution of $(Y_1, \ldots, Y_T)$, given α and λ, can be written as

$$p(y_1, \ldots, y_T \mid \alpha, \theta, \lambda, w) = \prod_{t=2}^{T} p(y_t \mid y_{t-1}, \alpha, \theta, \lambda, w).$$

- The joint distribution of $(Y_1, \ldots, Y_T)$, given α and λ, can be written as

$$p(y_1, \ldots, y_T \mid \alpha, \theta, \lambda, w) = \prod_{t=2}^{T} p(y_t \mid y_{t-1}, \alpha, \theta, \lambda, w).$$

- The likelihood function of $y = (y_2, \ldots, y_T)$ is directly derived: Hence, the AdINAR(1) model likelihood function is given by

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

Reparameterizaton

Let us introduce some new items.

- Let $M = (M_2, \ldots, M_T)$ be the set of maturations.
- Let the model be augmented by the latent varables

$$u = (u_2, \ldots, u_T)$$

such that

$$u_t = 1, \text{if } z_t \mid \theta \sim Geometric(\theta)$$

or

$$u_t = 0, \text{if } z_t \mid \lambda \sim Poisson(\lambda),$$

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

Conditionally conjugate priors

$$\begin{aligned}
\text{Thinning: } \alpha &\sim Beta(a_0^{(\alpha)}, b_0^{(\alpha)}) \\
\text{Weight: } w &\sim Beta(a_0^{(w)}, b_0^{(w)}) \\
\text{Geometric: } \theta &\sim Beta(a_0^{(\theta)}, b_0^{(\theta)}) \\
\text{Poisson: } \lambda &\sim Gamma(a_0^{(\lambda)}, b_0^{(\lambda)})
\end{aligned}$$

Simpler conditional distributions

- Postulate that:

$$p(y_t \mid m_t, u_t = 1) \;=\; \theta(1-\theta)^{y_t - m_t}\mathbb{I}_{\{m_t, m_{t+1}, \dots\}}(y_t),$$

$$p(y_t \mid m_t, u_t = 0) \;=\; \frac{e^{-\lambda}\lambda^{y_t - m_t}}{(y_t - m_t)!}\mathbb{I}_{\{m_t, m_{t+1}, \dots\}}(y_t),$$

$$p(m_t \mid \alpha, y_{t-1}) = \binom{y_{t-1}}{m_t}\alpha^{m_t}(1-\alpha)^{y_{t-1} - m_t}.$$

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

- It is possible to show that using these conditional distributions, we recover the original likelihood.

Full conditionals

The full conditional distributions are simply derived:

$$(\alpha \mid \ldots) \sim \text{Beta}\left(a_0^{(\alpha)} + \sum_{t=2}^{T} m_t, b_0^{(\alpha)} + \sum_{t=2}^{T}(y_{t-1} - m_t)\right)$$

$$(w \mid \ldots) \sim \text{Beta}\left(a_0^{(w)} + \sum_{t=2}^{T} u_t, b_0^{(w)} + (T-1) - \sum_{t=2}^{T} u_t\right)$$

$$(\theta \mid \ldots) \sim \text{Beta}\left(a_0^{(\theta)} + \sum_{t=2}^{T} u_t, b_0^{(\theta)} + \sum_{\{t:u_t=1\}} (y_t - m_t)\right)$$

$$(\lambda \mid \ldots) \sim \text{Gamma}\left(a_0^{(\lambda)} + \sum_{\{t:u_t=0\}} (y_t - m_t), b_0^{(\lambda)} + (T-1) - \sum_{t=2}^{T} u_t\right)$$

Full conditionals

Additionally,

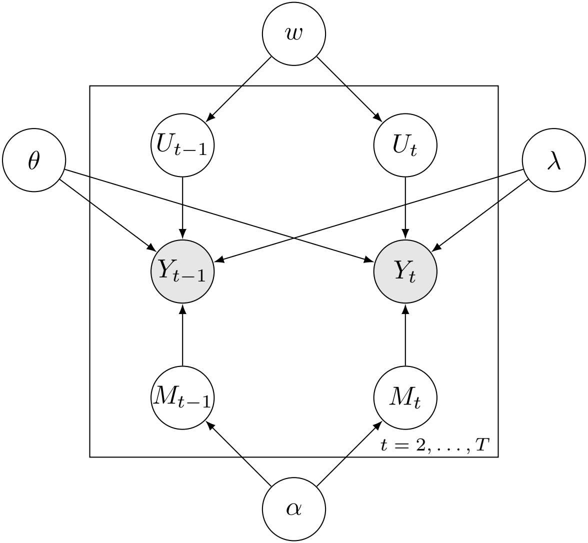
$$\Pr\{U_t = 1 \mid \ldots\} \propto w\ \theta(1-\theta)^{y_t - m_t};$$
$$\Pr\{U_t = 0 \mid \ldots\} \propto (1-w)\ \frac{e^{-\lambda}\lambda^{y_t - m_T}}{(y_t - m_t)!},$$

and

$$\Pr\{M_t = m_t \mid \ldots\}$$
$$\propto \begin{cases} \dfrac{1}{(y_{t-1} - m_t)!\ m_t!} \left(\dfrac{\alpha}{(1-\theta)(1-\alpha)}\right)^{m_t} & \text{if } u_t = 1 \\ \dfrac{1}{(y_t - m_t)!\ (y_{t-1} - m_t)!\ m_t!} \left(\dfrac{\alpha}{\lambda\ (1-\alpha)}\right)^{m_t} & \text{if } u_t = 0 \end{cases}$$

for $t = 2, \ldots, T,\ m_t = 0, 1, \ldots, \min\{y_t, y_{t-1}\}$.

Direct acyclic graph



- This mixture distribution allows the model to account for overdispersion in a time series of counts and accommodate inflation of zeros.
- In what follows, we extend the 2-component mixture of distributions by a generalized, DP-based version of the INAR(1) model.

Learning the latent pattern of heterogeneity in time series of counts

The Dirichlet Process

Given a measurable space $(\mathscr{X}, \mathscr{B})$ and a probability space $(\Omega, \mathscr{F}, \Pr)$, a random probability measure $\mathbb{G}$ is a mapping $\mathbb{G} : \mathscr{B} \times \Omega \to [0, 1]$.

Definition (Ferguson, 1973): Let α be a finite non-null measure on $(\mathscr{X}, \mathscr{B})$. We say $\mathbb{G}$ is a Dirichlet process if, for every measurable partition $\{B_1, \ldots, B_k\}$ of $\mathscr{X}$, the random vector $(P(B_1), \ldots, P(B_k))$ follows a Dirichlet distribution with parameter vector $(\alpha(B_1), \ldots, \alpha(B_k))$.

Let $\tau = \alpha(\mathscr{X})$ be the concentration parameter and, for every $B \in \mathscr{B}$, $G_0(B) = \alpha(B)/\alpha(\mathscr{X})$ the base measure which leads to a suitable parametrization in terms of a probability measure. Under this formulation, we denote $\mathbb{G} \sim DP(\tau\, G_0)$.

The Dirichlet Process

1. $\mathrm{E}(\mathbb{G}(B)) = G_0(B)$.

2. $\mathrm{Var}(\mathbb{G}(B)) = \frac{G_0(B)(1-G_0(B))}{\tau+1}$.

3. Assume that, given a Dirichlet process $\mathbb{G}$ with parameter α, $X_1, \ldots, X_n$ are conditionally independent and identically distributed such that $P(X_i \in B \mid \mathbb{G}) = \mathbb{G}(B) \quad i = 1, \ldots, n$, then $\mathbb{G} \mid X_1, \ldots, X_n \sim DP(\beta)$, where $\beta(C) = \alpha(C) + \sum_{i=1}^{n} \mathbb{I}_C(X_i)$.

4. As shown by [Blackwell and MacQueen, 1973] the predictive distribution of X_{n+1}, $n \geq 1$, given $X_1, \ldots, X_n$ may be obtained integrating out $\mathbb{G}$, which entails that

$$X_{n+1} \mid X_1, \ldots, X_n \sim \frac{\tau}{\tau + n} G_0 + \frac{1}{\tau + n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes a point mass on x.

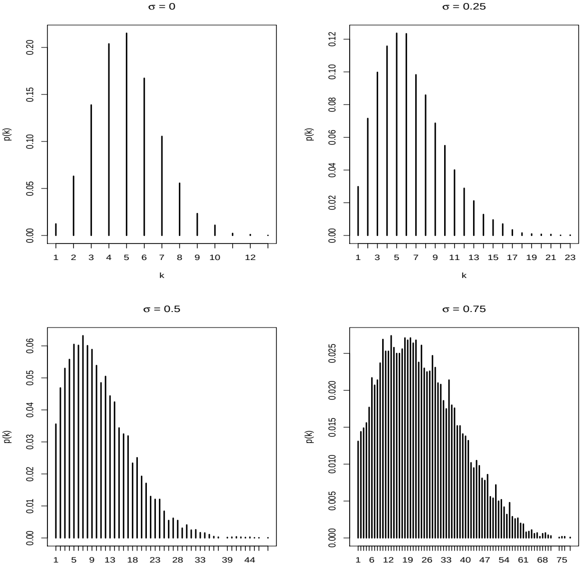
Dirichlet and Pitman-Yor Processes

- The discrete parcel in the predictive distribution implies the clustering property of the Dirichlet process, which induces a probability distribution on the number of distinct values in $(X_1, \dots, X_n)$, which we denote by k.
- [Pitman and Yor, 1997] generalized the Dirichlet process introducing a discount parameter σ, The predictive distribution for the Pitman-Yor process is given by:

$$X_{n+1} \mid X_1, \dots, X_n \sim \frac{\tau + k\sigma}{\tau + n} G_0 + \frac{1}{\tau + n} \sum_{i=1}^{n} \left(1 - \frac{\sigma}{n_i}\right) \delta_{X_i},$$

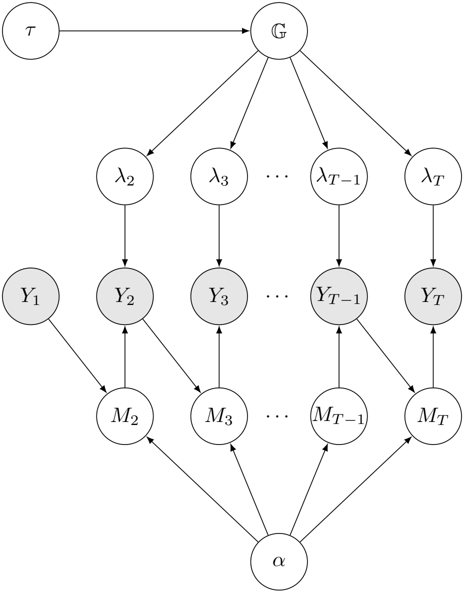
where n_i is the number of elements in $(X_1, \dots, X_n)$ equal to X_i, $\sigma \in [0, 1]$.

- The Pitman-Yor process with high σ induces less informative prior distributions for K [Pitman and Yor, 1997, De Blasi et al., 2013].



- In the INAR(1) structure, we now assume the innovation process is time-varying, i.e., $E(Z_t) = \lambda_t$.
- From a realization of the process $y_1, \ldots, y_T$, we want to learn the distribution of each λ_t and represent our uncertainties about the future steps $Y_{T+1}, \ldots, Y_{T+h}$ in order to forecast them.
- We create clusters of innovation rates as a means to learn the latent patterns of heterogeneity in the count time series.

DAG



τ
$\mathbb{G}$
λ_2
λ_3
$\cdots$
λ_{T-1}
λ_T
Y_1
Y_2
Y_3
$\cdots$
Y_{T-1}
Y_T
M_2
M_3
$\cdots$
M_{T-1}
M_T
α

Let $y = (y_1, \ldots, y_T)$ and $m = (m_2, \ldots, m_T)$. To obtain the posterior $p(\alpha, \lambda, m)$ we integrate out the random distribution P. From the parametric part in the graph, we have that:

$$p(y, m, \alpha, \lambda) = \int p(y, m, \alpha, \lambda \mid G) \, d\mu_{\mathbb{G}}(G)$$

$$= \left\{ \prod_{t=2}^{T} p(y_t \mid m_t, \lambda_t) \; p(m_t \mid y_{t-1}, \alpha) \right\} \times$$

$$\pi(\alpha) \times \int \prod_{t=2}^{T} p(\lambda_t \mid G) \, d\mu_{\mathbb{G}}(G).$$

- The random vector $(\lambda_2, \ldots, \lambda_T)$ has an exchangeable distribution.

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

$$\lambda_t \mid \text{all others} \sim w_0 \times \text{Gamma}(y_t - m_t + a_0^{(G_0)}, b_0^{(G_0)} + 1) + \sum_{r \neq t} \lambda_r^{y_t - m_t} e^{-\lambda_r} \delta_{\{\lambda_r\}},$$

in which $w_0 = \frac{\tau \cdot (b_0^{(G_0)})^{a_0^{(G_0)}} \Gamma(y_t - m_t + a_0^{(G_0)})}{\Gamma(a_0^{(G_0)})(b_0^{(G_0)} + 1)^{y_t - m_t + a_0^{(G_0)}}}$ and $\delta_{\{\lambda_r\}}$ denotes a point mass at λ_r.

- Recall that the full conditional of λ_t is a combination of the joint prior $p(\lambda_2, \ldots, \lambda_T)$ with $p(y_t \mid m_t, \lambda_t)$.
- The weights in the expression above are not normalized.

Choice of prior parameters

[Dorazio, 2009] choose the parameters $a_0^{(\tau)}$ and $b_0^{(\tau)}$ 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.

The marginal probability function of K can be computed as

$$\pi(k) = \int_0^\infty \Pr\{K = k \mid \tau\}\, \pi(\tau)\, d\tau = \frac{b_0^{(\tau)} S(T-1, k)}{\Gamma(a_0^{(\tau)})} I(a_0^{(\tau)}, b_0^{(\tau)}; k),$$

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

$$I(a_0^{(\tau)}, b_0^{(\tau)}; k) = \int_0^\infty \frac{\tau^{k+a_0^{(\tau)}-1}\ e^{-b_0^{(\tau)}\tau}\ \Gamma(\tau)}{\Gamma(\tau + T - 1)}\, d\tau.$$

Choice of prior parameters

Let q be the probability function of a uniform discrete distribution on $\{1, \ldots, T-1\}$, that is

$$q(k) = \frac{1}{T-1} \mathbb{I}_{\{1,\ldots,T-1\}}(k),$$

we find, by numerical integration and optimization, the values of $a_0^{(\tau)}$ and $b_0^{(\tau)}$ that minimize the Kullback-Leibler divergence

$$\mathrm{KL}[\pi \parallel q] = \sum_{k=1}^{T-1} q(k) \log\left(\frac{q(k)}{\pi(k)}\right).$$

Choice of prior parameters

Similarly, we choose the hyperparameters $a_0^{(G_0)}$ and $b_0^{(G_0)}$ of the base probability density g_0 minimizing the Kullback-Leibler divergence between g_0 and a uniform distribution on a suitable range $[0, \lambda_{\max}]$, where $\lambda_{\max}$ is chosen by taking into consideration the available information on the studied phenomena.

Choice of prior parameters

Let h be a uniform density on $[0, \lambda_{\text{max}}]$, that is

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

we find, by numerical optimization, the values of $a_0^{(G_0)}$ and $b_0^{(G_0)}$ that minimize the Kullback-Leibler divergence

$$\begin{aligned}\text{KL}[g_0 \parallel h] &= \int_0^{\lambda_{\text{max}}} \left(\frac{1}{\lambda_{\text{max}}}\right) \log\left(\frac{1/\lambda_{\text{max}}}{g_0(\lambda)}\right) d\lambda \\ &= -\log\lambda_{\text{max}} - a_0^{(G_0)} \log b_0^{(G_0)} + \log\Gamma(a_0^{(G_0)}) - \\ &(a_0^{(G_0)} - 1)(\log\lambda_{\text{max}} - 1) + \frac{b_0^{(G_0)}\lambda_{\text{max}}}{2}.\end{aligned}$$

Choosing the parameters for the α prior is more straightforward, with $a_0^{(\alpha)} = b_0^{(\alpha)} = 1$ being a natural choice.

Pitman-Yor case

The full conditional of each λ_t is slightly modified:

$$\lambda_t \mid \text{all others} \sim w_0^* \times \text{Ga}(y_t - m_t + a_0^{(G_0)}(b_0^{(G_0)} + 1) + \sum_{i \neq t} \left(1 - \frac{\sigma}{n_i}\right) \lambda_i^{y_t - m_t} e^{-\lambda_i} \delta_{\{\lambda_i\}},$$

$$w_0^* = \frac{(\tau + k_{\setminus t}\,\sigma) \cdot (b_0^{(G_0)})^{a_0^{(G_0)}} \Gamma(y_t - m_t + a_0^{(G_0)})}{\Gamma(a_0^{(G_0)})(b_0^{(G_0)} + 1)^{y_t - m_t + a_0^{(G_0)}}}.$$

- To improve efficiency, we remix the vector of distinct rates λ^* after every step of the sampler [Escobar and West, 1998].
- Let $(\lambda_1^*, \ldots, \lambda_k^*)$ be the k unique values among $(\lambda_2, \ldots, \lambda_T)$. Let $c_t = \sum_{j=1}^k j \cdot \mathbb{I}_{\{\lambda_j^*\}}(\lambda_t)$ be the cluster indicator of λ_t, and define the number of occupants of cluster j by $n_j = \sum_{t=2}^T \mathbb{I}_{\{j\}}(c_t)$:

$$\lambda_j^* \mid \text{all others} \sim \text{Gamma}\left(a_0^{(G_0)} + \sum_{\substack{t=2 \\ c_t=j}}^{T} (y_t - m_t), b_0^{(G_0)} + n_j \right).$$

for $j = 1, \ldots, k$.

- Also, the full conditionals for α and m_t are:

$$\alpha \mid \text{all others} \sim \text{Beta}\left(a_0^{(\alpha)} + \sum_{t=2}^{T} m_t, b_0^{(\alpha)} + \sum_{t=2}^{T} (y_{t-1} - m_t) \right).$$

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

- This Gibbs sampler yields, marginally, a sample $\{\alpha^{(n)}, \lambda^{(n)}\}_{n=1}^{N}$ from the posterior distribution.

The DP-INAR(1) Model

We extend [Freeland, 1998] original INAR(1) model.

Proposition

The probability function of Y_{t+h} given $Y_t = y_t$ and $\theta = (\alpha, \lambda_{t+1}, \ldots, \lambda_{t+h})$, can be writen as the convolution of a $Bin(y_t, \alpha^h)$ distribution and a $Poisson(\mu_h)$ distribution.

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

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

Pólya-Blackwell-MacQueen urn

Using [Blackwell and MacQueen, 1973] urn process recursively, for $n = 1 \ldots, N$, we draw a sample $\{\lambda_{T+1}^{(n)}, \ldots, \lambda_{T+h}^{(n)}\}_{n=1}^{N}$ from $\prod_{i=1}^{h} p(\lambda_{T+i} \mid \lambda_2, \ldots, \lambda_{T+i-1})$ sequentially as follows:

$$\lambda_{T+1}^{(n)} \sim \frac{\tau}{\tau + T} G_0 + \frac{1}{\tau + T} \sum_{t=2}^{T} \delta_{\{\lambda_t^{(n)}\}};$$

$$\lambda_{T+2}^{(n)} \sim \frac{\tau}{\tau + T + 1} G_0 + \frac{1}{\tau + T + 1} \sum_{t=2}^{T+1} \delta_{\{\lambda_t^{(n)}\}};$$

$$\vdots$$

$$\lambda_{T+h}^{(n)} \sim \frac{\tau}{\tau + T + h - 1} G_0 + \frac{1}{\tau + T + h - 1} \sum_{t=2}^{T+h-1} \delta_{\{\lambda_t^{(n)}\}}.$$

DP-INAR(1) posterior predictive

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

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

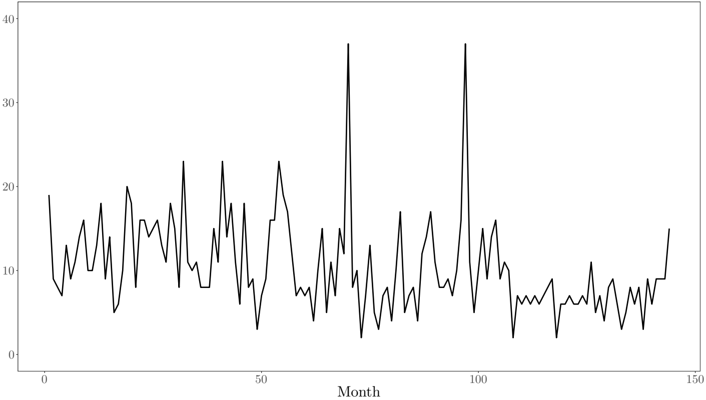
for $y_{T+h} \geq 0$.

- As a pointwise forecast, we compute the generalized median of the h-steps-ahead posterior predictive distribution, defined by

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

Monthly counts of burglary events

We analyze monthly time series of burglary events in Pittsburgh, USA, from January 1990 to December 2001. In this dataset, each time series has a length of 144 months and corresponds to a certain patrol area. The Figure below shows the series for patrol area 58.



http://www.forecastingprinciples.com/index.php/crimedata

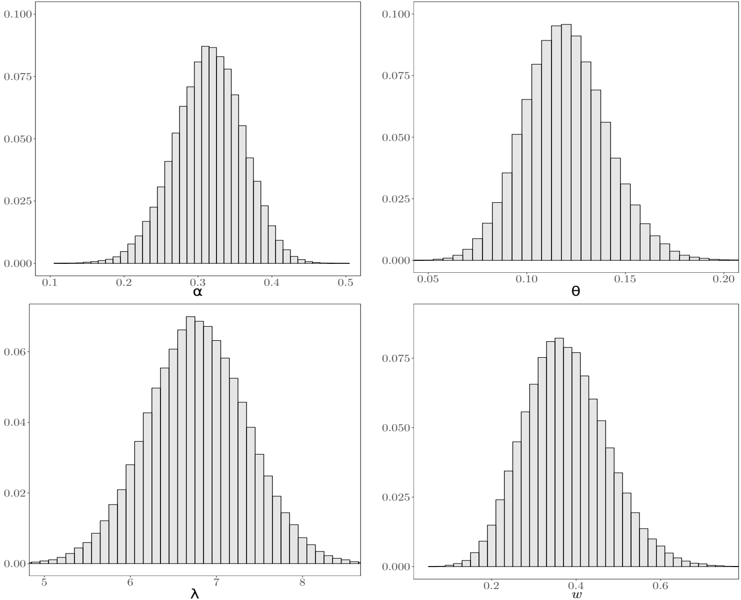
Posterior distributions - AdINAR(1)

For the AdINAR(1) model hyperparameters, we make the choices $a_\alpha = 1$, $b_\alpha = 1$, $a_\lambda = 1$, $b_\lambda = 0.01$, $a_\theta = 1$, $b_\theta = 1$, $a_w = 1$, and $b_w = 1$, which correspond to reasonably flat priors.

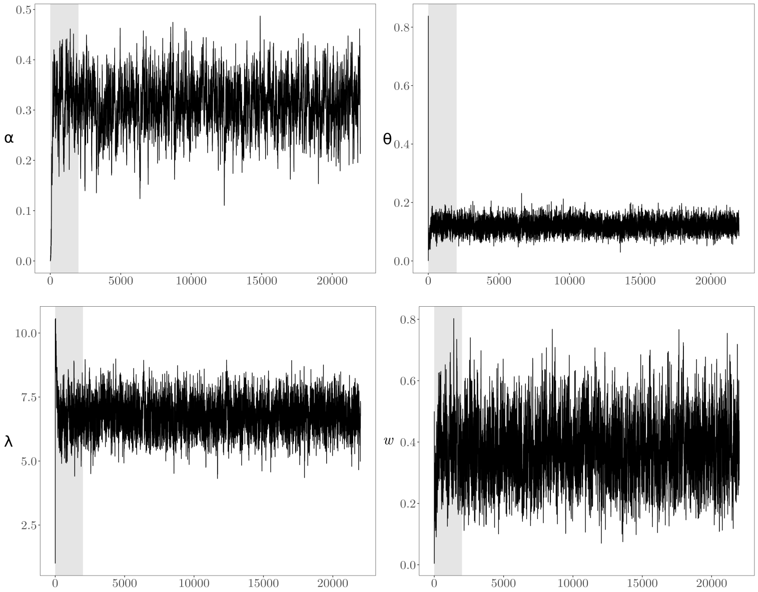
Figure 2 displays the marginal posterior distributions of the AdINAR(1) model parameters.

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.

Posterior distributions - AdINAR(1)



Posterior distributions - AdINAR(1)



Posterior distributions - DP-INAR(1)

For the DP-INAR(1) model, we specify the hyperparameters as follows. To determine a_τ and b_τ, the optimization procedure described, 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.

We control the support of G_0 by choosing the value of $\lambda_{\max}$ to be the maximum observed count. The level curves of $\mathrm{KL}[g_0 \parallel h]$ are given below. The minimum is attained at $a_0^{(G_0)} = 1.778$ and $b_0^{(G_0)} = 0.096$.

For the thinning parameter α, we adopt a uniform prior, choosing $a_0^{(\alpha)} = b_0^{(\alpha)} = 1$.

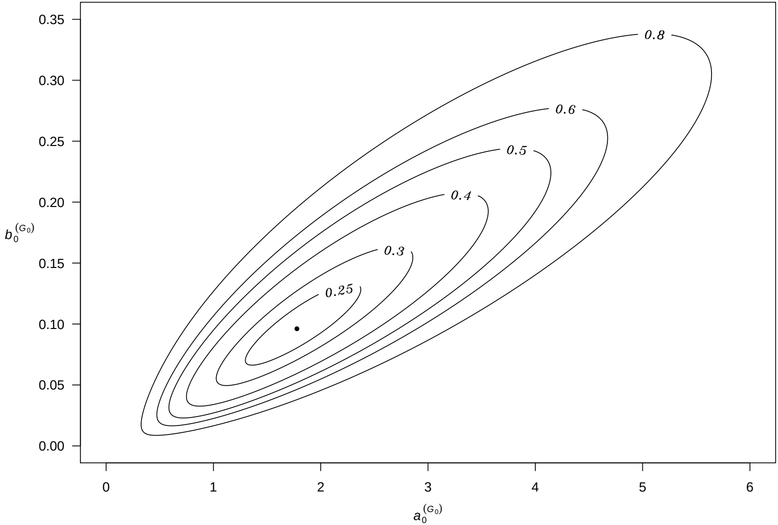


Figure: Level curves of the Kullback-Leibler divergence associated with the optimization of the base measure hyperparameters for patrol area 58.

Posterior distributions - DP-INAR(1)

The marginal posterior distributions of parameters α, λ_3, λ_{18}, and λ_{96} are displayed below. The posterior distribution of the thinning parameter α is reasonably concentrated, with posterior mean 0.19, showing that the autoregressive component is not negligible. The posterior distributions of λ_3, λ_{18} and λ_{96} are fairly concentrated as well, with posterior means equal to 6.50, 13.61 and 32.01, respectively, showing that different regimes of innovation rates were captured in the learning process.

Posterior distributions - DP-INAR(1)

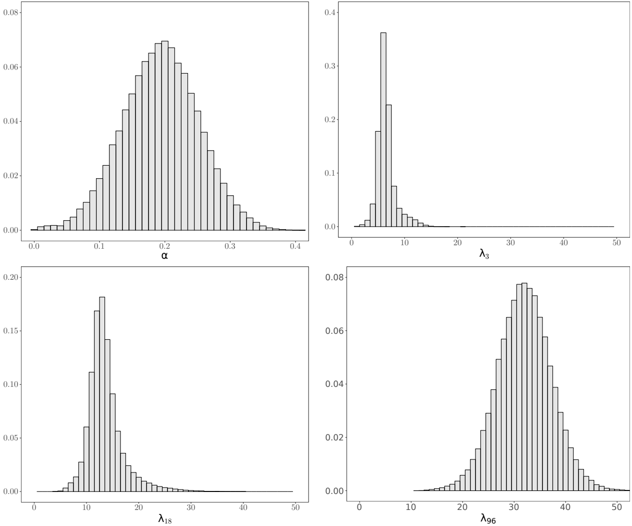


Figure: Marginal posterior distributions of parameters α, λ_3, λ_{18}, and λ_{96}, for patrol area 58.

Posterior distributions - DP-INAR(1)

The posterior means of the innovation rates follow the same pattern of heterogeneity of the series.

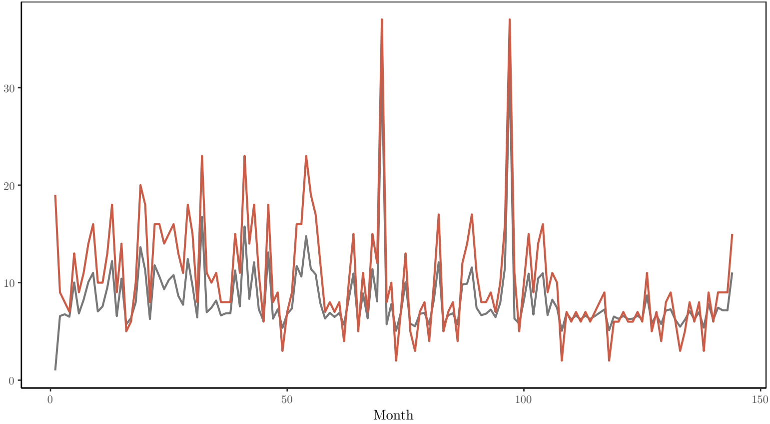


Figure: Posterior means of the innovation rates (in grey) and the observed counts for patrol area 58 (in red).

The Markov chains in Figure 7 indicate that proper mixing is achieved by the Gibbs sampler.

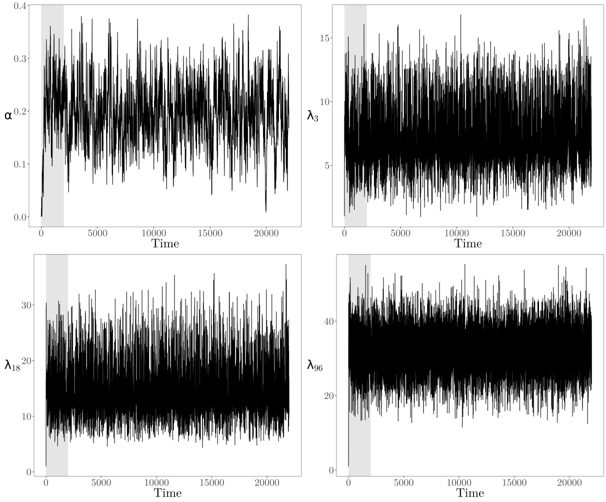


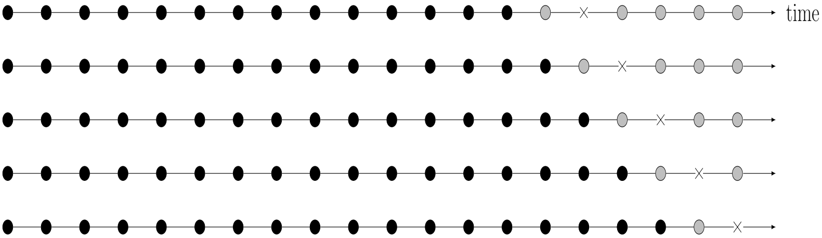
Figure: Markov chains associated with the marginal posterior distributions of parameters α, λ_3, λ_{18}, and λ_{96}, for patrol area 58. The gray rectangles indicate the burn-in periods.

Cross-validation procedure for time series

- We use a form of cross-validation to evaluate the forecasting performance of the model.
- For an observed time series $y_1, \ldots, y_T$, pick some $T^* < T$:

1. Treat the counts $y_{T^*}, \ldots, y_T$ as a test sample
2. For $t \geq T^*$,train the model using the values of $y_1, \ldots, y_{t-1}$ making an h-steps-ahead prediction $\hat{y}_{t+h}$.
3. Compute the median deviation $|\hat{y}_{t+h} - y_{t+h}|$
4. Take the average over all predictions.

Illustration of the 2-step ahead forecasting



Forecasting performance

In terms of forecasting performance:

- The AdINAR(1) and the DP-INAR(1) models outperform the INAR(1) model in 75% of the patrol areas
- 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.

- In terms of learning K, we fixed τ such that the prior expected number of clusters is equal to $\{3, 8, 13\}$.
- The Dirichlet Process is highly influenced by the prior specification of τ.
- Robustness is achieved in the Pitman-Yor case specifying high values for σ.

	DP $\sigma = 0$	Pitman-Yor σ		
	0	0.25	0.50	0.75
$E(K) = 3$	4.67	5.76	6.41	7.44
$E(K) = 8$	9.37	8.37	7.89	7.31
$E(K) = 13$	13.93	10.76	9.25	7.15

Future work

Future work

- Extend the DP-INAR(1) and PY-INAR(1) to higher order Markov process.
- Generalize these ideas for multivariate time series by using Hierarchical Dirichlet Processes.
- Incorporate covariates in the model.
- Multivariate extensions (work with Refik Soyer)
- Dynamic modeling extensions (with Refik as well)

References I

Alzaid, A. and Al-Osh, M. (1987).
First-order integer-valued autoregressive (inar(1)) processes.
Journal of Time Series Analysis.

Blackwell, D. and MacQueen, J. B. (1973).
Ferguson distributions via polya urn schemes.
Annals of Statistics, 1(2).

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013).
Are gibbs-type priors the most natural generalization of the dirichlet process?
IEEE transactions on pattern analysis and machine intelligence, 37(2):212–229.

References II

Dorazio, R. M. (2009).
On selecting a prior for the precision parameter of dirichlet process mixture models.
Journal of Statistical Planning and Inference.

Escobar, M. and West, M. (1998).
Computing nonparametric hierarchical models.
In Dey, D., Müller, P., and Sinha, D., editors, *Practical nonparametric and semiparametric Bayesian statistics*, chapter 1, pages 1–22. Springer-Verlag.

Freeland, R. K. (1998).
Statistical analysis of discrete time series with application to the analysis of workers' compensation data.
PhD thesis, University of British Columbia, Vancouver.

References III

McKenzie, E. (1985).
Some simple models for discrete variate time series.
JAWRA Journal of the American Water Resources Association, 21(4):645–650.

Pitman, J. and Yor, M. (1997).
The two-parameter poisson-dirichlet distribution derived from a stable subordinator.
Annals of Probability, 25:855–900.