Statistics: A Journal of Theoretical and Applied Statistics
Vol. 00, No. 00, Month-Month 200x, 1–18

On the construction of stationary AR(1) models via random distributions.

Alberto Contreras-Cristán†, Ramsés H. Mena*†
and Stephen G. Walker‡
†*Universidad Nacional Autónoma de México, Ciudad de México, A.P. 20-726 México*
‡*University of Kent, Canterbury, CT2 7NZ, UK*
(January 2007)

We explore a method for constructing first order stationary autoregressive type models with given marginal distributions. We impose the underlying dependence structure in the model using Bayesian nonparametric predictive distributions. This approach allows for non-linear dependency and at the same time works for any choice of marginal distribution. In particular, we look at the case of discrete-valued models, that is the marginal distributions are supported on the non-negative integers.

Keywords: AR model, Beta Stacy process, Bayesian nonparametrics, discrete-valued time series, Pólya trees, stationary process.

1 Introduction

In the literature, many of the constructions of stationary time series models with fixed marginal distributions usually rely on a simple dependence structure, typically being dominated by certain linearity conditions, e.g. $\mathbb{E}[X_t \mid X_{t-1} = x] = a\,x + b$. In this paper we aim to further explore a novel approach, introduced in a parametric framework in [25] and generalised to the nonparametric case in [27]. In particular, we will focus on the discrete-valued case.

These authors suggest a method to construct stationary autoregressive-type models of first order (AR(1)-type) with arbitrary, but fixed, stationary distributions.

Given a specific stationary distribution with distribution function $Q(x)$, the proposed model has the one-step transition distribution driving the AR(1)-

*Correspondence author: R. H. Mena. Departamento de probabilidad y estadística, IIMAS-UNAM, A.P. 20-726, Mexico, D.F. 01000. Mexico. Email: ramses@sigma.iimas.unam.mx

type model $\{X_t\}$ as

$$\mathrm{P}_{x_{t-1}}(x_t) = \mathrm{E}\{G_y(x_t) \,|\, x_{t-1}\} = \int G_y(x_t)\, G^*_{x_{t-1}}(\mathrm{d}y), \tag{1}$$

where $\mathrm{P}_{x_{t-1}}(x_t) = \Pr[X_t \leq x_t \mid X_{t-1} = x_{t-1}]$, $G_y(x_t) = \Pr[X_t \leq x_t \mid Y = y]$ and $G^*_{x_{t-1}}(y) = \Pr[Y \leq y \mid X_{t-1} = x_{t-1}]$. Notice that, in a Bayesian context, $G_y(x_t)$ denotes the corresponding posterior of $G^*_{x_{t-1}}(y)$ under $Q(x)$.

Here the two conditional distributions, $G_y(x_t)$ and $G^*_{x_{t-1}}(y)$, arise from a single joint distribution, $G(x, y) := \Pr[X \leq x, Y \leq y]$, such that

$$\int G^*_x(y)\, Q(\mathrm{d}x) \, = Q^*(y).$$

Equation (1) can also be thought as the Bayesian predictive distribution with likelihood $G_y(x)$ with certain prior $Q^*(y)$ and based on the single observation x_{t-1}.

The main part in this construction, imposing the dependence structure between X_t and X_{t-1}, is the choice of the parametric family $G^*_x(y)$, which given the choice of a stationary distribution $Q(x)$ and using Bayes' theorem, leads to $G_y(x)$. Notice that, a transition distribution constructed as in (1) satisfies

$$\int \mathrm{P}_{x_{t-1}}(x_t)\, Q(\mathrm{d}x_{t-1}) = Q(x_t),$$

that is, the distribution Q remains invariant and therefore an AR(1)-type model driven by (1) is strictly stationary having Q as its stationary distribution.

Clearly the range of possible choices for $G^*_x(y)$ is too wide and setting it to a specific parametric form results in a different model, namely a different dependence structure, but with the same stationary distribution.

In order to circumvent the potential problem of a misspecified dependence, Mena and Walker [27] proposed a nonparametric choice for $G^*_x(y)$, which makes the dependence structure more flexible. Effectively, they replace the latent variable "y" with a random distribution $\mathcal{G}$, with the probability measure written as $\mathcal{P}(\mathrm{d}\mathcal{G})$. That is, $\mathcal{P}$ denotes a probability measure on the space of probability measures on $\mathbb{R}$, say $(\mathcal{F}, \mathcal{B}_{\mathcal{F}})$. Applying a similar construction as in (1), we shall consider the "posterior" distribution corresponding to the nonparametric "prior" $\mathcal{P}$, in order to construct $\mathrm{P}_{x_{t-1}}(x_t)$. Specifically, for a set $A \in \mathcal{B}_{\mathcal{F}}$, start with the joint distribution

$$\Pr(X \leq x; \mathcal{G} \in A) = \mathbb{E}_{\mathcal{P}}[\mathcal{G}(x)\mathbb{I}(\mathcal{G} \in A)] = \int_A \mathcal{G}(x)\, \mathcal{P}(\mathrm{d}\mathcal{G}), \tag{2}$$

where $\mathcal{G} \in \mathcal{F}$. Note that for $A = \mathcal{F}$ we obtain $\Pr(X \leq x) = \mathbb{E}_{\mathcal{P}}[\mathcal{G}(x)]$. Hence, this implies that the posterior probability is given by

$$\mathcal{P}(\mathcal{G} \in A \mid X \leq x) = \frac{\mathbb{E}_{\mathcal{P}}[\mathcal{G}(x)\mathbb{I}(\mathcal{G} \in A)]}{\mathbb{E}_{\mathcal{P}}[\mathcal{G}(x)]}. \tag{3}$$

In a similar way we can condition on a singleton and obtain $\mathcal{P}_x(A) = \mathcal{P}(\mathcal{G} \in A \mid X = x)$ as the nonparametric Bayesian posterior distribution for the random distribution $\mathcal{G}$ given the single observation x. With these elements we proceed to "sweep out" the randomness in the nonparametric component to obtain the Markovian transition distribution given by

$$\mathrm{P}_{x_{t-1}}(x_t) = \int \mathcal{G}(x_t)\, \mathcal{P}_{x_{t-1}}(\mathrm{d}\mathcal{G}). \tag{4}$$

Even though, we have used a random distribution in this construction, notice that the resulting transition distribution is no longer random. However, due to the grater dimensionality of the support of $\mathcal{G}$ over the support of y, the transition resulting from (4) can encompass a wider dependence structure than the one arising from (1). Analogously to the parametric setting, the transition distribution can be interpreted as the Bayesian predictive distribution based on the prior $\mathcal{P}(\mathrm{d}\mathcal{G})$ and a single observation. This generalisation connects us with the area of nonparametric Bayesian methods; see [24] for a recent review. Note that the similarity to the parametric approach is given by replacing $\mathrm{G}_y(x)$ by $\mathcal{G}(x)$ and $G^*_x(y)$ by $\mathcal{P}_x(\mathrm{d}\mathcal{G})$.

The advantage of this generalisation is that there is no need to specify $G^*_x(y)$ in a parametric way. The dependence in the model is then determined by the choice of measure $\mathcal{P}(\cdot)$ and the stationary distribution is defined as $Q(x) := \Pr(X \leq x) = \mathbb{E}_{\mathcal{P}}\{\mathcal{G}(x)\}$. Therefore, whereas in the Bayesian nonparametric literature, Q is known as the baseline, in our construction Q will act as the stationary distribution for the AR(1)-type model.

The issue here is which random measure should we use in order to produce certain types of dependence or to match certain features underlying to the data, for instance an AR(1)-type model being discrete-valued and non-linear.

A simple example arises when $\mathcal{P}(\cdot)$ is the Dirichlet process. Denote by $\mathscr{D}(c\,Q)$ a Dirichlet process driven by the measure $c\,Q(\cdot)$, where $c > 0$. A random distribution function chosen by $\mathcal{G} \sim \mathscr{D}(c\,Q)$ satisfies

$$\mathbb{E}_{\mathcal{P}}\{\mathcal{G}(x)\} = Q(x),$$

for any $x \in \mathbb{R}$. See Ferguson [1]. The parameter $c > 0$ is commonly associated with the variability of the random distributions $\mathcal{G}$ about Q. In this case, the

well-known conjugacy property of the Dirichlet process leads to

$$\mathcal{G} \mid [X = x] \sim \mathscr{D}(cQ + \delta_x),$$

where δ_x denotes the point mass at x; see [1]. Following the ideas described above, we can construct the following transition distribution driving the AR(1)-type model $\{X_t\}_{t=1}^{\infty}$

$$\begin{aligned} \mathrm{P}_{x_{t-1}}(x_t) &= \mathbb{E}_{\mathcal{P}} \left\{ \mathcal{G}(x_t) \mid X_{t-1} = x_{t-1} \right\} \\ &= \frac{c}{c+1} Q(x_t) + \frac{1}{c+1} \mathbb{1}(x_t = x_{t-1}), \end{aligned} \tag{5}$$

which remains invariant with respect to Q.

In [27] the case of continuous-valued models was undertaken by choosing $\mathcal{P}$ to be the generalised log-Gaussian process [10]. The main objective of this paper is to explore this construction in two cases not covered in [27], namely that devoted to model discrete valued data and that being able to capture negative correlations. This lead us to two different choices of probability measures $\mathcal{P}(\cdot)$. First, in order to construct a model for discrete-valued AR(1)-type data we should use a measure which puts positive probability to discrete distributions, for this purpose we use a discrete-version of the Beta-Stacy process [21]. Under a different choice of $\mathcal{P}$, we will explore the Pólya tree distribution which leads us to AR(1)-type models with an appealing dependence structure, this is easily extendible to model negative correlation among observations.

Describing the layout of the paper; Section 2 provides with a brief discussion on the problem of constructing discrete-valued models. In Section 3 the nonparametric approach to construct discrete-valued AR(1) models using the Beta-Stacy process is undertaken. We also illustrate the capabilities of the proposed model with a example based on simulated data. Section 4 discusses the construction of stationary AR(1) models via Pólya trees. In particular, we address the issue of obtaining models for negatively correlated observations.

2 Discrete AR(1)-type models

Discrete-valued time series are found in many applications in statistics. This has encouraged researchers to develop adequate models for such data. McDonald and Zucchini [19] and McKenzie [26] review many models available in the literature. Most of the constructions of such models available in the literature are devoted to a specific parametric form, e.g. Binomial distribution, Poisson distribution. Furthermore, they are devoted to linear dependence structures, fact which makes them unsuitable for some applications.

One of the first efforts to tackle the modelling of discrete-valued time series is found in Jacobs and Lewis [3–5]. The idea in these papers can be easily stated in the AR(1)-type case given by

$$X_t = V_t\, X_{t-1} + (1 - V_t)\, Z_t,$$

where the $\{V_t\}$ are i.i.d. binary variables with $P(V_t = 1) = \rho$, $0 < \rho < 1$, and the $\{Z_t\}$ are i.i.d. with distribution Q. This model, known as the discrete autoregressive (DAR(1)) model, leads to a stationary model with Q as the stationary distribution. After a suitable re-parametrisation, model (5), can be seen as the DAR(1) model. Take $c = \rho^{-1} - 1$ to obtain the parametrisation in [5]. Although this approach encompasses a wide choice of marginal distributions, the simple construction based on a linear combination of i.i.d. discrete random variables leads to a very simple dependence structure.

Another way to construct stationary discrete-valued AR models is based on thinning operators. The most common of these operators is the binomial thinning; if N is a non-negative integer and $\rho \in [0, 1]$ then $\rho * N = \sum_{i=1}^{N} B_i(\rho)$, where $\{B_i(\rho)\}$ is a sequence of i.i.d. Bernoulli random variables, independent of N, satisfying $\Pr(B_i(\rho) = 1) = \rho$. This operator was proposed by [6] to generalise self-decomposable random variables to the discrete case. Many authors used this idea to construct models of the type

$$X_t = \rho * X_{t-1} + Z_t,$$

with specific marginal distributions. For example, McKenzie [7–9,11] proposed models for binomial, negative binomial and poisson marginals. See also [12, 14, 15].

An extension of the concept of thinning was also utilised to construct models of the type

$$X_t = A_t(X_{t-1}) + Z_t, \tag{6}$$

where $A_t(x)$ denotes a random operator defined by the law of the conditional distribution of $X_1 \mid [X_1 + X_2 = x]$, where $X_1 + X_2 \stackrel{\mathrm{d}}{=} Q$, the required marginal for the constructed AR(1) model.

An example of this extension can be found in [13], where AR(1) models with Binomial(N, p) marginals were introduced. In this case $Z_t \sim$Binomial$(N - M, \rho)$ and

$$A_t(X) \mid [X = x] \sim \text{Hypergeometric}(N, x, M),$$

which defines a hypergeometric thinning. For this model the resulting auto-

correlation function (ACF) is given by ρ^h with $\rho = M/N$, which depends on a parameter from the marginal distribution. This approach was also used in a more general setting [18] to construct AR models with convolution-closed infinitely divisible marginal distributions.

A common factor among all the models mentioned in this section is the linear dependence in the conditional mean, given by

$$\mathrm{E}(X_t \mid X_{t-1}) = \rho\, X_{t-1} + (1-\rho)\,\mu, \tag{7}$$

where $\mu = \mathrm{E}_Q(X)$ denotes the mean of the stationary density. This property implies an exponentially decaying ACF.

3 A discrete time version of the Beta-Stacy process

In this section we will apply the construction described in the introduction by choosing $\mathcal{P}(\cdot)$ to be the Beta-Stacy process [21]. We will briefly review this random distribution in the case where its support coincides with the space of distribution functions with support on $\{0, 1, 2, \ldots\}$, and denote this space by $\mathcal{F}$.

Denote by $B(\alpha, \beta)$ the beta function and by $\mathcal{C}(\alpha, \beta, \xi)$ the Beta-Stacy distribution with density function given by

$$\frac{1}{B(\alpha,\beta)} y^{\alpha-1} \frac{(\xi - y)^{\beta-1}}{\xi^{\alpha+\beta-1}} \mathbb{1}(y \in (0, \xi)).$$

Let $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ be a set of non-negative integers. Consider the sequence of positive random variables $\{Y_k; k \in \mathbb{N}_0\}$ given through

$$\begin{aligned} Y_1 &\sim \mathcal{C}(\alpha_1, \beta_1, 1), \\ Y_2 | Y_1 &\sim \mathcal{C}(\alpha_2, \beta_2, 1 - Y_1), \\ &\vdots \\ Y_k | Y_{k-1}, \ldots, Y_1 &\sim \mathcal{C}(\alpha_k, \beta_k, 1 - \mathcal{G}(k-1)), \end{aligned} \tag{8}$$

where $\{\alpha_k\}$ and $\{\beta_k\}$ are sequences of of positive real numbers and

$$\mathcal{G}(k) = \sum_{j=1}^{k} Y_j.$$

Proposition 1 in [21] states that almost surely the discrete time Beta-Stacy process

$$\mathcal{G}(k) = \begin{cases} 0 & \text{if } k = 0, \\ \sum_{j \leq k} Y_j & \text{if } k > 0 \end{cases}$$

is an element of $\mathcal{F}$. Thus, the random size of the jump of $\mathcal{G}$ at k is given by Y_k, and for each $m = 1, 2, \ldots$ the joint probability distribution of the vector $(Y_1, \ldots, Y_m)$ is the generalised Dirichlet distribution, presented in [21].

By considering the sets and $A_k = \{0, \ldots, k\}$, it is possible to center the process on a particular distribution $Q \in \mathcal{F}$ with mass function $q(k) > 0$, for all k. To this end, choose

$$\alpha_k = c_k q(k) \text{ and } \beta_k = c_k\{1 - Q(k)\} = c_k \left\{1 - \sum_{l=0}^{k} q(l)\right\}, \tag{9}$$

where $\{c_k\}$ is a sequence of positive real numbers. See [21] for details.

Given a random sample $X_1, \ldots, X_n$ from an unknown distribution $\mathcal{G}$ with discrete support, if $\mathcal{G}$ comes from a discrete Beta-Stacy process then the predictive mass function based on one observation is given by

$$\Pr(X_t = x_t \mid X_{t-1} = x_{t-1}) = h(x_t \mid x_{t-1}) \times \prod_{\xi < x_t} \{1 - h(\xi \mid x_{t-1})\}, \tag{10}$$

where

$$\begin{aligned} h(\xi \mid x_{t-1}) = {} & \frac{\alpha_\xi}{\alpha_\xi + \beta_\xi} \mathbb{1}(\xi > x_{t-1}) + \frac{\alpha_\xi + 1}{\alpha_\xi + \beta_\xi + 1} \mathbb{1}(\xi = x_{t-1}) \\ & + \frac{\alpha_\xi}{\alpha_\xi + \beta_\xi + 1} \mathbb{1}(\xi < x_{t-1}). \end{aligned}$$

It is worth emphasizing the parametric nature of the transition mechanism (10), which is due to the fact that the randomness in the nonparametric component has been integrated. In particular, if we use the choice of α's and β's given by (9) the transition mass function (10) has Q as the invariant distribution. Therefore, by imposing such a dependency, we have constructed a discrete-valued stationary AR(1) with transition function given by (10) and having Q as the stationary distribution. We will name this model the *Beta-Stacy AR*(1) *model*. Given a particular set of observations modelled through the Beta-Stacy AR(1) model, the required fitting translates to estimate the sequence $\{c_k\}$, and possible unknown parameters contained in the stationary

distribution Q. Notice that, from (10), the dependence in the regressor x_{t-1} is only affected by its relative level with respect to values less than x_t.

For the moment, let us assume that Q has a finite support on $\{x_0, x_2, \ldots, x_l\}$, so there is a finite number of parameters $\{c_0, c_1, \ldots c_l\}$. If $c_k = c > 0$ for every k, then the corresponding Beta-Stacy process turns out to be the Dirichlet process, see [20]. Furthermore, in this case, from equation (5) and the stationarity of $\{X_t\}$ the corresponding auto-correlation sequence is given by $\rho_k = (1+c)^{-k}$. Thus, for the Beta-Stacy process, different patterns of auto-correlation functions can arise depending on the values of $\{c_0, c_1, \ldots c_l\}$. As a way to illustrate that different patterns for the auto-correlation sequence of the process can be obtained, Figure 1 shows different auto-correlation sequences. The autocorrelations of the underlying AR(1) model, with transition probability given by (10), were computed by setting different values of the sequence $\{c_k; k = 0, 1, \ldots, 9\}$ and having a Binomial$(9, 0.3)$ distribution as the stationary distribution, Q. The upper panel was obtained using $c_k = \beta \exp\{\alpha k\}, k = 0, 1, \ldots, 9$, where $\beta = 0.001$ and $\alpha = 0.1$. The middle panel was obtained by changing these parameters to $\beta = 2.4$ and $\alpha = -0.1$. The lower panel was obtained by using $c_{k-1} = \gamma\,(1 - k/11)^{\alpha}\,(k/11)^{\beta} + c\,(1 - k/11)^{a}\,(k/11)^{b}$, $k = 1, 2, \ldots, 10$, where $\alpha = 0.9$, $\beta = 4.0$, $\gamma = 3$, $a = 4.0$, $b = 0.9$ and $c = 3$.

3.1 *Simulated Data*

In this section we aim to show how the Beta-Stacy AR(1) model performs to capture a known dependence contained in a simulated data set. For this illustration we have considered 200 simulated data points from the AR(1) model with binomial marginal distribution, introduced in [13], equation (6). For the choice of the marginal distribution we set $N = 5$, assumed to be known, and $p = 0.5$. The one-lag autocorrelation in this model is given by $\rho = M/N$, hence we set $\rho = 0.2$ by fixing $M = 1$.

In order to estimate the parameters in the Beta-Stacy AR(1) model, that is $\{c_0, \ldots, c_5\}$ and p, we used numerical maximum likelihood estimation via the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimisation algorithm. For the complete specification of the Beta-Stacy AR(1) we have chosen Q is binomial, so its support is finite. Denote by $i^\star$ the value of the largest state in the support of the stationary distribution Q, in this example $i^\star = 5$. It is not difficult to show that the transition probabilities (10) do not depend on $c_{i^\star}$ and consequently the likelihood function will not depend on $c_{i^\star}$. Thus we are actually estimating $\{c_0, \ldots, c_4\}$ and p.

In order to present the estimations result based on a fair simulation procedure we use the following criterion proposed by Walden *et al.* [23]. For a given sample $\{X_1^{(i)}, \ldots, X_{200}^{(i)}\}$ of the process, simulated from model (6), define the

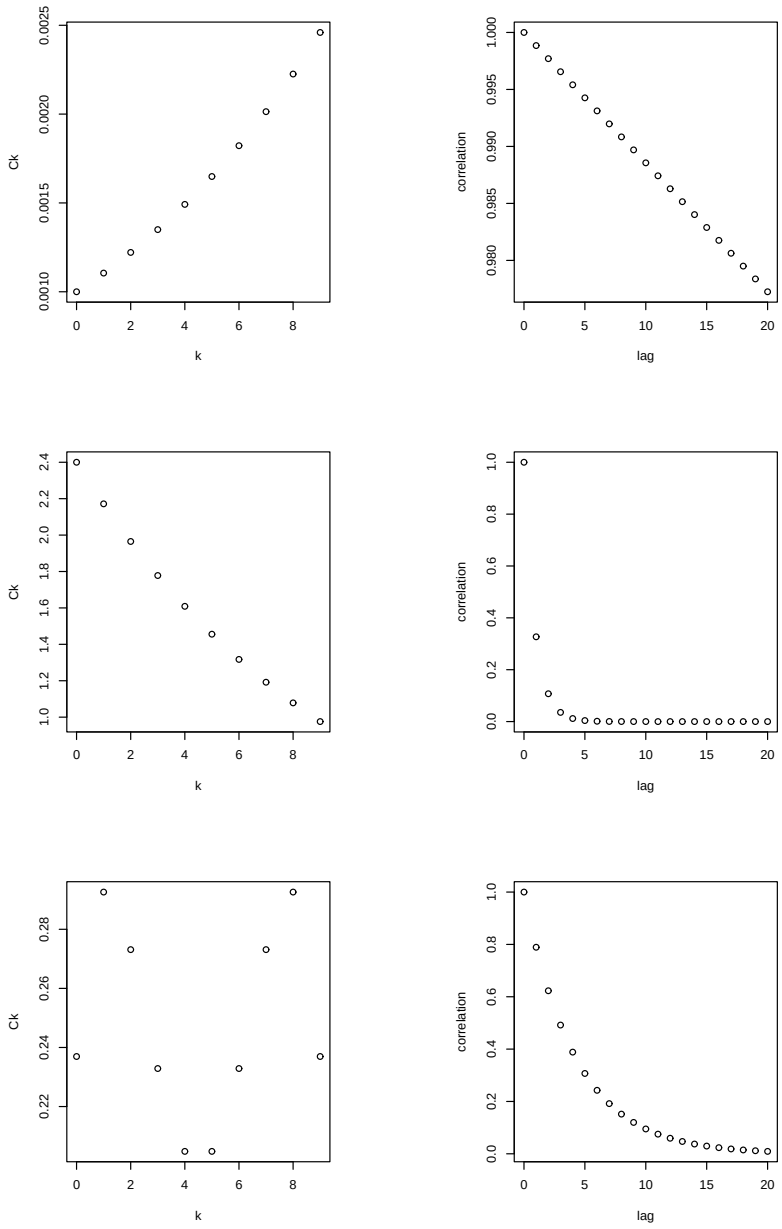

Figure 1. On the Left: The parameter sequence $\{c_k\}$, On the Right: Associated auto-correlation sequence $\{\rho_k\}$

root mean square error (RMSE) of the spectral density as

$$\left\{ \frac{1}{W/2+1} \sum_{k=0}^{W/2} \left[\widehat{S}^i(k/W) - S(k/W) \right]^2 \right\}^{1/2}, \tag{11}$$

where $W = 200$, S denotes the spectral density and $\widehat{S}^i$ is the estimated spectral density corresponding to the i-th sample. We computed this RMSE for $i = 1, 2, \ldots, 500$ samples.

We report the results obtained from the sample whose RMSE is located at the 50 % quantile. The maximum likelihood estimators are $\hat{c} =$ (3.73, 3.02, 8.81, 13.79, 13.67) and $\hat{p} = 0.488$. For the BFGS optimisation algorithm we used $p_0 = 0.1$ and $c_0 = (1, 1, 1, 1, 1)$ as initial values. Figure 2 shows the simulated data together with the estimated spectral density. From the spectral densities in Figure 2 is clear that the Beta-Stacy AR(1) model is able to capture the dependence up to a second-moment degree.

In order to see how our model is able to capture other dependencies on higher moments we have plotted, in Figure 3, the bivariate cumulative distribution functions $G_{X_t, X_{t-1}}$ corresponding to both the model and our estimate using the Beta-Stacy AR(1) model.

3.2 *Infinite support of Q*

For cases in which the stationary distribution has infinite support the BFGS optimisation algorithm is not feasible since the model becomes overparameterised. In order to overcome this issue we could reparameterise the c_k's to lower dimensions as we did for the illustrations in Figure 1. However, such an approach would limit the underlying dependence in the model.

Assuming that the parameters underlying to the chosen stationary distribution are known we are able to compute the likelihood function explicitly for each c_k and maximise the resulting expression to obtain an estimator. In a similar fashion of that followed to find MLE of Markov chains, we count the number of relevant transitions within the data. Let us first define n_{ib} to be the number of transitions which move from the state i to a state bigger than i (hence ib). Define also, in an obvious way, n_{bi}, n_{bb} and n_{ii}.

The likelihood for c_i is then given by

$$l_i \propto \left[1 - \frac{c_i q(i)}{c_i \tilde{Q}(i) + 1} \right]^{n_{bb}} \left[\frac{c_i q(i)}{c_i \tilde{Q}(i) + 1} \right]^{n_{bi}} \left[1 - \frac{c_i q(i) + 1}{c_i \tilde{Q}(i) + 1} \right]^{n_{ib}} \left[\frac{c_i q(i) + 1}{c_i \tilde{Q}(i) + 1} \right]^{n_{ii}},$$

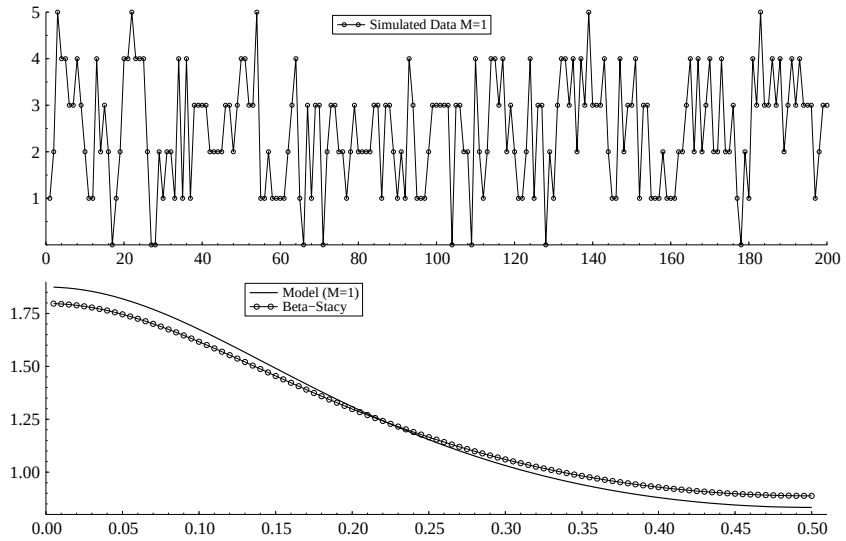

Figure 2. Top: 200 simulated data from the stationary AR(1) model of [13] with stationary distribution Binomial$(5, 0.5)$ and $M = 1$. Below: Spectral density corresponding to the model and its estimation using the Beta-Stacy AR(1) model.

where $q(i) = Q(i) - Q(i-1)$ and $\tilde{Q}(i) = (1 - q(0) - \cdots - q(i-1)) = \sum_{l=i}^{\infty} q(l)$.

Hence estimation via maximum likelihood is straightforward. We can go from $i = 0, 1, 2, \ldots$ maximising l_i to obtain $\widehat{c}_i$.

4 A stationary AR(1) model defined via Pólya trees

In this section we explore the use of the Pólya tree distribution as the choice for $\mathcal{P}(\cdot)$. Pólya tree distributions are an important ingredient in the development of Bayesian nonparametric techniques. Accounts regarding their construction and properties can be found in [16,20]. In what follows, we shortly review the features relevant for our approach.

For each $m = 0, 1, \ldots$, let $\{0,1\}^m \equiv \prod_{j=1}^{m} \{0,1\}$ and $\mathscr{B} = \{B_\epsilon\}$ be a binary partition of the state-space $(\mathrm{E}, \mathcal{E})$ where $\epsilon \in \{0,1\}^m$ this is $\epsilon = \epsilon_1 \cdots \epsilon_m$, $\epsilon_j \in \{0,1\}$. The subindex ϵ allocates the set B_ϵ in the tree while keeping the branch information. The partition mechanism in a Pólya tree is given as follows: in the mth level, partition B_ϵ splits into $(B_{\epsilon 0}, B_{\epsilon 1})$, then $B_{\epsilon 0}$ into $(B_{\epsilon 00}, B_{\epsilon 01})$ and so forth until infinity. Random mass is allocated to the sets via independent beta random variables $Y_{\epsilon 0} \sim \mathrm{Be}(\alpha_{\epsilon 0}, \alpha_{\epsilon 0})$, $Y_{\epsilon 1} = 1 - Y_{\epsilon 0}$ for non-negative numbers $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 0}$. Then, at a given level m the random mass

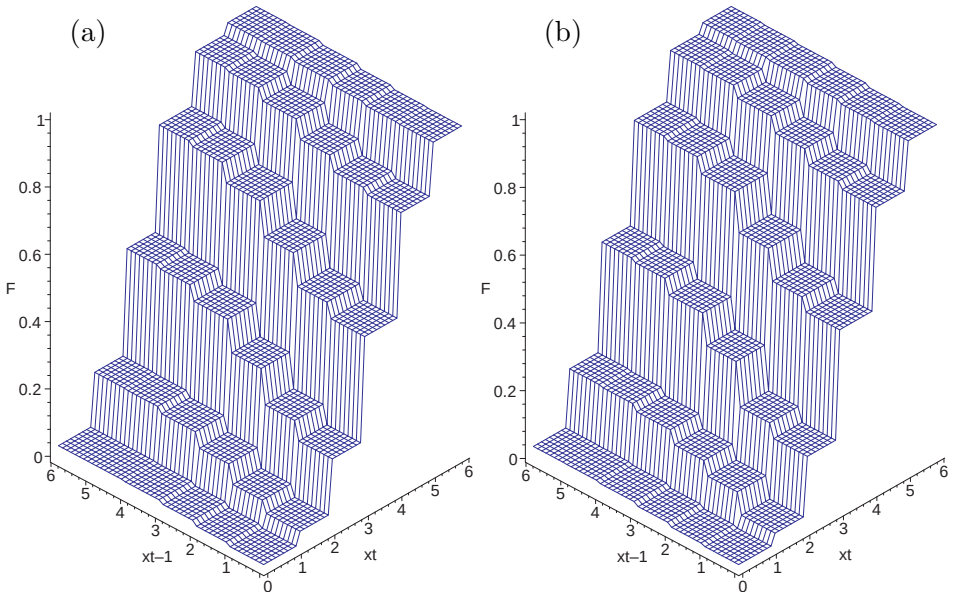

Figure 3. a) Bivariate distribution corresponding to the Binomial model by [13] with parameters $M = 1$, $p = 0.5$ and $N = 5$. b) Estimated bivariate distribution using the Beta-Stacy AR(1) model.

allocated to a particular set is given by

$$\mathcal{G}(B_\epsilon) = \left(\prod_{j=1;\, \epsilon_j=0}^{m} Y_{\epsilon_1 \cdots \epsilon_{j-1}0} \right) \left(\prod_{j=1;\, \epsilon_j=1}^{m} 1 - Y_{\epsilon_1 \cdots \epsilon_{j-1}0} \right),$$

where $\epsilon = \epsilon_1 \cdots \epsilon_m$. In theory the number of levels required is infinity, however an approximation is commonly used by terminating the process at a finite level m. For $\mathscr{A} \equiv \{\alpha_\epsilon, \epsilon \in \{0,1\}^\infty\}$, we use the notation $\mathcal{G} \sim \mathrm{PT}(\mathscr{B}, \mathscr{A})$ to denote a Pólya tree distribution. The Dirichlet process arises when $\alpha_{\epsilon 0} + \alpha_{\epsilon 1} = \alpha_\epsilon$ for all ϵ [2].

It is possible to center the process in a specific distribution Q by choosing the sets B_ϵ in the partition at level m as

$$\left[Q^{-1}\left(\frac{j-1}{2^m}\right), Q^{-1}\left(\frac{j}{2^m}\right) \right) \tag{12}$$

for $j = 1, \ldots, 2^m$, see [2, 16, 17, 24].

Under an exchangeable sampling scheme, the one-data based posterior prob-

ability of $\mathcal{G}$ given X_t is also a Pólya tree distribution $\mathrm{PT}(\mathscr{B}, \mathscr{A}|X_t)$, where

$$\mathscr{A} \mid X_t = \begin{cases} \alpha_\epsilon + 1 & \text{if } X_t \in B_\epsilon, \\ \alpha_\epsilon & \text{otherwise.} \end{cases} \tag{13}$$

Given the stationary distribution Q and a level m, if the random measure in (4) is modelled by $\mathcal{G} \sim \mathrm{PT}(\mathscr{B}, \mathscr{A})$ and all α_ϵ are fixed to be a constant $c > 0$, then the predictive distribution based on one observation is given by

$$\Pr(X_t \in \mathrm{B}_\epsilon \mid X_{t-1} = x) = \begin{cases} \left(\frac{c+1}{2c+1}\right)^{k_t} \frac{c}{2c+1} \left(\frac{1}{2}\right)^{m-k_t-1} & 0 \leq k_t < m, \\ \left(\frac{c+1}{2c+1}\right)^{k_t} & k_t = m, \end{cases} \tag{14}$$

where $\epsilon = \epsilon_1 \cdots \epsilon_m$ and k_t denotes the number of levels in which both X_{t-1} and X_t share the same partition set. We will use (14) as the transition probability leading to our stationary AR(1)-type model.

It is worth mentioning that in a similar, but different, approach Sarno [22] used Pólya tree distributions to model the dependence in autoregressive models of first order. The difference in our approach lies in that we use predictive distributions which are always invariant when used as transition probabilities. Sarno's model is not always strictly stationary. She also raised, but did not study, the question of how to include negative dependence between observations. We shall address this issue in the rest of this section.

In order to construct a stationary AR(1) model with Q invariant distribution via Pólya trees, we fix the partitions to match the percentiles of Q as described in (12). Therefore the transition mechanism driving the underlying stationary model is approximated by (14).

Regarding the estimation of the parameter c, let the number of levels m approach to infinity in the transition (14). The score for c corresponding to a sample $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is given by

$$\frac{\partial \log L_{\mathbf{x}}(c)}{\partial c} = \frac{N-1}{c(2c+1)} - \frac{1}{(c+1)(2c+1)} \sum_t k_t.$$

By equating the above quantity to zero and solving for c we get the MLE for c, given by

$$\hat{c} = \frac{1}{\overline{k} - 1}, \tag{15}$$

where $\overline{k}$ denotes the mean of the number of levels shared for the consecutive

observations in the sample.

4.1 *Correlation structure*

At this point we need to study admissible values for $\hat{c}$. Despite the condition $c > 0$ allows us to contruct a Pólya tree distribution from which (14) arises by using equation (4), note that (14) is also well defined for negative values of c (e.g. $c < -1$). Thus, for such negative values of c, we would not have an associated random probability measure, however, still we can define an autoregressive process with fixed marginal Q.

Now, the condition $-1 < \hat{c} < 0$ leads to $\overline{k} < 0$, which is contradictory since $\overline{k}$ is an average of non-negative numbers. Therefore, we shall consider values of $\hat{c}$ in $(-\infty, -1) \cup (0, \infty)$, this matches the domain for c such that (14) defines a transition probability.

Notice that the value of c affects the dependence between observations in the sample. A natural question to ask is how the estimator $\hat{c}$ changes as the correlation in the sample varies. In the following example we use simulations to depict the latter relation.

Let us consider the Gaussian AR(1) model given by

$$Y_t = \rho Y_{t-1} + \sqrt{1-\rho^2}\,\epsilon_t, \tag{16}$$

where $0 < |\rho| < 1$ and $\epsilon_1, \epsilon_2, \ldots$ are independent and $\mathrm{N}(0,1)$ distributed. It is easy to verify that the above model is stationary with $\mathrm{Corr}(Y_t, Y_{t-1}) = \rho$.

In order to illustrate the dependence on the correlation ρ, of the parameter c, we have simulated series, with 10000 observations each, from the autoregressive model (16) ranging in a grid of values of ρ. For the resulting simulations we fitted a stationary Pólya tree AR(1) model with invariant distribution $Q = \mathrm{N}(0,1)$ and transition probability (14). In other words, given a ρ we simulate from model (16) and compute $\hat{c} = 1/(\overline{k} - 1)$. Figure 4 shows the results.

In general, negative correlation at lag 1 of the samples corresponds to negative values of $\hat{c}$ and positive correlation at lag 1 of the samples corresponds to positive values of $\hat{c}$. Let us consider the partition in two sets of the support of Q introduced by the mean of the distribution. If correlation at lag 1 is positive then consecutive observations tend to stay on the same side of the real line with respect to the mean. Thus, we expect $\overline{k} > 1$ and therefore $\hat{c} > 0$. On the other hand if correlation at lag 1 is negative then consecutive observations tend stay on opposite sides with respect to the mean. Thus, we expect $\overline{k} < 1$ so that $\hat{c} < 0$.

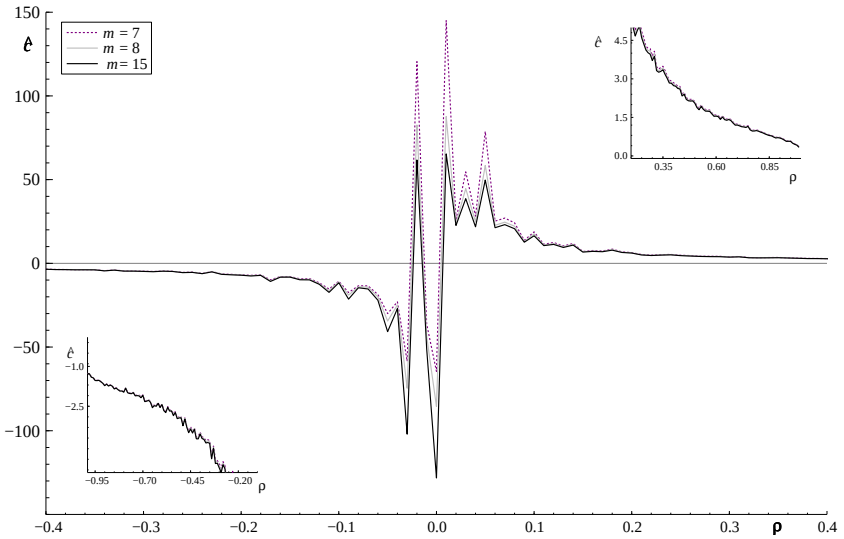

Figure 4. Estimator $\hat{c}$ as the correlation ρ varies. The estimator was computed for simulated data over for $\rho = -0.99, -0.98, \ldots 0.98, 0.99$. The central plot only shows results for $\rho = -0.4, -0.39, \ldots 0.39, 0.4$. In the upper-right corner, the behaviour of $\hat{c}$ for values of ρ close to 0.99 is shown. In the lower-left corner, a magnified plot presents the behaviour of $\hat{c}$ for values of ρ close to -0.99.

For positive values of ρ we note that the estimator $\hat{c}$ decreases as ρ increases. For large and positive correlation more levels are shared between consecutive observations (in mean), then from (15) $\hat{c}$ turns to be small.

When $|\rho|$ approaches to zero $\hat{c}$ is very unstable. For small correlation values, consecutive observations only share one level (in average), thus we expect $\overline{k} \approx 1$ and $|\hat{c}|$ can be very large.

As an instance, consider Q to be the uniform distribution over the set $[0, 1]$, $m = 10$ and $c = -2$. We used (14) to simulate a realisation of an autoregressive process with uniform marginal density and negative correlation. Figure 5, in the upper-left panel, shows the first 1000 samples of the realisation. In upper-right panel we show a histogram based on 10000 samples. Finally, in the lower panels we can appraise that the sample autocorrelation of order 1 is negative and significant.

We can apply the same principle to obtain a negative correlated Beta-Stacy AR(1) process. To this end define $u_k = \max\{1/q(k), 1/\tilde{Q}(k+1)\}$, where $\tilde{Q}(k) = \sum_{l=k}^{\infty} q(l)$. Then, for a positive number ϵ let us consider $c_k = -(1+\epsilon)u_k$. Again, because originaly c_k should be positive, we do not have a random probability measure from which (10) arises. However, this choice of $\{c_k\}$ enables us to use (10) as a transition probability and then to define an autoregressive process

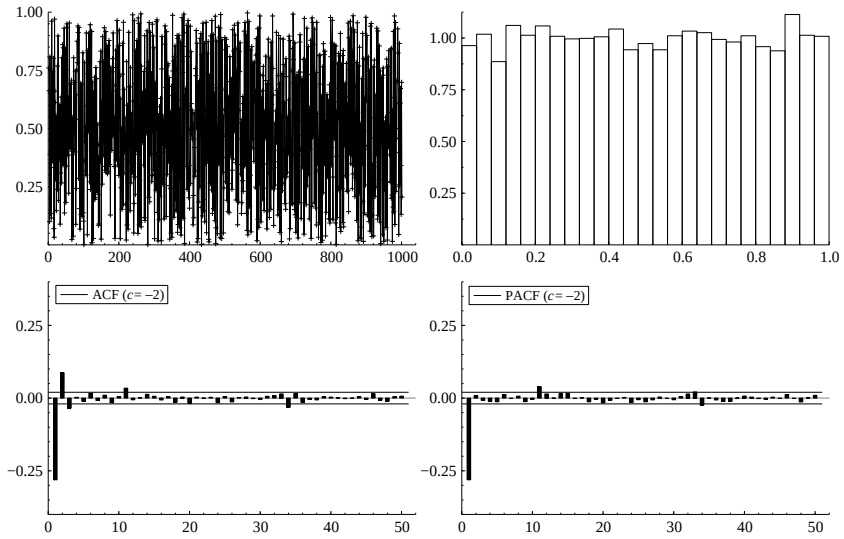

Figure 5. Simulation of the AR(1) Polya tree process. On the upper-left panel: first 1000 Simulated samples. On the upper-right panel: histogram of 10000 observations. On the lower-left panel: ACF of the sample. On the lower-right panel: PACF of the sample.

with negative correlation.

We have implemented this idea for $\epsilon = 0.2$, Q the Binomial probability distribution with parameters $N = 5$ and $p = 0.5$. Figure 6 shows the first 1000 samples of the realisation in the upper-left panel. In upper-right panel we show a histogram based on 10000 samples. Finally, in the lower panels the sample autocorrelation of order 1 is negative and significant.

Acknowledgments

Alberto Contreras-Cristán and Ramsés H. Mena are grateful for the support PAPIIT grant IN109906 and CONACyT grant J48538, UNAM, México. The research of Stephen G. Walker was partially supported by an EPSRC Advanced Research Fellowship.

References

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

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

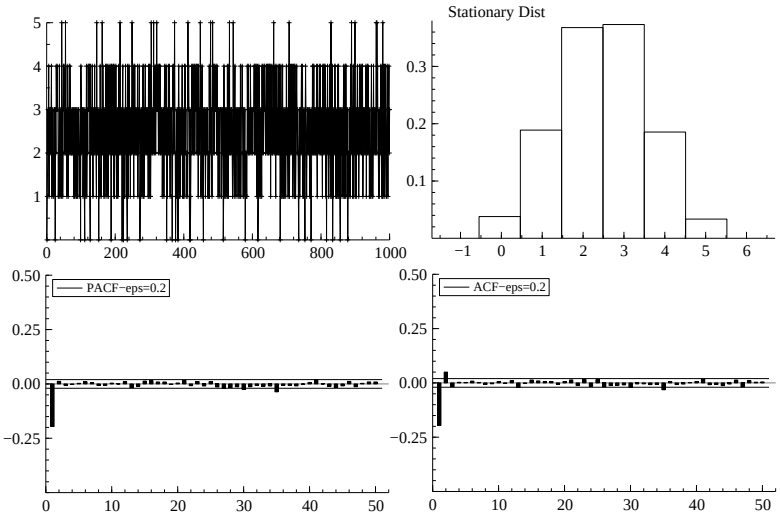

Figure 6. Simulation of the negative correlated Beta Stacy-AR(1) process. On the upper-left panel: first 1000 Simulated samples. On the upper-right panel: histogram of 10000 observations. On the lower-left panel: PACF of the sample. On the lower-right panel: ACF of the sample.

[3] Jacobs, P. A. and Lewis, P. A. W., 1978, Discrete time series generated by mixtures. I: Correlational and Runs properties. *Journal of the Royal Statistical Society. Series B,* **40**, 94–105.
[4] Jacobs, P. A. and Lewis, P. A. W., 1978, Discrete time series generated by mixtures. II: Asymptotic properties. *Journal of the Royal Statistical Society. Series B,* **40**, 222–228.
[5] Jacobs, P. A. and Lewis, P. A. W., 1978, Discrete time series generated by mixtures. III: Autoregressive processes (DAR(p)). Technical report NPS55-78-022, Naval Posgraduate School, Monterey, California.
[6] Steutel, F. W. and van Harn, K., 1979, Discrete analogues of self-decomposability and stability. *Annals of Probability,* **7,** 893–99.
[7] McKenzie, Ed., 1985, Some simple models for discrete variate time series. *Water Resources Bulletin,* **21,** 645–650.
[8] McKenzie, Ed., 1986, Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Advances in Applied Probability,* **18,** 679–705.
[9] McKenzie, Ed., 1987, First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analisys,* **8,** 261–275.
[10] Lenk, P. J., 1988, The logistic normal distribution for Bayesian nonparametric, predictive densities. *Journal of the American Statistical Association,* **83,** 509–516.
[11] McKenzie, Ed., 1988, Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability,* **20,** 822–835.
[12] Alzaid, A. A. and Al-Osh, M. A., 1990, An integer-valued pth order autoregressive structure (INAR(p)) process. *Journal of Applied Probability,* **27,** 314–324.
[13] Al-Osh, M. A. and Alzaid, A. A., 1991, Binomial autoregressive moving average models. *Communications in Statistics Stochastic Models,* **7,** 261–282.
[14] Du, J. G. and Li, Y., 1991, The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analisys,* **12,** 129–142.
[15] Al-Osh, M. A. and Aly, E. A. A., 1992, First order autoregressive time series with negative binomial and geometric marginals. *Communications in Statistics. Theory and Methods,* **21,** 2483–2492.
[16] Lavine, M., 1992, Some aspects of Pólya tree distributions for statistical modelling, *Annals of Statistics,* **20,** 1203–1221.
[17] Mauldin, R. and Sudderth, W. and Williams, S., 1992, Pólya trees and random distributions.

Annals of Statistics, **20,** 1203–1221.
[18] Joe, H., 1996, Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability,* **33,** 664–677.
[19] McDonald, I. L. and Zucchini, W., 1996, *Hidden Markov and other models for discrete-valued time series* (London: Chapman and Hall/CRC Press).
[20] Muliere, P. and Walker, S. G., 1997, A Bayesian Non-parametric approach to survival analysis using Pólya trees. *Scandinavian Journal of Statistics,* **24,** 331–340.
[21] Walker, S. G. and Muliere, P., 1997, Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Annals of Statistics,* **25,** 1762–1780.
[22] Sarno, E., 1998, Dependence structures of Pólya tree autoregressive models. *Statistica LVIII,* **3,** 363–373.
[23] Walden, A. T., Percival, D. B. and McCoy, E. J., 1998, Spectrum Estimation by Wavelet Thresholding of Multitaper Estimators, *IEEE Transactions on Signal Processing,* **46,** 3153–3165.
[24] Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M., 1999, Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society. Series B,* **61,** 485–527.
[25] Pitt, M. K. and Chatfield, C. and Walker, S. G., 2002, Constructing first order autoregressive models via latent processes, *Scandinavian Journal of Statistics,* **29,** 657–663.
[26] McKenzie, Ed., 2003, *Discrete variate time series,* Stochastic processes: modelling and simulation, Handbook of Statist., **21**, 573–606. (Amsterdam: North-Holland)
[27] Mena, R. H. and Walker, S. G., 2005, Stationary Autoregresive models via a Bayesian nonparametric approach, *Journal of Time Series Analisys,* **26,** 789–805.