entropy

MDPI

Article

Prior Sensitivity Analysis in a Semi-Parametric Integer-Valued Time Series Model

Helton Graziadei [1,*], Antonio Lijoi [2], Hedibert F. Lopes [3] and Paulo C. Marques F. [3] and Igor Prünster [2]

1 Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo 05508-090, Brazil
2 Department of Decision Sciences and BIDSA, Bocconi University, via Röntgen 1, 20136 Milano, Italy; antonio.lijoi@unibocconi.it (A.L.); igor.pruenster@unibocconi.it (I.P.)
3 Insper Institute of Education and Research, Rua Quatá 300, São Paulo 04546-042, Brazil; HedibertFL@insper.edu.br (H.F.L.); PauloCMF1@insper.edu.br (P.C.M.F.)
* Correspondence: hltgraziadei@gmail.com

Received: 26 November 2019; Accepted: 3 January 2020; Published: 6 January 2020

Abstract: We examine issues of prior sensitivity in a semi-parametric hierarchical extension of the INAR(p) model with innovation rates clustered according to a Pitman–Yor process placed at the top of the model hierarchy. Our main finding is a graphical criterion that guides the specification of the hyperparameters of the Pitman–Yor process base measure. We show how the discount and concentration parameters interact with the chosen base measure to yield a gain in terms of the robustness of the inferential results. The forecasting performance of the model is exemplified in the analysis of a time series of worldwide earthquake events, for which the new model outperforms the original INAR(p) model.

Keywords: time series of counts; Bayesian hierarchical modeling; Bayesian nonparametrics; Pitman–Yor process; prior sensitivity; clustering; Bayesian forecasting

1. Introduction

Integer-valued time series are relevant to many fields of knowledge, ranging from finance and econometrics to ecology and meteorology. An extensive number of models for this kind of data has been proposed since the introduction of the INAR(1) model in the pioneering works of McKenzie [1] and Al-Osh and Alzaid [2] (see also the book by Weiss [3]). A higher-order INAR(p) model was considered in the work of Du and Li [4].

In this paper, we generalize the Bayesian version of the INAR(p) model studied by Neal and Kypraios [5]. In our model, the innovation rates are allowed to vary through time, with the distribution of the innovation rates being modeled hierarchically by means of a Pitman–Yor process [6]. In this way, we account for potential heterogeneity in the innovation rates as the process evolves through time, and this feature is automatically incorporated in the Bayesian forecasting capabilities of the model.

The semi-parametric form of the model demands a robustness analysis of our inferential conclusions as we vary the hyperparameters of the Pitman–Yor process. We investigate this prior sensitivity issue carefully and find ways to control the hyperparameters in order to achieve robust results.

This paper is organized as follows. In Section 2, we construct a generalized INAR(p) model with variable innovation rates. The likelihood function of the generalized model is derived and a data augmentation scheme is developed, which gives a specification of the model in terms of conditional distributions. This data augmented representation of the model enables the derivation in Section 4 of full conditional distributions in simple analytical form, which are essential for the stochastic simulations

in Section 5. Section 3 recollects the main properties of the Pitman–Yor process which are used to define the PY-INAR(p) model in Section 4, including its clustering properties. In building the PY-INAR(p), we propose a form for the prior distribution of the thinning parameters vector which improves on the choice made for the Bayesian INAR(p) model studied in [5]. In Section 5, we investigate the robustness of the inference with respect to changes in the Pitman–Yor process hyperparameters. Using the full conditional distributions of the innovation rates derived in Section 4, we inspect the behavior of the model as we concentrate or spread the mass of the Pitman–Yor base measure. This leads us to a graphical criterion that identifies an elbow in the posterior expectation of the number of clusters as we vary the hyperparameters of the base measure. Once we have control over the base measure, we study its interaction with the concentration and discount hyperparameters, showing how to make choices that yield robust results. In the course of this development, we use geometrical tools to inspect the clustering of the innovation rates produced by the model. Section 6 puts the graphical criterion to work for simulated data. In Section 7, using a time series of worldwide earthquake events, we finish the paper comparing the forecasting performance of the PY-INAR(p) model against the original INAR(p) model, with favorable results.

2. A Generalization of the INAR(p) Model

We begin by generalizing the original INAR(p) model of Du and Li [4] as follows.

Let $\{Y_t\}_{t\geq 1}$ be an integer-valued time series, and, for some integer $p \geq 1$, let the *innovations* $\{Z_t\}_{t\geq p+1}$, given positive parameters $\{\lambda_t\}_{t\geq p+1}$, be a sequence of conditionally independent Poisson(λ_t) random variables. For a given vector of parameters $\alpha = (\alpha_1, \ldots, \alpha_p) \in [0,1]^p$, let $\mathscr{F}_i = \{B_{ij}(t) : j \geq 0, t \geq 2\}$ be a family of conditionally independent and identically distributed Bernoulli(α_i) random variables. For $i \neq k$, suppose that $\mathscr{F}_i$ and $\mathscr{F}_k$ are conditionally independent, given α. Furthermore, assume that the innovations $\{Z_t\}_{t\geq p+1}$ and the families $\mathscr{F}_1, \ldots, \mathscr{F}_p$ are conditionally independent, given α and λ. The generalized INAR(p) model is defined by the functional relation

$$Y_t = \alpha_1 \circ Y_{t-1} + \cdots + \alpha_p \circ Y_{t-p} + Z_t,$$

for $t \geq p+1$, in which $\circ$ denotes the binomial thinning operator, defined by $\alpha_i \circ Y_{t-i} = \sum_{j=1}^{Y_{t-i}} B_{ij}(t)$, if $Y_{t-i} > 0$, and $\alpha_i \circ Y_{t-i} = 0$, if $Y_{t-i} = 0$. In the homogeneous case, when all the λ_t's are assumed to be equal, we recover the original INAR(p) model.

When $p = 1$, 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 $t-1$ (see [7] for an interpretation of the order p case as a birth-and-death process with immigration).

Let $y = (y_1, \ldots, y_T)$ denote the values of an observed time series. For simplicity, we assume that $Y_1 = y_1, \ldots, Y_p = y_p$ with probability one. The joint distribution of $Y_1, \ldots, Y_T$, given parameters α and $\lambda = (\lambda_{p+1}, \ldots, \lambda_T)$, can be factored as

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

Since, with probability one, $\alpha_i \circ Y_{t-i} \leq Y_{t-i}$ and $Z_t \geq 0$, the likelihood function of the generalized INAR(p) model is given by

$$L_y(\alpha,\lambda)=\prod_{t=p+1}^{T}\sum_{m_{1,t}=0}^{\min\{y_t,y_{t-1}\}}\cdots\sum_{m_{p,t}=0}^{\min\{y_t-\sum_{j=1}^{p-1}m_{j,t},\,y_{t-p}\}}\left(\prod_{i=1}^{p}\binom{y_{t-i}}{m_{i,t}}\alpha_i^{m_{i,t}}(1-\alpha_i)^{y_{t-i}-m_{i,t}}\right)\times\left(\frac{e^{-\lambda_t}\lambda_t^{\,y_t-\sum_{j=1}^{p}m_{j,t}}}{(y_t-\sum_{j=1}^{p}m_{j,t})!}\right).$$

For some epoch t and $i=1,\ldots,p$, suppose that we could observe the values of the latent *maturations* $M_{i,t}$. Postulate that $M_{i,t}\mid Y_{t-i}=y_{t-i},\alpha_i\sim\text{Binomial}(y_{t-i},\alpha_i)$, so that the conditional probability function of $M_{i,t}$ is given by

$$\begin{aligned}p(m_{i,t}\mid y_{t-i},\alpha_i)&=\Pr\{M_{i,t}=m_{i,t}\mid Y_{t-i}=y_{t-i},\alpha_i\}\\&=\binom{y_{t-i}}{m_{i,t}}\alpha_i^{m_{i,t}}(1-\alpha_i)^{y_{t-i}-m_{i,t}}\,\mathbb{I}_{\{0,\ldots,y_{t-i}\}}(m_{i,t}).\end{aligned}$$

Furthermore, suppose that

$$\begin{aligned}p(y_t\mid m_{1,t},\ldots,m_{p,t},\lambda_t)&=\Pr\{Y_t=y_t\mid M_{1,t}=m_{1,t},\ldots,M_{p,t}=m_{p,t},\lambda_t\}\\&=\frac{e^{-\lambda_t}\lambda_t^{\,y_t-\sum_{j=1}^{p}m_{j,t}}}{(y_t-\sum_{j=1}^{p}m_{j,t})!}\,\mathbb{I}_{\{\sum_{j=1}^{p}m_{j,t},\,\sum_{j=1}^{p}m_{j,t}+1,\ldots\}}(y_t).\end{aligned}$$

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

$$\begin{aligned}p(y_t\mid y_{t-1},\ldots,y_{t-p},\alpha,\lambda_t)&=\sum_{m_{1,t}=0}^{y_{t-1}}\cdots\sum_{m_{p,t}=0}^{y_{t-p}}p(y_t,m_{1,t},\ldots,m_{p,t}\mid y_{t-1},\ldots,y_{t-p},\alpha,\lambda_t)\\&=\sum_{m_{1,t}=0}^{y_{t-1}}\cdots\sum_{m_{p,t}=0}^{y_{t-p}}p(y_t\mid m_{1,t},\ldots,m_{p,t},\lambda_t)\times\prod_{i=1}^{p}p(m_{i,t}\mid y_{t-i},\alpha_i).\end{aligned}$$

Since

$$\begin{aligned}\mathbb{I}_{\{\sum_{j=1}^{p}m_{j,t},\,\sum_{j=1}^{p}m_{j,t}+1,\ldots\}}(y_t)&=\mathbb{I}_{\{0,\ldots,y_t\}}\left(\textstyle\sum_{j=1}^{p}m_{j,t}\right)\\&=\mathbb{I}_{\{0,\ldots,y_t\}}(m_{1,t})\times\cdots\times\mathbb{I}_{\{0,\ldots,y_t-\sum_{j=1}^{p-1}m_{j,t}\}}(m_{p,t})\end{aligned}$$

and

$$\mathbb{I}_{\{\sum_{j=1}^{p}m_{j,t},\,\sum_{j=1}^{p}m_{j,t}+1,\ldots\}}(y_t)\times\mathbb{I}_{\{0,\ldots,y_{t-i}\}}(m_{i,t})=\mathbb{I}_{\{0,1,\ldots,\min\{y_t-\sum_{j\neq i}m_{j,t},\,y_{t-i}\}\}}(m_{i,t}),$$

we recover the original likelihood of the generalized INAR(p), showing that the introduction of the latent maturations $M_{i,t}$ with the specified distributions is a valid data augmentation scheme (see [8,9] for a general discussion of data augmentation techniques).

In the next section, we review the needed definitions and properties of the Pitman–Yor process.

3. Pitman–Yor Process

Let the random probability measure $\mathbb{G}\sim\text{DP}(\tau,G_0)$ be a Dirichlet process [10–12] with concentration parameter τ and base measure G_0. If the random variables $X_1,\ldots,X_n$, given $\mathbb{G}=G$, are conditionally independent and identically distributed as G, then it follows that

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

for every Borel set B. If we imagine the sequential generation of the X_i's, for $i = 1, \ldots, n$, the former expression shows 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. Therefore, almost surely, realizations of a Dirichlet process are discrete probability measures, perhaps with denumerable infinite support, depending on the nature of G_0. Also, this data-generating process, known as the Pólya–Blackwell–MacQueen urn, implies that the X_i's are "softly clustered", in the sense that in one realization of the process the elements of a subset of the X_i's may have exactly the same value.

The Pitman–Yor process [6] is a generalization of the Dirichlet process which results in a model with added flexibility. Essentially, the Pitman–Yor process modifies the expression of the probability associated with the Pólya-Blackwell-MacQueen urn introducing a new parameter so that the posterior predictive probability becomes

$$\Pr\{X_{n+1} \in B \mid X_1 = x_1, \ldots, X_n = x_n\} = \frac{\tau + k\sigma}{\tau + n} G_0(B) + \frac{1}{\tau + n} \sum_{i=1}^{n} \left(1 - \frac{\sigma}{n_i}\right) I_B(x_i),$$

in which $0 \leq \sigma < 1$ is the discount parameter, $\tau > -\sigma$, k is the number of distinct elements in $\{X_1, \ldots, X_n\}$, and n_i is the number of elements in $\{X_1, \ldots, X_n\}$ which are equal to X_i, for $i = 1, \ldots, n$. It is well known that $\mathrm{E}[\mathbb{G}(B)] = G_0(B)$ and

$$\mathrm{Var}[\mathbb{G}(B)] = \left(\frac{1 - \sigma}{\tau + 1}\right) G_0(B)(1 - G_0(B)),$$

for every Borel set B. Hence, $\mathbb{G}$ is centered on the base probability measure G_0, while τ and σ control the concentration of $\mathbb{G}$ around G_0. We use the notation $\mathbb{G} \sim \mathrm{PY}(\tau, \sigma, G_0)$. When $\sigma = 0$, we recover the Dirichlet process as a special case. The PY process is also defined for $\sigma < 0$ and $\tau = |\sigma| m$, for some positive integer m. For our purposes, it is enough to consider the case of non-negative σ.

Pitman [6] derived the distribution of the number of clusters K (the number of distinct X_i's), conditionally on both the concentration parameter τ and the discount parameter σ, as

$$\Pr\{K = k \mid \tau, \sigma\} = \frac{\prod_{i=1}^{k-1} (\tau + i\sigma)}{\sigma^k \times (\tau + 1)_{n-1}} \times \mathscr{C}(n, k; \sigma),$$

in which $(x)_n = \Gamma(x + n)/\Gamma(x)$ is the rising factorial and $\mathscr{C}(n, k; \sigma)$ is the generalized factorial coefficient [13].

In the next section, we use a Pitman–Yor process to model the distribution of the innovation rates in the generalized INAR(p) model.

4. PY-INAR(p) Model

The PY-INAR(p) model is as a hierarchical extension of the generalized INAR(p) model defined in Section 2. Given a random measure $\mathbb{G} \sim \mathrm{PY}(\tau, \sigma, G_0)$, in which G_0 is a Gamma(a_0, b_0) distribution, let the innovation rates $\lambda_{p+1}, \ldots, \lambda_T$ be conditionally independent and identically distributed with distribution $\Pr\{\lambda_t \in B \mid \mathbb{G} = G\} = G(B)$.

To complete the PY-INAR(p) model, we need to specify the form of the prior distribution for the vector of thinning parameters $\alpha = (\alpha_1, \ldots, \alpha_p)$. By comparison with standard results from the theory of the AR(p) model [14], Du and Li [4] found that in the INAR(p) model the constraint $\sum_{i=1}^{p} \alpha_i < 1$ must be fulfilled to guarantee the non-explosiveness of the process. In their Bayesian analysis of the INAR(p) model, Neal and Kypraios [5] considered independent beta distributions for the α_i's. Unfortunately, this choice is problematic. For example, in the particular case when the α_i's have independent uniform distributions, it is possible to show that $\Pr\{\sum_{i=1}^{p} \alpha_i < 1\} = 1/p!$, implying that we would be concentrating most of the prior mass on the explosive region even for moderate values of the model order p. We circumvent this problem using a prior distribution for α that places all of its

mass on the nonexplosive region and still allows us to derive the full conditional distributions of the α_i's in simple closed form. Specifically, we take the prior distribution of α to be a Dirichlet distribution with hyperparameters $(a_1, \ldots, a_p; a_{p+1})$, and corresponding density

$$\pi(\alpha) = \frac{\Gamma\left(\sum_{i=1}^{p+1} a_i\right)}{\prod_{i=1}^{p+1} \Gamma(a_i)} \prod_{i=1}^{p+1} \alpha_i^{a_i - 1},$$

in which $a_i > 0$, for $i = 1, \ldots, p+1$, and $\alpha_{p+1} = 1 - \sum_{i=1}^{p} \alpha_i$.

Let $m = \{m_{i,t} \colon i = 1, \ldots, p, t = p+1, \ldots, T\}$ denote the set of all maturations, and let $\mu_{\mathbb{G}}$ be the distribution of $\mathbb{G}$. Our strategy to derive the full conditionals distributions of the model parameters and latent variables is to consider the marginal distribution

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

From this expression, using the results in Section 3, the derivation of the full conditional distributions is straightforward. In the following expressions, the symbol $\propto$ denotes proportionality up to a suitable normalization factor, and the label "all others" designate the observed counts y and all the other latent variables and model parameters, with the exception of the one under consideration.

Let $\lambda_{\setminus t}$ denote the set $\{\lambda_{p+1}, \ldots, \lambda_T\}$ with the element λ_t removed. Then, for $t = p+1, \ldots, T$, we have

$$\lambda_t \mid \text{all others} \sim w_t \times \text{Gamma}(y_t - m_t + a_0, b_0 + 1) + \sum_{r \neq t} \left(1 - \frac{\sigma}{n_r}\right) \lambda_r^{y_t - m_t} e^{-\lambda_r} \delta_{\{\lambda_r\}},$$

in which the weight

$$w_t = \frac{(\tau + k_{\setminus t}\, \sigma) \times b_0^{a_0} \times \Gamma(y_t - m_t + a_0)}{\Gamma(a_0) \times (b_0 + 1)^{y_t - m_t + a_0}},$$

n_r is the number of elements in $\lambda_{\setminus t}$ which are equal to λ_r, and $k_{\setminus t}$ is the number of distinct elements in $\lambda_{\setminus t}$. In this mixture, we suppressed the normalization constant that makes all weights add up to one.

Making the choice $a_{p+1} = 1$, we have

$$\alpha_i \mid \text{all others} \sim \text{TBeta}\left(a_i + \sum_{t=p+1}^{T} m_{i,t}, 1 + \sum_{t=p+1}^{T} (y_{t-i} - m_{i,t}), 1 - \sum_{j \neq i} \alpha_j\right),$$

for $i = 1, \ldots, p$, in which TBeta denotes the right truncated Beta distribution with support $(0, 1 - \sum_{j \neq i}^{p} \alpha_j)$.

For the latent maturations, we find

$$\begin{aligned} p(m_{i,t} \mid \text{all others}) &\propto \frac{1}{(m_{i,t})!(y_t - \sum_{j=1}^{p} m_{j,t})!(y_{t-i} - m_{i,t})!} \left(\frac{\alpha_i}{\lambda_t(1 - \alpha_i)}\right)^{m_{i,t}} \\ &\quad \times \mathbb{I}_{\{0, 1, \ldots, \min\{y_t - \sum_{j \neq i} m_{j,t}, y_{t-i}\}\}}(m_{i,t}). \end{aligned}$$

To explore the posterior distribution of the model, we build a Gibbs sampler [15] using these full conditional distributions. Escobar and West [16] showed, in a similar context, that we can improve mixing by resampling simultaneously the values of all λ_t's inside the same cluster at the end of each iteration of the Gibbs sampler. Letting $(\lambda_1^*, \ldots, \lambda_k^*)$ be the k unique values among $(\lambda_{p+1}, \ldots, \lambda_T)$, define the number of occupants of cluster j by $\nu_j = \sum_{t=p+1}^{T} \mathbb{I}_{\{\lambda_j^*\}}(\lambda_t)$, for $j = 1, \ldots, k$. It follows that

$$\lambda_j^* \mid \text{all others} \sim \text{Gamma}\left(a_0 + \sum_{t=p+1}^{T}\left(y_t - \sum_{i=1}^{p} m_{i,t}\right) \cdot \mathbb{I}_{\{\lambda_j^*\}}(\lambda_t), b_0 + \nu_j\right).$$

for $j = 1, \ldots, k$. At the end of each iteration of the Gibbs sampler, we update the values of all λ_t's inside each cluster by the corresponding λ_j^* using this distribution.

5. Prior Sensitivity

As it is often the case for Bayesian models with nonparametric components, a choice of the prior parameters for the PY-INAR(p) model which yields robustness of the posterior distribution is nontrivial [17].

The first aspect to be considered is the fact that the base measure G_0 plays a crucial role in the determination of the posterior distribution of the number of clusters K. This can be seen directly by inspecting the form of the full conditional distributions derived in Section 4. Recalling that G_0 is a gamma distribution with mean a_0/b_0 and variance a_0/b_0^2, from the full conditional distribution of λ_t one may note that the probability of generating, on each iteration of the Gibbs sampler, a value for λ_t anew from G_0 is proportional to

$$\frac{(\tau + k_{\setminus t}\,\sigma) \times b_0^{a_0} \times \Gamma(y_t - m_t + a_0)}{\Gamma(a_0)(b_0+1)^{y_t - m_t + a_0}}.$$

Therefore, supposing that all the other terms are fixed, if we concentrate the mass of G_0 around zero by making $b_0 \to \infty$, this probability decreases to zero. This is not problematic, because it is hardly the case that we want to make such a drastic choice for G_0. The behavior in the other direction is more revealing, since taking $b_0 \downarrow 0$, in order to spread the mass of G_0, also makes the limit of this probability to be zero. Due to this behavior, we need to establish a criterion to choose the hyperparameters of the base measure which avoids these extreme cases.

In our analysis, it is convenient to have a single hyperparameter regulating how the mass of G_0 is spread over its support. For a given $\lambda_{\max} > 0$, we find numerically the values of a_0 and b_0 which minimize the Kullback-Leibler divergence between G_0 and a uniform distribution on the interval $[0, \lambda_{\max}]$. This Kullback-Leibler divergence can be computed explicitly as

$$-\log \lambda_{\max} - a_0 \log b_0 + \log \Gamma(a_0) - (a_0 - 1)(\log \lambda_{\max} - 1) + \frac{b_0 \lambda_{\max}}{2}.$$

In this new parameterization, our goal is to make a sensible choice for $\lambda_{\max}$. It is worth emphasizing that by this procedure we are not truncating the support of G_0, but only using the uniform distribution on the interval $[0, \lambda_{\max}]$ as a reference for our choice of the base measure hyperparameters a_0 and b_0.

Our proposal to choose $\lambda_{\max}$ goes as follows. We fix some value $0 \leq \sigma < 1$ for the discount parameter and choose an integer k_0 as the prior expectation of the number of clusters K, which, using the results at the end of Section 3, can be computed explicitly as

$$\mathrm{E}[K] = \begin{cases} \tau \times (\psi(\tau + T - p) - \psi(\tau)) & \text{if} \quad \sigma = 0; \\ ((\tau + \sigma)_{T-p} / (\sigma \times (\tau + 1)_{T-p-1})) - \tau/\sigma & \text{if} \quad \sigma > 0, \end{cases}$$

in which $\psi(x)$ is the digamma function (see [6] for a derivation of this result). Next, we find the value of the concentration parameter τ by solving $\mathrm{E}[K] = k_0$ numerically. After this, for each $\lambda_{\max}$ in a grid of values, we run the Gibbs sampler and compute the posterior expectation of the number of clusters $\mathrm{E}[K \mid y]$. Finally, in the corresponding graph, we look for the value of $\lambda_{\max}$ located at the "elbow" of the curve, that is, the value of $\lambda_{\max}$ at which the values of $\mathrm{E}[K \mid y]$ level off.

6. Simulated Data

As an explicit example of the graphical criterion in action, we used the functional form of a first-order model with thinning parameter $\alpha = 0.15$ to simulate a time series of length $T = 1000$, for which the distribution of the innovations is a symmetric mixture of three Poisson distributions with parameters 1, 8, and 15. Figure 1 shows the formations of the elbows for two values of the discount parameter: $\sigma = 0.5$ and $\sigma = 0.75$.

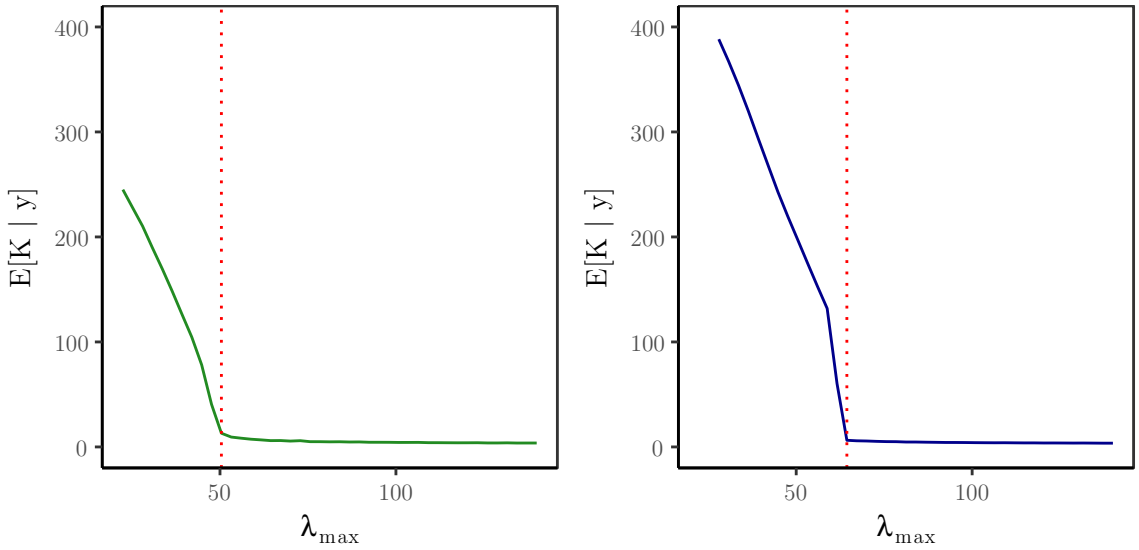

Figure 1. Formation of the elbows for $\sigma = 0.5$ (left) and $\sigma = 0.75$ (right). The red dotted lines indicate the chosen values of $\lambda_{\max}$.

For the simulated time series, Figures 2–5 display the behavior of the posterior distributions obtained using the elbow method for $(k_0, \sigma) \in \{4, 10, 16, 30\} \times \{0, 0.25, 0.5, 0.75\}$. These figures make the relation between the choice of the value of the discount parameter σ and the achieved robustness of the posterior distribution quite explicit: as we increase the value of the discount parameter σ, the posterior becomes insensitive to the choice of k_0. In particular, for $\sigma = 0.75$, the posterior mode is always near 3, which is the number of components used in the distribution of the innovations of the simulated time series.

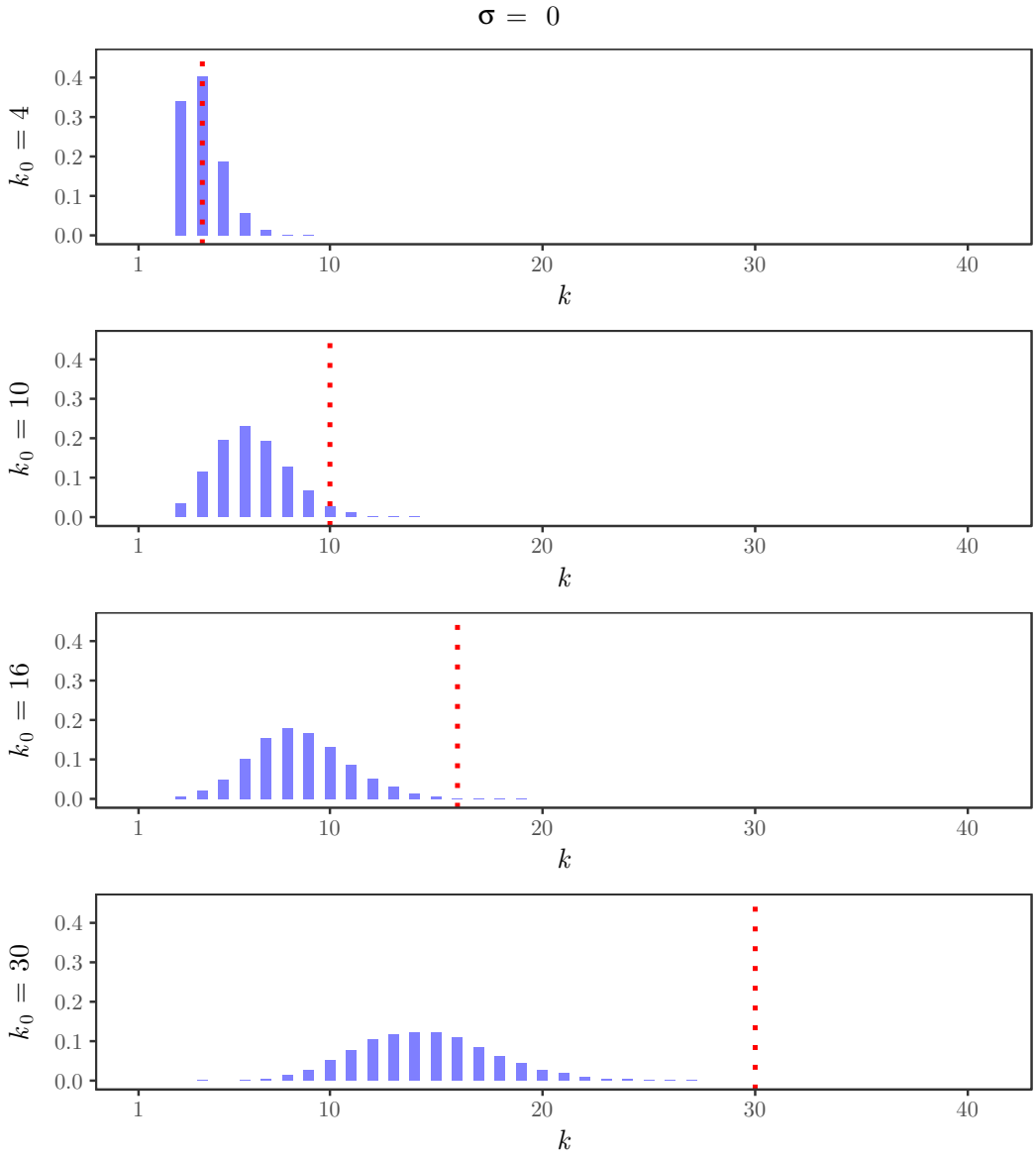

Figure 2. Posterior distributions of the number of clusters K for the simulated time series with $\sigma = 0$ and $k_0 = 4, 10, 16, 30$. The red dotted lines indicate the value of k_0.

Once we understand the influence of the prior parameters on the robustness of the posterior distribution, an interesting question is how to get a point estimate for the distribution of clusters, in the sense that each λ_t, for $t = p+1, \ldots, T$, would be assigned to one of the available clusters.

From the Gibbs sampler, we can easily get a Monte Carlo approximation for the probabilities $d_{rt} = \Pr\{\lambda_r \neq \lambda_t \mid y\}$, for $r, t = p+1, \ldots, T$. These probabilities define a dissimilarity matrix $D = (d_{rt})$ among the innovation rates. Although D is not a distance matrix, we can use it as a starting point to represent the innovation rates in a two-dimensional Euclidean space using the technique of metric multidimensional scaling (see [18] for a general discussion). From this two-dimensional representation, we use hierarchical clustering techniques to build a dendrogram, which is appropriately cut in order to define three clusters, allowing us to assign a single cluster label to each innovation rate.

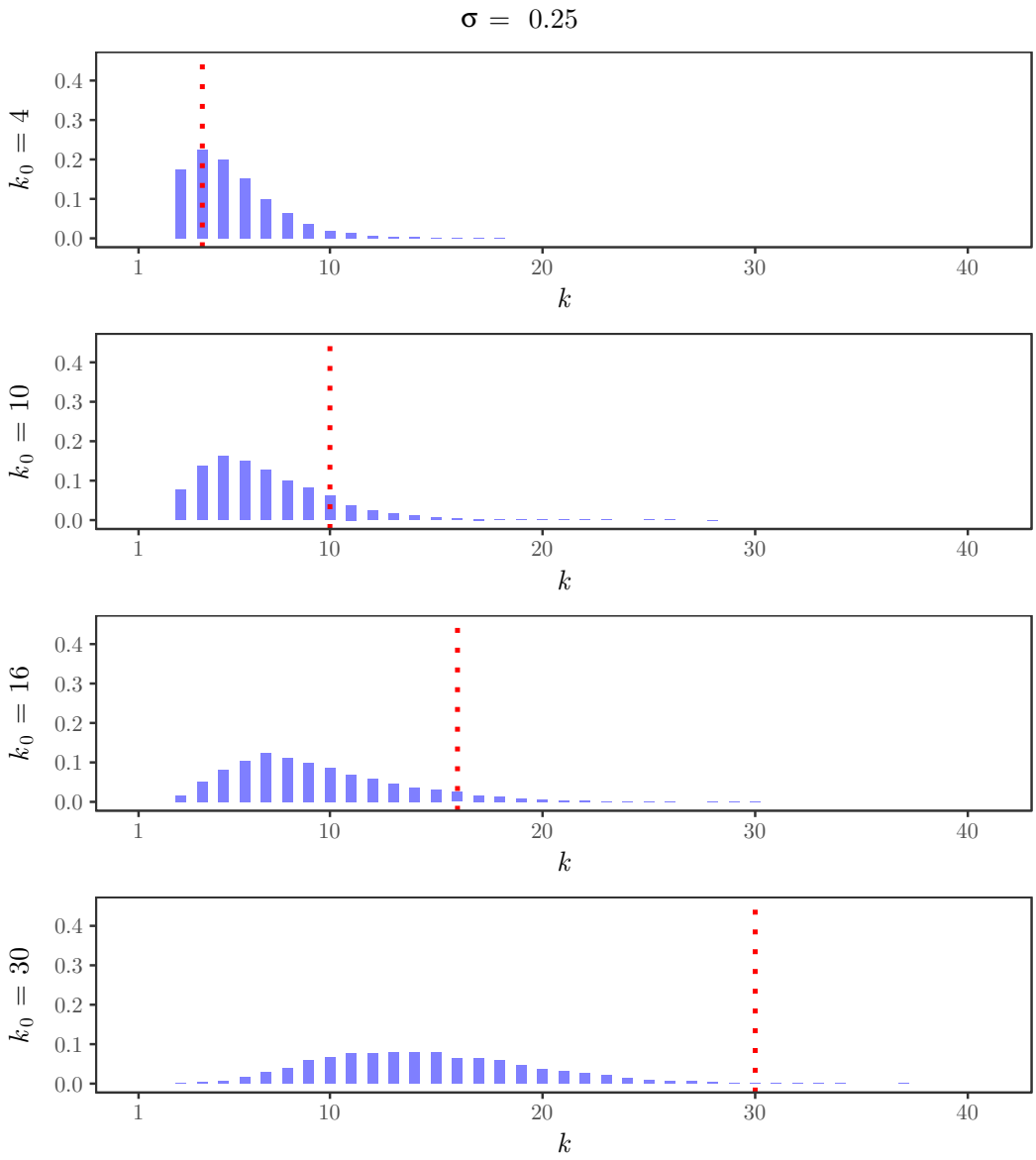

Figure 3. Posterior distributions of the number of clusters K for the simulated time series with $\sigma = 0.25$ and $k_0 = 4, 10, 16, 30$. The red dotted lines indicate the value of k_0.

Table 1 displays the confusion matrix of this assignment, showing that 83% of the innovations were grouped correctly in the clusters which correspond to the mixture components used to simulate the time series.

Table 1. Confusion matrix for the cluster assignments.

	True		
Predicted	1	2	3
1	297	32	0
2	11	217	42
3	0	84	316

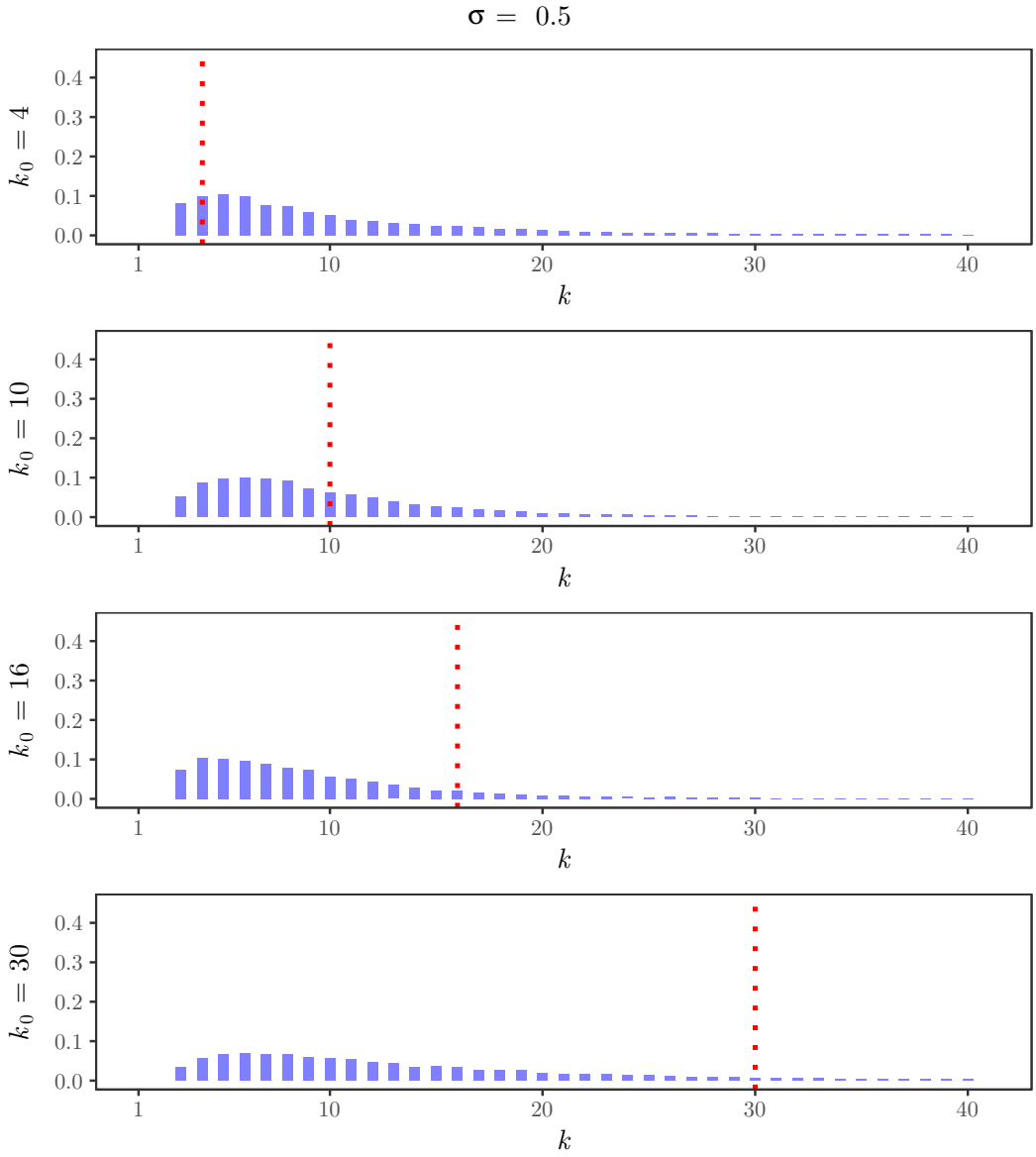

Figure 4. Posterior distributions of the number of clusters K for the simulated time series with $\sigma = 0.5$ and $k_0 = 4, 10, 16, 30$. The red dotted lines indicate the value of k_0.

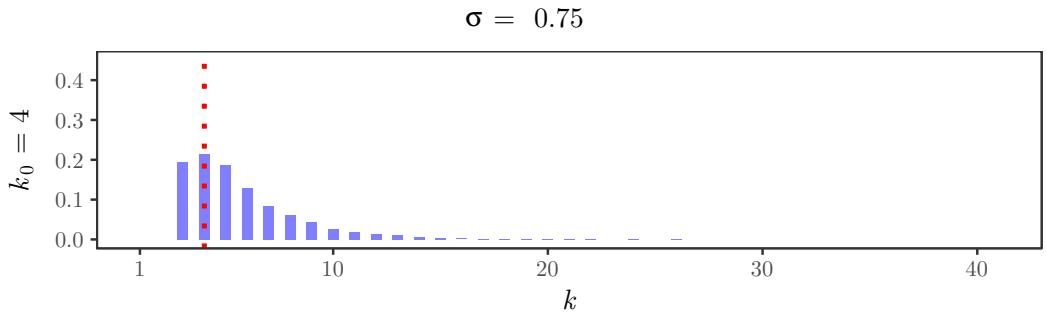

Figure 5. *Cont.*

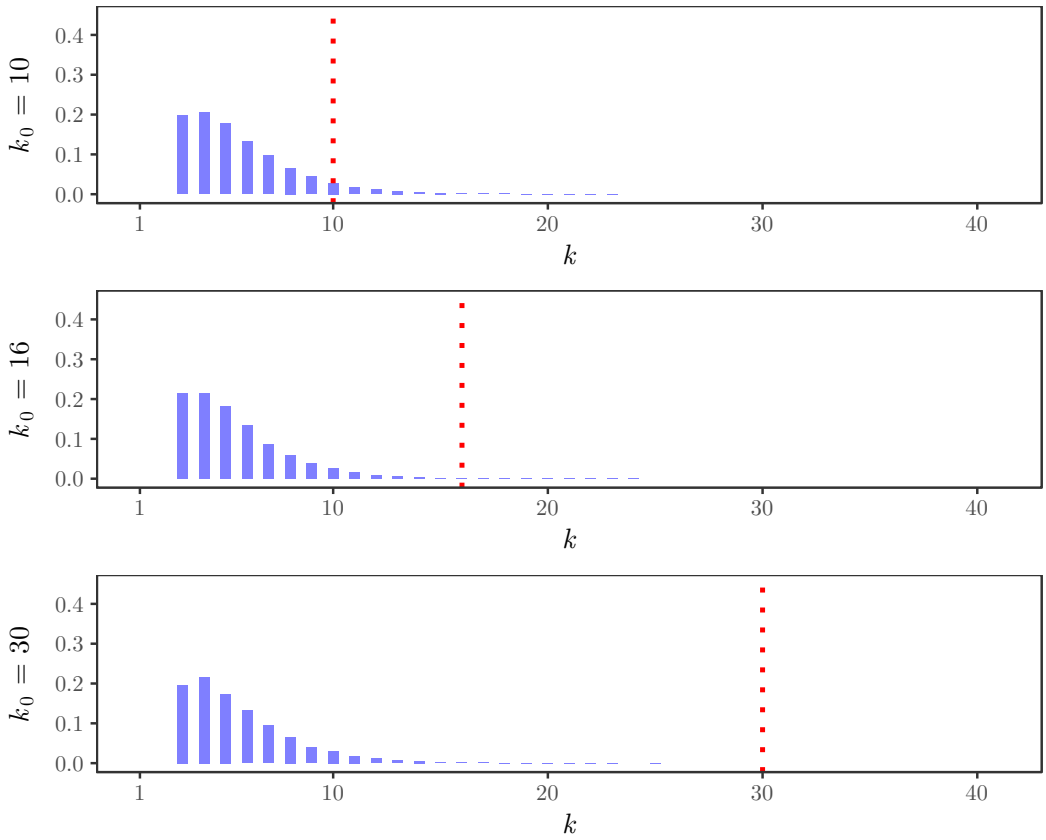

Figure 5. Posterior distributions of the number of clusters K for the simulated time series with $\sigma = 0.75$ and $k_0 = 4, 10, 16, 30$. The red dotted lines indicate the value of k_0.

7. Earthquake Data

In this section, we analyze a time series of yearly worldwide earthquakes events of substantial magnitude (equal or greater than 7 points on the Richter scale) from 1900 to 2018 (http://www.usgs.gov/natural-hazards/earthquake-hazards/earthquakes).

The forecasting performances of the INAR(p) and the PY-INAR(p) models are compared using a cross-validation procedure in which the models are trained with data ranging from the beginning of the time series up to a certain time, and predictions are made for epochs outside this training range.

Using this cross-validation procedure, we trained the INAR(p) and the PY-INAR(p) models with orders $p = 1, 2,$ and 3, and made one-step-ahead predictions. Table 2 shows the out-of-sample mean absolute errors (MAE) for the INAR(p) and the PY-INAR(p) models. In this table, the MAE's are computed predicting the counts for the last 36 months. For the three model orders, the PY-INAR(p) model yields a smaller MAE than the original INAR(p) model.

Table 2. Out-of-sample MAE's for the INAR(p) and the PY-INAR(p) models, with orders $p = 1, 2,$ and 3. The last column shows the relative variations of the MAE's for the PY-INAR(p) models with respect to the corresponding MAE's for the INAR(p) models.

	INAR	**PY-INAR**	$\Delta_{\text{PY-INAR}}$
$p = 1$	3.861	3.583	-0.072
$p = 2$	3.583	3.417	-0.046
$p = 3$	3.972	3.305	-0.202

Author Contributions: Theoretical development: H.G., A.L., H.F.L., P.C.M.F, I.P. Software development: H.G., P.C.M.F. All authors have read and agreed to the published version of the manuscript.

Funding: Helton Graziadei and Hedibert F. Lopes thank FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for financial support through grants numbers 2017/10096-6 and 2017/22914-5. Antonio Lijoi and Igor Prünster are partially supported by MIUR, PRIN Project 2015SNS29B.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McKenzie, E. Some simple models for discrete variate time series. *J. Am. Water Resour. Assoc.* **1985**, *21*, 645–650. [CrossRef]
2. Al-Osh, M.; Alzaid, A. First-order integer-valued autoregressive (INAR(1)) process: Distributional and regression properties. *Stat. Neerl.* **1988**, *42*, 53–61. [CrossRef]
3. Weiß, C. *An Introduction to Discrete-Valued Time Series*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
4. Du, J.G.; Li, Y. The integer-valued autoregressive (INAR(p)) model. *J. Time Ser. Anal.* **1991**, *12*, 129–142. [CrossRef]
5. Neal, P.; Kypraios, T. Exact Bayesian inference via data augmentation. *Stat. Comput.* **2015**, *25*, 333–347. [CrossRef]
6. Pitman, J. Combinatorial stochastic processes. Technical Report 621; Springer: Berlin/Heidelberg, Germany, 2002. [CrossRef]
7. Dion, J.; Gauthier, G.; Latour, A. Branching processes with immigration and integer-valued time series. *Serdica Math. J.* **1995**, *21*, 123–136.
8. Van Dyk, D.; Meng, X.L. The art of data augmentation. *J. Comput. Graph. Stat.* **2001**, *10*, 1–50. [CrossRef]
9. Tanner, M.; Wong, W. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–540. [CrossRef]
10. Ferguson, T. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1973**, *1*, 209–230. [CrossRef]
11. Schervish, M.J. *Theory of Statistics*; Springer: Berlin/Heidelberg, Germany, 1995; pp. 52–60.
12. Hjort, N.; Holmes, C.; Müller, P.; Walker, S. *Bayesian Nonparametrics*; Cambridge University Press: Cambridge, UK, 2010; Voume 28.
13. Lijoi, A.; Mena, R.H.; Prünster, I. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **2007**, *94*, 769–786. [CrossRef]
14. Hamilton, J. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 1994; Volume 2, pp. 43–71.
15. Gamerman, D.; Lopes, H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006.
16. Escobar, M.; West, M. Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*; Dey, D., Müller, P., Sinha, D., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; Chapter 1, pp. 1–22. [CrossRef]
17. Canale, A.; Prünster, I. Robustifying Bayesian nonparametric mixtures for count data. *Biometrics* **2017**, *73*, 174–184. [CrossRef] [PubMed]
18. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 570–572.