entropy

MDPI

Article

Regime-Switching Discrete ARMA Models for Categorical Time Series

Christian H. Weiß

Department of Mathematics and Statistics, Helmut Schmidt University, 22043 Hamburg, Germany; weissc@hsu-hh.de; Tel.: +49-40-6541-2779

Received: 18 March 2020; Accepted: 16 April 2020; Published: 17 April 2020

Abstract: For the modeling of categorical time series, both nominal or ordinal time series, an extension of the basic discrete autoregressive moving-average (ARMA) models is proposed. It uses an observation-driven regime-switching mechanism, leading to the family of RS-DARMA models. After having discussed the stochastic properties of RS-DARMA models in general, we focus on the particular case of the first-order RS-DAR model. This RS-DAR(1) model constitutes a parsimoniously parameterized type of Markov chain, which has an easy-to-interpret data-generating mechanism and may also handle negative forms of serial dependence. Approaches for model fitting are elaborated on, and they are illustrated by two real-data examples: the modeling of a nominal sequence from biology, and of an ordinal time series regarding cloudiness. For future research, one might use the RS-DAR(1) model for constructing parsimonious advanced models, and one might adapt techniques for smoother regime transitions.

Keywords: categorical time series; discrete ARMA models; parsimonious Markov chain; regime-switching models

1. Introduction

Since the pioneering textbook on time series analysis by Box & Jenkins [1], this topic has attracted an immense interest in research and applications. To put it more precisely, *real-valued* time series (having a range consisting of real numbers or vectors) have been in the limelight of scientists and practitioners since that time. Besides many approaches for analyzing real-valued time series, also enumerable models have been developed, starting with the basic autoregressive moving-average (ARMA) models [1]. ARMA models are characterized by a linear conditional mean and an autocorrelation function (ACF) satisfying the so-called Yule–Walker equations. Their stochastic properties are well understood, but their actual potential for application is limited. Therefore, many alternatives and extensions have been developed, which cover more realistic time series patterns [2]. For example, if being concerned with a time series exhibiting sudden jumps and a piecewise behavior, then it is more appropriate to consider regime-switching models like the self-exciting threshold (SET) AR models proposed by Tong & Lim [3], also see the survey in Tong [4].

During the last decades, *discrete-valued* time series received more and more attention, see Weiß [5] for a recent survey. Especially count time series, i.e., quantitative time series with a range included in the set $\mathbb{N}_0 = \{0, 1, \ldots\}$ of non-negative integers, have been studied intensively. A large number of models have been developed for this type of discrete data, not only integer-valued counterparts to the basic ARMA model, but also to more advanced models like the aforementioned SETAR models. The latter include the proposals by Möller [6], Möller et al. [7], Monteiro et al. [8], Thyregod et al. [9], Wang et al. [10].

In this article, we consider another type of discrete-valued time series, which is somewhat neglected in the time series literature: *categorical* time series $x_1, \ldots, x_T$ with $T \in \mathbb{N} = \{1, 2, \ldots\}$ and a qualitative range $\mathcal{S}$ consisting of a finite number of categories [5]. There are fundamental

differences between quantitative scales (interval or ratio) and qualitative scales (ordinal or nominal) of measurement, see Table 1 in Stevens [11]. In particular, quantitative scales (such as the aforementioned count data, or real-valued measurements such as temperatures or prices) allow us to use the basic arithmetic operations, which, however, is not permitted for qualitative scales. Qualitative (categorical) ranges $\mathcal{S}$ are further distinguished into the cases that $\mathcal{S}$ exhibits a natural order among the categories (*ordinal* range), and that the categories in $\mathcal{S}$ are unordered (*nominal* range). In what follows, we are interested in both types of categorical time series. To simplify notations, we always assume the possible outcomes to be arranged in a certain order (either lexicographical or natural order), i.e., we denote the range as $\mathcal{S} = \{s_0, s_1, \ldots, s_m\}$ with some $m \in \mathbb{N}$. Categorical time series require a tailor-made treatment in any sense. All the moment-based tools developed for quantitative time series cannot be applied due to the inadmissibility of the basic arithmetic operations. Instead, we have to express the location in terms of the mode (for ordinal time series, also the median can be used), and to use one of the customized measures of dispersion and serial dependence from Table 1, see Klein et al. [12], Weiß [13,14] for further details. Even the basic time series plot can only be done in the ordinal case, whereas we may use the rate evolution graph as a substitute for nominal time series [5].

Table 1. Some measures of dispersion and serial dependence for categorical time series.

	Nominal Range	**Ordinal Range**
Dispersion	*Index of qualitative variation* (*Gini index*):	*Index of ordinal variation*:
	$\text{IQV} = \frac{m+1}{m}\left(1 - \sum_{i=0}^{m} p_i^2\right)$,	$\text{IOV} = \frac{4}{m}\sum_{i=0}^{m-1} f_i(1-f_i)$,
	Entropy:	*Cumulative paired entropy*:
	$\text{En} = \frac{-1}{\ln(m+1)}\sum_{i=0}^{m} p_i \ln p_i$;	$\text{CPE} = \frac{-1}{m \ln 2}\sum_{i=0}^{m-1}\left(f_i \ln f_i + (1-f_i)\ln(1-f_i)\right)$;
	where $p_i = P(X = s_i)$, $\boldsymbol{p} = (p_0, \ldots, p_m)^\top$	where $f_i = P(X \leq s_i)$, $\boldsymbol{f} = (f_0, \ldots, f_{m-1})^\top$
Serial dependence	*Cohen's* κ:	*Ordinal Cohen's* κ:
	$\kappa(h) = \frac{\sum_{i=0}^{m}\left(p_{ii}(h) - p_i^2\right)}{1 - \sum_{i=0}^{m} p_i^2}$,	$\kappa_{\text{ord}}(h) = \frac{\sum_{i=0}^{m-1}\left(f_{ii}(h) - f_i^2\right)}{\sum_{i=0}^{m-1} f_i(1-f_i)}$,
	where	where
	$p_{ij}(h) = P(X_t = s_i, X_{t-h} = s_j)$ for time lag $h \in \mathbb{N}$	$f_{ij}(h) = P(X_t \leq s_i, X_{t-h} \leq s_j)$ for time lag $h \in \mathbb{N}$

Finally, the selection of possible models for categorical time series is yet limited, see Weiß [5]. The quite flexible, higher-order Markov models suffer from a huge number of model parameters, which increases exponentially in the model order and polynomially in the number of categories, $m+1$. The latter might be reduced by amalgamating some categories, but this causes a loss of information. The extremely parsimonious discrete ARMA models by Jacobs & Lewis [15], in contrast, have a rather narrow scope of application, see the discussion in Section 2 for further details. Therefore, as a compromise between flexibility and parsimony, we extend the discrete ARMA models for categorical time series by an observation-driven regime-switching mechanism, see Section 3. For ordinal time series, it might be implemented in analogy to the SET approach for quantitative time series. However, the regime-switching can also be applied to nominal time series to capture, e.g., structures or similarities within the categorical range. As an important special case, we obtain a family of parsimonious Markov chain (MC) models, see Section 4. The application potential of the new model family is demonstrated in Section 5 with two real-data applications, where we also illustrate how model fitting might be done. Finally, we conclude in Section 6 and outline issues for future research.

2. About Discrete ARMA Models

Jacobs & Lewis [15] proposed two families of discrete ARMA models. The first of these families (labeled by the acronym "DARMA") is defined in a nested way and has therefore found less attention in the literature. The second family (labeled as "NDARMA"), in contrast, directly imitates the ordinary ARMA recursion and has been considered in several subsequent works. Both families agree in their boundary cases (i.e., pure AR- and pure MA-type models). We concentrate on the "NDARMA family" in the sequel, and we refer to these models simply as discrete ARMA models. The regime-switching mechanism to be proposed in Section 3 could be applied to the "DARMA family" in an analogous way. According to Weiß & Göb [16], the discrete ARMA models ("NDARMA") by Jacobs & Lewis [15] might be defined as follows (Jacobs & Lewis [15] provide an equivalent definition based on backshift operators).

Definition 1. *Let* $(X_t)_{\mathbb{Z}}$ *and* $(\epsilon_t)_{\mathbb{Z}}$ *be categorical processes with range* $\mathcal{S}$*, where* $(\epsilon_t)_{\mathbb{Z}}$ *is independent and identically distributed (i. i. d.) with marginal distribution* $\boldsymbol{p}$ *(*$\epsilon_t \sim \boldsymbol{p}$*), and where* ϵ_t *is independent of* $(X_s)_{s<t}$*. Let*

$$\boldsymbol{D}_t = (\alpha_{t,1}, \ldots, \alpha_{t,\mathrm{p}}, \beta_{t,0}, \ldots, \beta_{t,\mathrm{q}}) \quad \sim \mathrm{Mult}(1; \boldsymbol{\rho})$$

be i. i. d. multinomial random vectors with $\boldsymbol{\rho} = (\phi_1, \ldots, \phi_{\mathrm{p}}, \varphi_0, \ldots, \varphi_{\mathrm{q}})$*, which are independent of* $(\epsilon_t)_{\mathbb{Z}}$ *and of* $(X_s)_{s<t}$*. So the probabilities in* $\boldsymbol{\rho}$*,* $\phi_1, \ldots, \varphi_{\mathrm{q}} \in (0;1)$*, sum up to one.*

Then, $(X_t)_{\mathbb{Z}}$ *is said to be a* discrete ARMA(p, q) process *if it follows the recursion*

$$X_t = \alpha_{t,1} \cdot X_{t-1} + \ldots + \alpha_{t,\mathrm{p}} \cdot X_{t-\mathrm{p}} + \beta_{t,0} \cdot \epsilon_t + \ldots + \beta_{t,\mathrm{q}} \cdot \epsilon_{t-\mathrm{q}}. \tag{1}$$

(Here, if the range $\mathcal{S}$ *is not numerically coded, then we assume* $0 \cdot s = 0$*,* $1 \cdot s = s$ *and* $s + 0 = s$ *for each* $s \in \mathcal{S}$*.)*

The boundary cases $\mathrm{q} = 0$ and $\mathrm{p} = 0$ are referred to as a *DAR*(p) *process* and *DMA*(q) *process*, respectively. The probability vector $\boldsymbol{p}$ is contained in the $(m+1)$-part unit simplex (recall that the range $\mathcal{S}$ consists of $m+1$ categories),

$$\mathbb{S}_{m+1} = \left\{ \boldsymbol{u} \in (0;1)^{m+1} \,\middle|\, u_0 + \ldots + u_m = 1 \right\},$$

and leads to m model parameters. Analogously, $\boldsymbol{\rho} \in \mathbb{S}_{\mathrm{p}+\mathrm{q}+1}$ leads to $\mathrm{p}+\mathrm{q}$ further model parameters. Based on a Markov-chain representation of the discrete ARMA(p, q) process according to Definition 1, Weiß [17] concluded on the ergodicity of the process as well as on the existence of a unique stationary distribution. The initial distribution for achieving stationarity is obtained by solving the invariance equation corresponding to the Markov-chain representation.

Although being denoted in an "ARMA style", the model recursion of Equation (1) implies that X_t is generated by doing nothing else than simply selecting either one of the past p observations $X_{t-1}, \ldots, X_{t-\mathrm{p}}$, or one of the available $\mathrm{q}+1$ innovations $\epsilon_t, \ldots, \epsilon_{t-\mathrm{q}}$. As a consequence, X_t and ϵ_t have the same stationary marginal distribution, namely $\boldsymbol{p}$, i.e., $P(X_t = s_i) = p_i = P(\epsilon_t = s_i)$ for all $i = 0, \ldots, m$. Furthermore, the random-selection mechanism leads to the following transition probabilities:

$$\begin{aligned} &P(X_t = i_0 \mid X_{t-1} = i_1, \ldots X_{t-\mathrm{p}} = i_{\mathrm{p}}, \epsilon_t = j_0, \epsilon_{t-1} = j_1, \ldots, \epsilon_{t-\mathrm{q}} = j_{\mathrm{q}}) \\ &\quad = \textstyle\sum_{r=1}^{\mathrm{p}} \delta_{i_0 i_r} \phi_r + \delta_{i_0 j_0} \varphi_0 + \sum_{r=1}^{\mathrm{q}} \delta_{i_0 j_r} \varphi_r \\ &\quad =: p_{(\mathrm{p},\mathrm{q})}(i_0 \mid i_1, \ldots, i_{\mathrm{p}}, j_0, \ldots, j_{\mathrm{q}}; \boldsymbol{\rho}), \end{aligned} \tag{2}$$

where empty sums (in case of $\mathrm{q} = 0$ or $\mathrm{p} = 0$) are assumed to take the value 0. Here, $\delta_{i,j}$ denotes the Kronecker delta, which takes the value 1 (0) iff $i = j$ ($i \neq j$).

Despite their quite extraordinary data-generating mechanism, the discrete ARMA processes according to Definition 1 have an ARMA-like serial dependence structure. If serial dependence is expressed in terms of Cohen's κ from Table 1, then the following Yule–Walker equations hold [16]:

$$\kappa(h) \;=\; \sum_{j=1}^{\mathrm{p}} \phi_j \, \kappa(|h-j|) \;+\; \sum_{i=0}^{\mathrm{q}-h} \varphi_{i+h} \, r(i) \qquad \text{for } h \geq 1, \tag{3}$$

where the $r(i)$ are given by $r(i) \;=\; \sum_{j=\max\{0,i-\mathrm{p}\}}^{i-1} \phi_{i-j} \cdot r(j) \;+\; \varphi_i \, \mathbb{1}(0 \leq i \leq \mathrm{q})$. Here, $\mathbb{1}(A)$ denotes the indicator function, which takes the value 1 (0) iff A is true (false). Furthermore, Weiß [17] showed that the discrete ARMA processes are ϕ-mixing with exponentially decreasing weights. As a consequence, one can apply the central limit theorem on p. 200 in Billingsley [18] to establish the asymptotic normality for statistics derived from such a process.

Remark 1. *Equation (3) also applies to the ordinal version of Cohen's κ in Table 1. The reason for this is given by the fact that the bivariate probabilities at lag h, $p_{ij}(h)$, satisfy $p_{ij}(h) \;=\; \big(1-\kappa(h)\big)\, p_i p_j \;+\; \kappa(h)\, \delta_{i,j}\, p_j$ for discrete ARMA processes. As a result, one obtains $f_{ij}(h) \;=\; \big(1-\kappa(h)\big)\, f_i f_j \;+\; \kappa(h)\, f_{\min\{i,j\}}$ in the ordinal case, so $f_{ii}(h) - f_i^2 \;=\; \kappa(h)\, f_i(1-f_i)$. Thus, for discrete ARMA processes, the identity $\kappa(h) = \kappa_{\mathrm{ord}}(h)$ has to hold, whereas these measures usually differ from each other for other data-generating processes. It should be pointed out that further identities with other measures of serial dependence exist, see Weiß [13], Weiß & Göb [16] for details.*

While the discrete ARMA models are very attractive in view of parameter parsimony (only $m + \mathrm{p} + \mathrm{q}$ parameters) and some of its model properties (e.g., Yule–Walker equations for $\kappa(h)$), they suffer from the fact that only positive forms of serial dependence are possible (in the sense that always $\kappa(h) \geq 0$). Furthermore, because of the simple selection mechanism in Equation (1), the sample paths generated by discrete ARMA models are characterized by long constant segments being finished by abrupt changes, which will often be inappropriate for real applications. For quantitative time series, a possible remedy is to use additional variation operators, see Möller & Weiß [19]. However, this solution cannot be applied to qualitative time series. Therefore, in Section 3, the novel regime-switching discrete ARMA models are proposed to offer solutions to both of the above drawbacks, the limitation to positive dependence and the piecewise constant sample paths.

3. Regime-Switching Discrete ARMA Models

Let $(X_t)_{\mathbb{Z}}$ be a categorical process with range $\mathcal{S} = \{s_0, s_1, \ldots, s_m\}$. If the range $\mathcal{S}$ is ordinal, then the states are strictly ordered imposing a distinct structure on $\mathcal{S}$. Even for a nominal range, the states are not necessarily free of any relations, but similarities might exist within it. An example is given by biological sequences such as deoxyribonucleic acid (DNA) and protein sequences. The four DNA bases 'a', 'c', 'g', and 't' (adenine, cytosine, guanine, and thymine, respectively) are divided into the group of pyrimidines (c, t) and purines (a, g). In this sense, c is more similar to t than to a or g. The twenty different amino acids exhibit an even more refined similarity structure, which is visualized by the Venn diagram in Figure 1, see Taylor [20] for further details. If developing a stochastic model for a biological sequence (also see Section 5.1 below), it is reasonable to try to account for the apparent structure within the range.

As a possible solution for this task, let us now introduce a novel regime-switching (RS) extension of the discrete ARMA model. It is defined with respect to a partition of the range $\mathcal{S} = \{s_0, s_1, \ldots, s_m\}$ into K non-empty subsets, where $1 \leq K \leq m+1$. Here, $\mathcal{S}_1, \ldots, \mathcal{S}_K$ constitute a partition of $\mathcal{S}$ iff these sets are pairwise disjoint and satisfy $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_K$. The K regimes shall be used as a means to account for a structure within the categorical range, such as the grouping of the DNA bases in pyrimidines and purines, or the ordering of the categories in the case of an ordinal time series. The current regime is determined by the last (or even earlier) observation: if $X_{t-1} \in \mathcal{S}_k$, then the process is in the kth regime at time t, and the upcoming observation X_t is generated according to a regime-specific model. So we are concerned with an observation-driven ("self-exciting") RS-mechanism, which is in contrast to, e.g., the Hidden-Markov model [21], where the regimes are defined by a latent process. As an example, for the ordinal cloudiness time series to be discussed in Section 5.2, we consider (among others) a two-regime model (so $K = 2$), where the "lower regime" refers to a sky with at most scattered clouds, and the "upper regime" to broken clouds or an even overcast sky. Being in the lower regime, the upcoming cloudiness state follows a different model as if being in the upper regime.

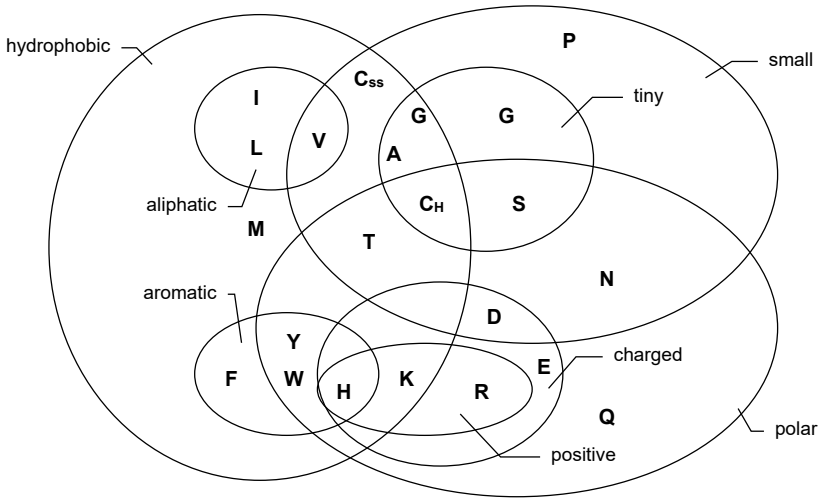


Figure 1. Venn diagram for the classification of amino acids, adapted from Figure 3a in Taylor [20].

Definition 2. *Like in Definition 1, let $(X_t)_{\mathbb{Z}}$ and $(\epsilon_t)_{\mathbb{Z}}$ be categorical processes, where the range $\mathcal{S}$ is partitioned into $\mathcal{S}_1, \ldots, \mathcal{S}_K$, and let $\boldsymbol{D}_t = (\alpha_{t,1}, \ldots, \beta_{t,\mathrm{q}})$ denote the multinomial mixture vectors.*

Let $\boldsymbol{p}_\epsilon^{(1)}, \ldots, \boldsymbol{p}_\epsilon^{(K)} \in \mathbb{S}_{m+1}$ and let $\boldsymbol{\rho}^{(1)}, \ldots, \boldsymbol{\rho}^{(K)} \in \mathbb{S}_{\mathrm{p+q}+1}$ be K state-dependent probability vectors. Then, the regime-switching discrete ARMA(p,q) process ("RS-DARMA") is defined by the recursive scheme

$$\begin{aligned} X_t \;=\; & \alpha_{t,1} \cdot X_{t-1} + \ldots + \alpha_{t,\mathrm{p}} \cdot X_{t-\mathrm{p}} \;+\; \beta_{t,0} \cdot \epsilon_t + \ldots + \beta_{t,\mathrm{q}} \cdot \epsilon_{t-\mathrm{q}} \\ & \textit{with } \boldsymbol{D}_t \sim \mathrm{Mult}(1, \boldsymbol{\rho}_t) \textit{ and } \epsilon_t \sim \boldsymbol{p}_t, \\ & \textit{where } \boldsymbol{\rho}_t = \boldsymbol{\rho}^{(k)} \textit{ and } \boldsymbol{p}_t = \boldsymbol{p}_\epsilon^{(k)} \textit{ iff } X_{t-1} \in \mathcal{S}_k. \end{aligned} \tag{4}$$

Note that the boundary case $K = 1$ leads to the ordinary discrete ARMA model. In Definition 2, we stated the most basic type of RS-condition, where the last observation X_{t-1} determines the current regime. Certainly, one may use other conditions as well, e.g., based on more delayed observations X_{t-d} with $d > 1$.

If $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_K$ denotes a partition, then each $s \in \mathcal{S}$ belongs to exactly one of the $\mathcal{S}_1, \ldots, \mathcal{S}_K$. To simplify notations, we introduce the mapping $\pi : \mathcal{S} \to \{1, \ldots, K\}$ defined by $\pi(s) = k$ iff $s \in \mathcal{S}_k$. So the element $s \in \mathcal{S}$ belongs to the $\pi(s)$-th subset. Then, the RS-condition in Equation (4) can be rewritten as $\boldsymbol{p}_t = \boldsymbol{p}_\epsilon^{(\pi(X_{t-1}))}$ and $\boldsymbol{\rho}_t = \boldsymbol{\rho}^{(\pi(X_{t-1}))}$.

Note that the number of possible partitions of a set of size $m+1$ into K subsets is equal to $S(m+1, K)$, a Stirling number of the second kind [22], p. 5. These numbers can be computed recursively according to $S(m+1, K) = K \cdot S(m, K) + S(m, K-1)$ with $S(0,0) = 1$ and $S(m,0) = 0 = S(0,m)$. The total number of all partitions of $\mathcal{S}$ into non-empty subsets, $B_{m+1} = \sum_{K=1}^{m+1} S(m+1, K)$, is the $(m+1)$-th Bell number [22], p. 5.

Example 1. *Recall the DNA example mentioned in the beginning of Section 3. There, the range consists of $m+1 = 4$ different states, $\mathcal{S} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. Since $m = 3$, there are $B_4 = 15$ possible partitions, namely $1, 7, 6, 1$ different partitions into $K = 1, 2, 3, 4$ subsets, respectively. One of them splits into the purines and pyrimidines, i.e., $\mathcal{S}$ is partitioned into $\mathcal{S}_1 = \{\mathtt{a}, \mathtt{g}\}$ and $\mathcal{S}_2 = \{\mathtt{c}, \mathtt{t}\}$. In this particular case, π maps $\mathtt{a} \mapsto 1$, $\mathtt{c} \mapsto 2$, $\mathtt{g} \mapsto 1$, and $\mathtt{t} \mapsto 2$.*

Let us now discuss the stochastic properties of the RS-DARMA(p,q) process according to Definition 2. The ordinary discrete ARMA's transition probabilities from Equation (2) change to

$$\begin{aligned} & P(X_t = i_0 \mid X_{t-1} = i_1, \ldots X_{t-\mathrm{p}} = i_{\mathrm{p}}, \epsilon_t = j_0, \epsilon_{t-1} = j_1, \ldots, \epsilon_{t-\mathrm{q}} = j_{\mathrm{q}}) \\ & \quad = \; p_{(\mathrm{p},\mathrm{q})}\big(i_0 \mid i_1, \ldots, i_{\mathrm{p}},\; j_0, \ldots, j_{\mathrm{q}};\; \boldsymbol{\rho}^{(\pi(i_1))}\big), \end{aligned} \tag{5}$$

i.e., depending on the outcome $X_{t-1} = i_1$, another set of dependence parameters $\boldsymbol{\rho}^{(\pi(i_1))}$ is plugged-in into Equation (2), otherwise, we just have the ordinary discrete-ARMA structure. Therefore, the MC-representation of the discrete ARMA process as well as the resulting proofs for existence, ergodicity, and mixing properties (ϕ-mixing with exponentially decreasing weights), as provided by Section 2.2 in Weiß [17], can be adapted to the RS-DARMA process, see Appendix A for details.

Example 2. *Let us consider the case* $\mathrm{q} = 0$*, i.e., the purely autoregressive RS-DARMA model according to Definition 2. It can be understood as the direct counterpart to the popular SETAR model. The RS-DAR*(p) *process constitutes a* p*th-order Markov process, where the transition probabilities compute as*

$$\begin{aligned} &P(X_t = s_i \mid X_{t-1} = s_{i_1}, \ldots X_{t-\mathrm{p}} = s_{i_\mathrm{p}}) \\ &\quad = \textstyle\sum_{j=0}^{m} P(X_t = s_i \mid X_{t-1} = s_{i_1}, \ldots X_{t-\mathrm{p}} = s_{i_\mathrm{p}}, \epsilon_t = s_j)\, P(\epsilon_t = s_j \mid X_{t-1} = s_{i_1}) \\ &\quad \overset{(5)}{=} \textstyle\sum_{j=0}^{m} p_{(\mathrm{p},\mathrm{q})}\big(s_i \mid s_{i_1}, \ldots, s_{i_\mathrm{p}},\, s_j;\, \boldsymbol{\rho}^{(\pi(s_{i_1}))}\big)\, p_{\epsilon,j}^{(\pi(s_{i_1}))} \\ &\quad \overset{(2)}{=} \textstyle\sum_{j=0}^{m} \Big(\sum_{r=1}^{\mathrm{p}} \delta_{s_i s_{i_r}} \phi_r^{(\pi(s_{i_1}))} + \delta_{s_i s_j} \varphi_0^{(\pi(s_{i_1}))} \Big)\, p_{\epsilon,j}^{(\pi(s_{i_1}))} \\ &\quad = \textstyle\sum_{r=1}^{\mathrm{p}} \delta_{s_i s_{i_r}} \phi_r^{(\pi(s_{i_1}))} + \varphi_0^{(\pi(s_{i_1}))}\, p_{\epsilon,i}^{(\pi(s_{i_1}))}. \end{aligned}$$

These can now be used for likelihood computations or forecasting purposes.

4. A Class of Parsimonious Markov Chains

A particularly important special case of the RS-DARMA family is obtained by setting $(\mathrm{p},\mathrm{q}) = (1,0)$, because such a RS-DAR$(1)$ process constitutes a parsimoniously parameterized Markov chain (MC), also see Example 2. MCs, in turn, are of great relevance, because (1) such a first-order memory is often sufficient in practice, and because (2) MCs may constitute the starting point for defining more complex time series models [5]. Well-known examples regarding (2) are the so-called mixture transition distribution (MTD) model proposed by Raftery [23], which extends an underlying MC to a higher-order Markov model with only one additional parameter for each increment of the model order, or the hidden-Markov model (HMM), where the observable process is controlled by a latent MC [21]. Recall that an HMM can be interpreted as a parameter-driven RS-model, whereas we consider observation-driven RS-models in this article.

For any of the extensions in the sense of (2), it would be of relevance to start with a maximally parsimonious MC model to keep the overall number of model parameters at a feasible level. A full MC model on $\mathcal{S}$ has $m\,(m+1)$ model parameters, which increases quadratically in m. The lower bound of model parameters is determined through the i. i. d.-case, where only the marginal distribution $\boldsymbol{p}$ has to be specified (which requires m parameters). So any non-i. i. d. model with unspecified marginal distribution must have $\geq m+1$ parameters. The ordinary DAR(1) model with its $m+1$ parameters reaches this lower bound (and also the so-called "Negative Markov model", see, e.g., Weiß [13] for details). It may sometimes be too simplistic for practice, recall the discussion in Section 2. In fact, an MTD(p) model relying on a DAR(1)-MC just leads to a DAR(p) model. Thus, using a (true) RS-DAR(1) model as a base for defining a HMM or MTD model, respectively, might turn out as a reasonable compromise between model flexibility and parameter parsimony.

The model recursion of an ordinary DAR(1) process can be denoted as

$$X_t = \alpha_t\, X_{t-1} + (1-\alpha_t)\, \epsilon_t \quad \text{with } \alpha_t \sim \mathrm{Bin}(1,\phi),\ \epsilon_t \sim \boldsymbol{p}_\epsilon, \tag{6}$$

where $\phi = \phi_1$ and $\boldsymbol{p}_\epsilon = \boldsymbol{p}$ according to Definition 1. Its transition probabilities equal $p_{i|j} = P(X_t = i \mid X_{t-1} = j) = (1-\phi)\, p_{\epsilon,i} + \phi\, \delta_{i,j}$. Let $\mathcal{S} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_K$ be a partition, then the *RS-DAR*(1) *model* is generally defined by

$$X_t = \alpha_t\, X_{t-1} + (1-\alpha_t)\, \epsilon_t \quad \text{with } \alpha_t \sim \mathrm{Bin}(1,\phi^{(\pi(X_{t-1}))}),\ \epsilon_t \sim \boldsymbol{p}_t^{(\pi(X_{t-1}))}, \tag{7}$$

see Example 2. For applications, however, it might be better to impose further restrictions such that the resulting model is better interpretable. Therefore, we shall now propose two special cases of Equation (7), where the regimes affect either the marginal distribution or the dependence parameter.

4.1. Marginal Regimes

A *RS-DAR*(1) *model with respect to the marginals* is defined by

$$\begin{aligned} X_t \;&=\; \alpha_t\, X_{t-1} \;+\; (1-\alpha_t)\,\epsilon_t \quad \text{with } \alpha_t \sim \mathrm{Bin}(1,\phi),\ \epsilon_t \sim \boldsymbol{p}_t, \\ &\text{where } \boldsymbol{p}_t = \boldsymbol{p}_\epsilon^{(k)} \text{ iff } X_{t-1} \in \mathcal{S}_k, \quad \text{i.e., } \boldsymbol{p}_t = \boldsymbol{p}_\epsilon^{(\pi(X_{t-1}))}. \end{aligned} \tag{8}$$

Here, $\boldsymbol{p}_\epsilon^{(1)}, \ldots, \boldsymbol{p}_\epsilon^{(K)} \in \mathbb{S}_{m+1}$ are the K state-dependent probability vectors for the innovations ϵ_t, implying $K\,m + 1$ parameters for model Equation (8). $K = m + 1$ would lead to a full MC model, but then ϕ would not be identifiable anymore. So model Equation (8) requires $K \leq m$. The transition probabilities follow as $p_{i|j} = (1-\phi)\, p_{\epsilon,i}^{(\pi(j))} + \phi\, \delta_{i,j}$, see Example 2. In contrast to the ordinary DAR(1) model, the RS-DAR(1) model Equation (8) also allows for negative serial dependence. This can be obtained by choosing the $\boldsymbol{p}_\epsilon^{(k)}$ such that $p_{\epsilon,i}^{(\pi(i))} \to 0$ for all $i = 0, \ldots, m$.

Example 3. *Let $m = 3$, and define the partition $\mathcal{S}_1 = \{s_0, s_1\}$ and $\mathcal{S}_2 = \{s_2, s_3\}$. Furthermore, let us consider the boundary case*

$$\boldsymbol{p}_\epsilon^{(1)} \;=\; \big(0, 0,\ p^{(1)}, 1-p^{(1)}\big)^\top, \qquad \boldsymbol{p}_\epsilon^{(2)} \;=\; \big(p^{(2)}, 1-p^{(2)},\ 0, 0\big)^\top,$$

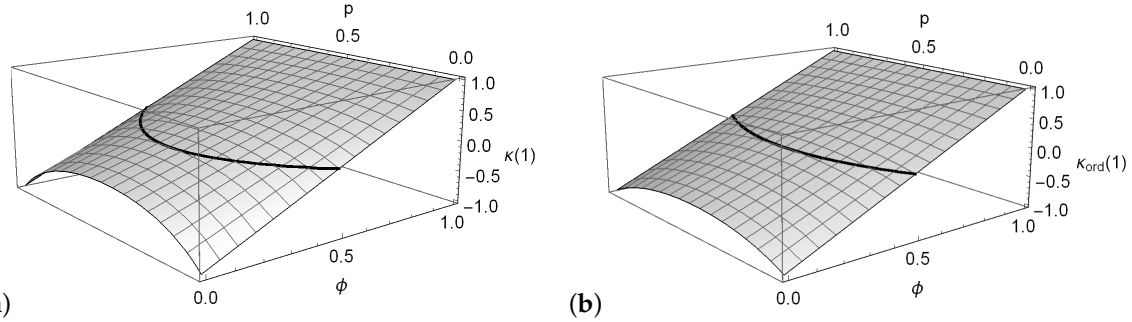
where $p^{(1)}, p^{(2)} \in (0;1)$. Then $p_{\epsilon,i}^{(\pi(i))} = 0$ for all $i = 0, \ldots, m$. The transition matrix $\mathbf{P} = (p_{i|j})_{i,j=0,\ldots,m}$ equals

$$\mathbf{P} \;=\; \begin{pmatrix} \phi & 0 & (1-\phi)\,p^{(2)} & (1-\phi)\,p^{(2)} \\ 0 & \phi & (1-\phi)\,(1-p^{(2)}) & (1-\phi)\,(1-p^{(2)}) \\ (1-\phi)\,p^{(1)} & (1-\phi)\,p^{(1)} & \phi & 0 \\ (1-\phi)\,(1-p^{(1)}) & (1-\phi)\,(1-p^{(1)}) & 0 & \phi \end{pmatrix}.$$

Solving the invariance equation $\mathbf{P}\,\boldsymbol{p} = \boldsymbol{p}$, the stationary marginal distribution, $X_t \sim \boldsymbol{p}$, equals $\boldsymbol{p} = \frac{1}{2}\,\big(p^{(2)}, 1-p^{(2)}, p^{(1)}, 1-p^{(1)}\big)^\top$. From the diagonal of $\mathbf{P}$, it becomes clear that the bivariate probabilities $p_{ii}(1) = \phi\, p_i$. Furthermore, $2\,\sum_{i=0}^m p_i^2 = 1 - p^{(1)}(1-p^{(1)}) - p^{(2)}(1-p^{(2)})$, so one computes Cohen's κ at lag 1, see Table 1, as

$$\kappa(1) \;=\; 1 \;-\; \frac{2\,(1-\phi)}{1 + p^{(1)}(1-p^{(1)}) + p^{(2)}(1-p^{(2)})}.$$

This expression might also become negative, which is illustrated in Figure 2a, where $\kappa(1)$ is plotted against ϕ and p with $p^{(1)} = p^{(2)} = p$.



(**a**) (**b**)

Figure 2. Example 3: Plot of (**a**) $\kappa(1)$ and (**b**) $\kappa_{\mathrm{ord}}(1)$ against ϕ and p with $p^{(1)} = p^{(2)} = p$. The black curves indicate those (ϕ, p), where $\kappa(1) = 0$ or $\kappa_{\mathrm{ord}}(1) = 0$, respectively.

Note that if the range is assumed ordinal, then $\kappa_{\mathrm{ord}}(1)$ differs from $\kappa(1)$ (in contrast to the case of ordinary DAR(1) models). With analogous computations as before, one obtains

$$\kappa_{\mathrm{ord}}(1) \;=\; 1 \;-\; \frac{2\,(1-\phi)\,(2+p^{(2)}-p^{(1)})}{2+p^{(2)}-p^{(1)} \;+\; p^{(1)}(1-p^{(1)})+p^{(2)}(1-p^{(2)})}.$$

But like for $\kappa(1)$*, also* $\kappa_{\mathrm{ord}}(1)$ *might take negative values, see Figure 2b.*

4.2. Dependence Regimes

As a possible alternative to Equation (8), a *RS-DAR*(1) *model with respect to the dependence parameter* is defined by

$$\begin{aligned} X_t \;&=\; \alpha_t\, X_{t-1} \;+\; (1-\alpha_t)\,\epsilon_t \quad \text{with } \alpha_t \sim \mathrm{Bin}(1,\phi_t),\ \epsilon_t \sim \boldsymbol{p}_\epsilon, \\ &\text{where } \phi_t = \phi^{(k)} \text{ iff } X_{t-1} \in \mathcal{S}_k, \quad \text{i.e., } \phi_t = \phi^{(\pi(X_{t-1}))}. \end{aligned} \tag{9}$$

Here, $\phi^{(1)},\ldots,\phi^{(K)} \in (0;1)$ are the K state-dependent dependence parameters, so model Equation (9) has altogether $m+K$ parameters. The transition probabilities compute as $p_{i|j} = (1-\phi^{(\pi(j))})\,p_{\epsilon,i} + \phi^{(\pi(j))}\,\delta_{i,j}$, see Example 2.

Example 4. *As a possible application of model Equation (9), consider an ordinal range* $\mathcal{S}$ *with* $m \geq 2$*, and let us assume individual dependence parameters for each state. More precisely, let us assume the partition* $\mathcal{S} = \bigcup_{k=0}^{m} \mathcal{S}_k$ *with* $\mathcal{S}_k = \{s_k\}$*, and the corresponding* $m+1$ *state-dependent dependence parameters* $\phi^{(0)},\ldots,\phi^{(m)} \in (0;1)$*. So we have* $2m+1$ *model parameters, whereas a full MC would have* $m(m+1)$ *parameters. The transition probabilities equal* $p_{i|j} = (1-\phi^{(j)})\,p_{\epsilon,i} + \phi^{(j)}\,\delta_{i,j}$*. Then* $X_t \sim \boldsymbol{p}$ *with*

$$p_j \;=\; \frac{\frac{p_{\epsilon,j}}{1-\phi^{(j)}}}{\sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}}, \qquad \textit{and} \qquad p_{ij} \;=\; p_{i|j}\,p_j \;=\; \frac{p_{\epsilon,i}\,p_{\epsilon,j} + \delta_{i,j}\,\frac{\phi^{(i)}}{1-\phi^{(i)}}\,p_{\epsilon,i}}{\sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}}.$$

The formula for p_i *is easily verified by computing*

$$\sum_{j=0}^{m} p_{ij} \;=\; \frac{p_{\epsilon,i} \;+\; \frac{\phi^{(i)}}{1-\phi^{(i)}}\,p_{\epsilon,i}}{\sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}} \;=\; \frac{\frac{p_{\epsilon,i}}{1-\phi^{(i)}}}{\sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}}.$$

The cumulative probabilities $\boldsymbol{f}$ *of* X_t *are given by* $f_i = \sum_{j=0}^{i}\frac{p_{\epsilon,j}}{1-\phi^{(j)}} \,/\, \sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}$*. It follows that*

$$p_{ii} \;=\; \frac{\frac{p_{\epsilon,i}}{1-\phi^{(i)}} - p_{\epsilon,i}\,(1-p_{\epsilon,i})}{\sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}}, \qquad f_{ii} \;=\; \sum_{r,s=0}^{i} p_{rs} \;=\; \frac{\sum_{j=0}^{i}\frac{p_{\epsilon,j}}{1-\phi^{(j)}} \;- f_{\epsilon,i}\,(1-f_{\epsilon,i})}{\sum_{l=0}^{m}\frac{p_{\epsilon,l}}{1-\phi^{(l)}}},$$

which can be used to compute the dependence measures $\kappa(1)$ *and* $\kappa_{\mathrm{ord}}(1)$ *from Table 1.*

4.3. Statistical Inference

Let $\boldsymbol{\theta}$ denote the vector of all model parameters, i.e., for the RS-DAR(1) model Equation (8), we have $\boldsymbol{\theta} = \big(\phi,\ p_{\epsilon,1}^{(1)},\ldots,p_{\epsilon,m}^{(K)}\big) \in (0;1)^{K\,m+1}$, whereas $\boldsymbol{\theta} = \big(\phi^{(1)},\ldots,\phi^{(K)},\ p_{\epsilon,1},\ldots,p_{\epsilon,m}\big) \in (0;1)^{K+m}$ for model Equation (9). To estimate $\boldsymbol{\theta}$ from a given time series $x_1,\ldots,x_T$, we use the maximum likelihood (ML) approach. The (conditional) ML estimate $\hat{\boldsymbol{\theta}}$ is obtained by numerically maximizing the log-likelihood function,

$$\ell(\boldsymbol{\theta}) \;=\; \textstyle\sum_{t=2}^{T} \ln p_{x_t|x_{t-1}}(\boldsymbol{\theta}). \tag{10}$$

We denote the maximized log-likelihood by $\ell_{\max} = \frac{T}{T-1}\,\ell\big(\hat{\boldsymbol{\theta}}\big)$, where the factor $\frac{T}{T-1}$ corrects for the conditioning on x_1 [5], p. 236. The existence, consistency, and asymptotic normality are easily established by proving that Condition 5.1 in Billingsley [24] holds. This condition requires that

1. the set $D = \big\{(k,l) \mid p_{k|l}(\boldsymbol{\theta}) > 0\big\}$ does not dependent on $\boldsymbol{\theta}$;
2. each $p_{k|l}(\boldsymbol{\theta})$ has continuous partial derivatives in θ;
3. the Jacobian matrix of $\big(\ldots,\partial p_{k|l}(\boldsymbol{\theta}),\ldots\big)_{(k,l)\in D}$ has full rank, i.e., the rank $n_{\mathrm{model}} = \dim(\boldsymbol{\theta})$;
4. the MC is irreducible.

Since the transition probabilities are quadratic polynomials in the model parameters, part 2 is always satisfied. Part 1 holds by restricting the model parameters to the open interval $(0;1)$, then all $p_{k|l}(\boldsymbol{\theta}) > 0$. This also implies the irreducibility of the transition matrix. Part 3 is ensured by an appropriate design of the model (identifiability of parameters).

Finally, if the model design is not fixed by the considered application scenario, then one will commonly try out multiple types of partitioning (between the boundary cases of an ordinary DAR(1) model and a full MC). In this case, the model selection might be done based on a certain type of information criterion, such as Akaike's information criterion (AIC) or the Bayesian information criterion (BIC), see Burnham & Anderson [25]. These are given by

$$\text{AIC} \;=\; -2\,\ell_{\max} \;+\; 2\,n_{\text{model}}, \qquad \text{BIC} \;=\; -2\,\ell_{\max} \;+\; n_{\text{model}}\,\ln T, \tag{11}$$

respectively. The performance of AIC and BIC if selecting among general Markov models was investigated by Katz [26]. It was shown that only the BIC is consistent while the AIC tends to overfitting.

5. Real-Data Applications

In what follows, we apply the RS-DAR(1) models to two data examples. The first one refers to a DNA sequence, which constitutes a nominal time series (Section 5.1). The second example, in contrast, is about an ordinal time series of cloudiness states (Section 5.2).

5.1. DNA Sequence Modeling

Let us pick up the discussion in the beginning of Section 3 as well as in Example 1, where we considered a nominal DNA sequence having the range $\mathcal{S} = \{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$ (so $m = 3$). Although such a sequence does not constitute a "time" series in the original sense, it is common practice to use models for stochastic processes as a tool for summarizing its main properties, see Churchill [27], Dehnert et al. [28]. In what follows, we consider the DNA sequence of the Bovine leukemia virus (length $T = 8419$), which is published by the National Center for Biotechnology Information at https://www.ncbi.nlm.nih.gov/nuccore/NC_001414?%3Fdb=nucleotide. Its rate evolution graph (a time series plot is not possible for nominal data, see Weiß [5]) and its sample Cohen's κ (recall Table 1) are plotted in Figure 3.

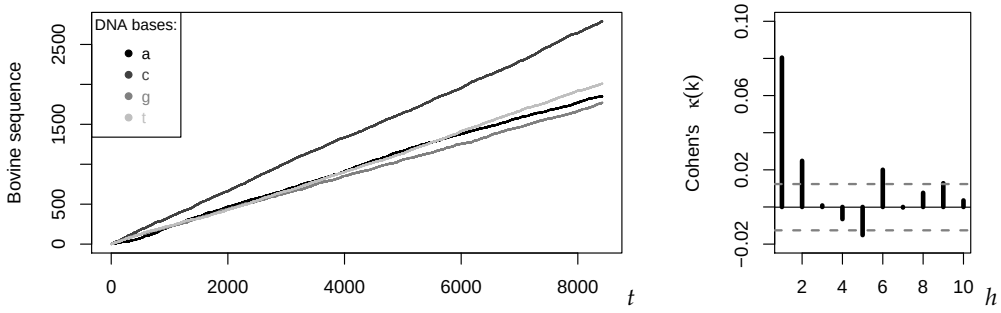


Figure 3. Rate evolution graph of Bovine sequence and of $\hat{\kappa}(h)$ against lag h.

The Bovine sequence was comprehensively analyzed in Section 7.2 of Weiß [5], where several types of models were fitted to the data. The most parsimonious model used was the ordinary DAR(1) model (four parameters), but the best fit was actually obtained by a full MC model (12 parameters). Weiß [5] also tried out a 2-state HMM (eight parameters), which improved over the DAR(1) model but was inferior to the full MC. Nevertheless, it is worth noting that the two hidden states that were constructed during model fitting closely matched the purines (a, g) and pyrimidines (c, t), respectively. This indicates a RS-approach might turn out to be appropriate for the data.

Therefore, we now fit several types of RS-DAR(1) model to these data. As the criterion for model selection, we use the BIC, which is very suitable for time series of large length T. First, we try out

the approach of Section 4.2, where the DAR(1)'s dependence parameter ϕ depends on the current regime. The obtained BIC results are summarized in Table 2. The RS-DAR(1) model with two regimes (purines $\mathcal{S}_1$ and pyrimidines $\mathcal{S}_2$) and, thus, five parameters, leads to a slight improvement compared to the DAR(1) model, but the four-regime model (seven parameters) performs even worse in terms of the BIC. Any of these models is much worse than the full MC, so a regime-dependent dependence structure does not seem to be appropriate for the data.

Table 2. Bovine DNA data: Bayesian information criterion (BIC) of RS-DAR(1) models with respect to dependence parameter ϕ, compared to those of ordinary DAR(1) and Markov chain (MC) model, respectively.

Model	DAR(1)	$\mathcal{S}_1 = \{\mathtt{a},\mathtt{g}\}$, $\mathcal{S}_2 = \{\mathtt{c},\mathtt{t}\}$	$\mathcal{S}_1 = \{\mathtt{a}\}, \mathcal{S}_2 = \{\mathtt{g}\}$, $\mathcal{S}_3 = \{\mathtt{c}\}, \mathcal{S}_4 = \{\mathtt{t}\}$	Full MC
BIC	22927.4	22926.3	22928.9	22824.6

Hence, let us now use the RS-DAR(1) model with regime-dependent innovations' distribution, see Section 4.1. Starting again with the two-regime model (purines vs. pyrimidines; seven parameters), we achieve a considerable improvement over the DAR(1) model, but still do not reach the full MC's BIC, see Table 3. Thus, we try splitting either the purine or the pyrimidine regime, leading to a three-regime model with ten parameters. If we split the purines into $\mathcal{S}_1 = \{\mathtt{a}\}, \mathcal{S}_2 = \{\mathtt{g}\}$, then the BIC deteriorates again. However, if splitting the pyrimidines into $\mathcal{S}_2 = \{\mathtt{c}\}, \mathcal{S}_3 = \{\mathtt{t}\}$, then the BIC improves and even becomes better than that of the full MC. So this type of RS-DAR(1) model is best among all candidate models.

Table 3. Bovine DNA data: BICs of RS-DAR(1) models with respect to marginals $\boldsymbol{p}_\epsilon$, compared to those of ordinary DAR(1) and MC model, respectively.

Model	DAR(1)	$\mathcal{S}_1 = \{\mathtt{a},\mathtt{g}\}$, $\mathcal{S}_2 = \{\mathtt{c},\mathtt{t}\}$	$\mathcal{S}_1 = \{\mathtt{a}\}$, $\mathcal{S}_2 = \{\mathtt{g}\}$, $\mathcal{S}_3 = \{\mathtt{c},\mathtt{t}\}$	$\mathcal{S}_1 = \{\mathtt{a},\mathtt{g}\}$, $\mathcal{S}_2 = \{\mathtt{c}\}$, $\mathcal{S}_3 = \{\mathtt{t}\}$	Full MC
BIC	22927.4	22869.6	22875.0	22822.0	22824.6

Thus, let us analyze this model fit in some more detail. The dependence parameter is estimated as $\hat{\phi} \approx 0.061$, and the three regime-dependent innovations' distributions are

$$\hat{\boldsymbol{p}}_\epsilon^{(1)} \approx (0.244, 0.299, 0.245, 0.211)^\top, \qquad \hat{\boldsymbol{p}}_\epsilon^{(2)} \approx (0.216, 0.352, 0.143, 0.289)^\top,$$
$$\hat{\boldsymbol{p}}_\epsilon^{(3)} \approx (0.181, 0.360, 0.240, 0.219)^\top.$$

This implies, for example, that if $X_{t-1} \in \mathcal{S}_3 = \{\mathtt{t}\}$, then the probability for $\epsilon_t \in \mathcal{S}_2 = \{\mathtt{c}\}$ is quite large, whereas $X_{t-1} \in \mathcal{S}_1 = \{\mathtt{a},\mathtt{g}\}$ also leads to a rather large probability for $\epsilon_t \in \mathcal{S}_1$. These "transition rules" can also be seen from the resulting transition matrix

$$\mathbf{P}_{\text{fit}} \approx \begin{pmatrix} 0.291 & 0.203 & 0.229 & 0.170 \\ 0.281 & 0.392 & 0.281 & 0.338 \\ 0.230 & 0.134 & 0.292 & 0.225 \\ 0.198 & 0.271 & 0.198 & 0.267 \end{pmatrix},$$

which implies the following stationary marginal distribution for X_t:

$$\boldsymbol{p}_{\text{fit}} \approx (0.220, 0.331, 0.210, 0.239)^\top.$$

For the given three-digit rounding, this perfectly agrees with the vector of relative frequencies computed from the data. This excellent agreement between fitted and observed marginal distribution carries over to the measures of dispersion given in Table 1, with an IQV value of ≈ 0.988 and an

entropy of ≈ 0.987. Both values are very close to 1, because the marginal distribution is quite close to a uniform distribution, which is considered as the maximally dispersed scenario for nominal data [13].

Finally, let us do a diagnostic check of the serial dependence structure. On p. 146 in Weiß [5], it was pointed out that the sample value of Cohen's κ at lag 1, $\hat{\kappa}(1) \approx 0.080$, deviates from the corresponding sample value of Cramer's v, $\hat{v}(1) \approx 0.113$. Here, Cramer's v is defined by $v(k) = \left(\frac{1}{m} \sum_{i,j \in \mathcal{S}} \left(p_{ij}(k) - p_i p_j\right)^2 / (p_i p_j)\right)^{1/2}$, constituting a so-called "unsigned" measure of serial dependence [5]. This discrepancy between $\hat{\kappa}(1)$ and $\hat{v}(1)$ contradicts a DAR(1) model, where $\kappa(h)$ and $v(h)$ exactly agree, but it was reproduced by the fitted full MC. For the fitted three-regime model, we obtain $\kappa_{\text{fit}}(1) \approx 0.080$ and $v_{\text{fit}}(1) \approx 0.111$. The fitted RS-DAR(1) model reproduces the discrepancy between κ and v, confirming its adequacy for the Bovine data.

5.2. Cloudiness Time Series

The amount of cloud coverage is measured in "okta", i.e., in eighths of the sky being covered by clouds. Then, a common classification of cloudiness consists of the following five ordinal states (so $m = 4$), ordered from lowest to highest: 'SKC' (sky clear, 0 oktas), 'FEW' (few, 1–2 oktas), 'SCT' (scattered, 3–4 oktas), 'BKN' (broken, 5–7 oktas), and 'OVC' (overcast, 8 oktas). We considered a time series obtained from the "DWD Climate Data Center" offered by the Deutscher Wetterdienst (German Weather Service) at https://cdc.dwd.de/portal/201912031600/mapview. The data refer to the hourly observations of cloudiness at the weather station in Schleswig (a town in the north of Germany) in May 2011. So the time series is of total length $T = 744$. A plot of the data as well as the corresponding sample ordinal Cohen's κ (recall Table 1) are shown in Figure 4. We have a rather strong degree of serial dependence ($\hat{\kappa}(1) \approx 0.741$). The marginal cumulative frequencies are given by $\hat{\boldsymbol{f}} \approx (0.050, 0.253, 0.433, 0.825)^\top$, leading to the dispersion values 0.626 for the sample IOV and 0.689 for the CPE (recall Table 1). Note that in the ordinal case, maximal dispersion does not go along with a uniform distribution, but with an extreme two-point distribution, i.e., $\boldsymbol{f} = (0.5, \ldots, 0.5)^\top$.

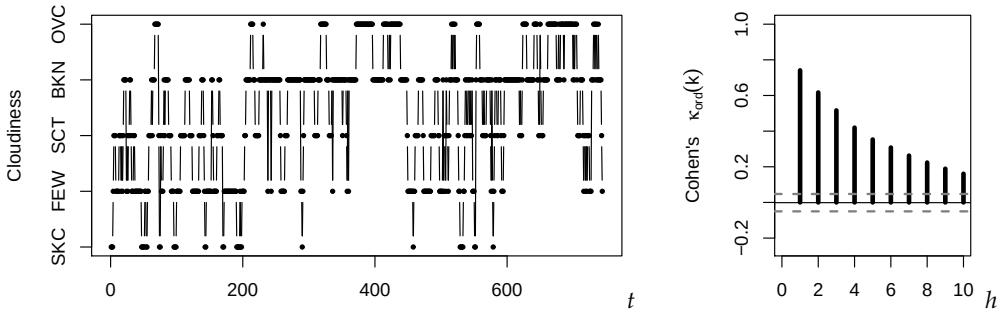


Figure 4. Plot of cloudiness time series and of $\hat{\kappa}_{\text{ord}}(h)$ against lag h.

In analogy to Section 5.1, we now fit types of MC model to the cloudiness data. The most parsimonious model is again the DAR(1) model, but this does not account for the ordinal structure of the range. Therefore, besides the full MC model, we also considered RS-DAR(1) models (with respect to the marginal distribution), which are designed such that they account for the natural order within the range (the RS-DAR(1) model with $m+1$ dependence regimes, see Example 4, performs clearly worse and is, thus, not reported here). This is achieved by imitating the threshold approach for quantitative time series: we considered the partitioning $\mathcal{S}_1 = \{\texttt{SKC}, \texttt{FEW}, \texttt{SCT}\}$, $\mathcal{S}_2 = \{\texttt{BKN}, \texttt{OVC}\}$ corresponding to a threshold at SCT, and the refinement $\mathcal{S}_1 = \{\texttt{SKC}, \texttt{FEW}\}$, $\mathcal{S}_2 = \{\texttt{SCT}\}$, $\mathcal{S}_3 = \{\texttt{BKN}, \texttt{OVC}\}$ corresponding to two thresholds at FEW, SCT. The BICs of the four candidate models are summarized in Table 4.

Table 4. Cloudiness data: BICs of RS-DAR(1) models with respect to marginals $\boldsymbol{p}_\epsilon$, compared to those of ordinary DAR(1) and MC model, respectively.

Model	**DAR(1)**	$\mathcal{S}_1 = \{\texttt{SKC}, \texttt{FEW}, \texttt{SCT}\}$, $\mathcal{S}_2 = \{\texttt{BKN}, \texttt{OVC}\}$	$\mathcal{S}_1 = \{\texttt{SKC}, \texttt{FEW}\}$, $\mathcal{S}_2 = \{\texttt{SCT}\}$, $\mathcal{S}_3 = \{\texttt{BKN}, \texttt{OVC}\}$	**Full MC**
BIC	1423.4	1345.5	1350.1	1392.6

According to the BIC, we select the two-regime model, where the dependence parameter is estimated as $\hat{\phi} \approx 0.547$, and where the two regime-dependent innovations' distributions are

$$\begin{aligned} \hat{\boldsymbol{p}}_\epsilon^{(1)} &\approx (0.091, 0.395, 0.238, 0.269, 0.006)^\top, \\ \hat{\boldsymbol{p}}_\epsilon^{(2)} &\approx (0.000, 0.050, 0.195, 0.560, 0.194)^\top. \end{aligned}$$

According to this fitted model, we mainly produce innovations from the lower regime if staying in the lower regime, and vice versa. This type of "inertia" can also be seen from the right part of Figure 5, where a time series was simulated according to the fitted two-regime model. This sample path looks much more similar to the original time series in Figure 4 than the simulated DAR(1) path in the left part of Figure 5, which does not exhibit a piecewise behavior. The (cumulative) stationary marginal distribution of the fitted two-regime model results as

$$\boldsymbol{f}_{\text{fit}} \approx (0.043, 0.256, 0.471, 0.894)^\top,$$

which is reasonably close to $\hat{\boldsymbol{f}}$. In fact, looking at the corresponding dispersion measures, we get $\text{IOV}_{\text{fit}} \approx 0.575$ and $\text{CPE}_{\text{fit}} \approx 0.640$, which is only slightly below the above sample values. The dependence structure is captured quite well, with $\kappa_{\text{ord, fit}}(1) \approx 0.697$.

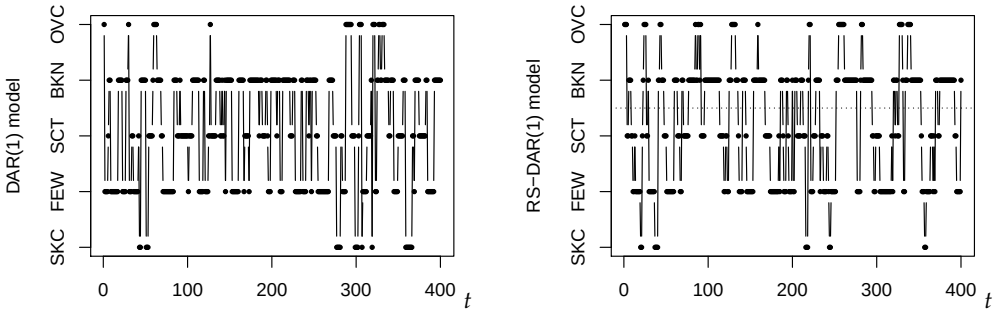


Figure 5. Plots of simulated cloudiness time series, generated according to the fitted DAR(1) model (**left**) and two-regime RS-DAR(1) model ((**right**); regimes separated by dotted line), respectively.

Remark 2. *Since $T = 744$ is not that large, one might also think of using the AIC for model selection. In this case, the three-regime model (*AIC ≈ 1290.1*) would be preferred over the two-regime model (*AIC ≈ 1304.0*). The parameter estimates are $\hat{\phi} \approx 0.592$ and*

$$\begin{aligned} \hat{\boldsymbol{p}}_\epsilon^{(1)} &\approx (0.154, 0.414, 0.302, 0.130, 0.000)^\top, \\ \hat{\boldsymbol{p}}_\epsilon^{(2)} &\approx (0.031, 0.486, 0.001, 0.464, 0.018)^\top, \\ \hat{\boldsymbol{p}}_\epsilon^{(3)} &\approx (0.000, 0.055, 0.204, 0.361, 0.380)^\top. \end{aligned}$$

It is interesting to note that the innovations' distribution of the central regime, $\mathcal{S}_2 = \{\texttt{SCT}\}$, will hardly produce the value SCT *itself. Instead, there is a nearly fifty-fifty chance of falling either above or below this value. The four additional model parameters are outweighed by a closer fit of the marginal distribution, now*

$$\boldsymbol{f}_{\text{fit}} \approx (0.052, 0.297, 0.490, 0.802)^\top$$

with stronger dispersion ($\mathrm{IOV}_{\mathrm{fit}} \approx 0.666$, $\mathrm{CPE}_{\mathrm{fit}} \approx 0.722$*), and by more serial dependence* $\kappa_{\mathrm{ord,\,fit}}(1) \approx 0.739$.

The fitted two-regime model can now be applied to forecasting the cloudiness. The one-step-ahead conditional mode or median is always equal to the given state, i.e., the point forecast for the next hour is equal to the current cloudiness state. But if using 95% prediction intervals for weather forecasting, then more complex rules of the form "$X_t = a \;\Rightarrow\; X_{t+1} \in B$" are obtained:

$$\texttt{SKC} \;\Rightarrow\; \{\texttt{SKC},\texttt{FEW},\texttt{SCT},\texttt{BKN}\}, \qquad \texttt{FEW or SCT} \;\Rightarrow\; \{\texttt{FEW},\texttt{SCT},\texttt{BKN}\},$$
$$\texttt{BKN or OVC} \;\Rightarrow\; \{\texttt{SCT},\texttt{BKN},\texttt{OVC}\}.$$

6. Conclusions

We extended the basic discrete ARMA model of Jacobs & Lewis [15] by an observation-driven regime-switching mechanism, leading to the family of RS-DARMA models. Particular attention was given to two instances of the RS-DAR(1) model, because they constitute an easy-to-interpret type of parsimoniously parameterized MC model. Furthermore, in contrast to the ordinary DAR(1) model, the RS-DAR(1) model may even handle negative forms of serial dependence. Model fitting was illustrated by two real-data examples: a nominal DNA sequence, and an ordinal time series of cloudiness states. Besides such an immediate application of the novel models, it was also pointed out that types of RS-DAR(1) model might serve as the base for constructing parsimonious advanced models, such as MTD models or HMMs. This direction deserves further attention by future research. Furthermore, the special case of regime-switching ordinal processes, which has various applications in practice, should be further elaborated. Due to the close connection to bounded-counts time series [14], one may try to adapt regime-switching techniques known from count processes, e.g., hysteresis zones for enabling smoother regime transitions.

Funding: This research received no external funding.

Acknowledgments: The author thanks the two reviewers for their useful comments on an earlier draft of this article.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Markov-Chain Representation

In analogy to the argumentation in Section 2.2 of Weiß [17], the RS-DARMA(p, q) process can be represented as the homogeneous MC $(\boldsymbol{Z}_t)_{\mathbb{Z}}$ defined by

$$\boldsymbol{Z}_t \;:=\; (X_t;\, X_{t-1},\ldots,X_{t-\mathrm{p}+1},\, \epsilon_t,\ldots,\epsilon_{t-\mathrm{q}+1})^{\top}. \tag{A1}$$

For $\mathrm{q} = 0$, no ϵ-terms need to be included in $\boldsymbol{Z}_t$, but X_t always has to be part of the vector $\boldsymbol{Z}_t$. The corresponding transition probabilities are

$$\begin{aligned} &P\big(\boldsymbol{Z}_t = (i_0,\ldots,i_{\mathrm{p}-1},j_0,\ldots,j_{\mathrm{q}-1}) \,\big|\, \boldsymbol{Z}_{t-1} = (i'_1,\ldots,i'_{\mathrm{p}},j'_1,\ldots,j'_{\mathrm{q}})\big) \\ &\quad= P(X_t = i_0 \mid X_{t-1} = i'_1,\ldots X_{t-\mathrm{p}} = i'_{\mathrm{p}}, \epsilon_t = j_0, \epsilon_{t-1} = j'_1,\ldots,\epsilon_{t-\mathrm{q}} = j'_{\mathrm{q}}) \\ &\qquad\cdot P(\epsilon_t = j_0 \mid X_{t-1} = i'_1) \\ &\qquad\cdot P(X_{t-1} = i_1 \mid X_{t-1} = i'_1)\cdots P(\epsilon_{t-\mathrm{q}+1} = j_{\mathrm{q}-1} \mid \epsilon_{t-\mathrm{q}+1} = j'_{\mathrm{q}-1}) \\ &\quad= p_{j_0}^{(\pi(i'_1))}\cdot\delta_{i_1 i'_1}\cdots\delta_{j_{\mathrm{q}-1}j'_{\mathrm{q}-1}}\cdot p_{(\mathrm{p},\mathrm{q})}\big(i_0 \mid i'_1,\ldots,i'_{\mathrm{p}},\, j_0,j'_1,\ldots,j'_{\mathrm{q}};\, \boldsymbol{\rho}^{(\pi(i'_1))}\big), \end{aligned} \tag{A2}$$

see Equation (5). The crucial point in the derivations of Weiß [17] is the Lemma in Appendix A.1, where it is shown that all $(\mathrm{p}+\mathrm{q}+1)$-step-ahead transition probabilities of $(\boldsymbol{Z}_t)_{\mathbb{Z}}$ are truly positive. From this property, it follows that $(\boldsymbol{Z}_t)_{\mathbb{Z}}$ is primitive, which implies the existence, ergodicity, and the mixing properties of $(X_t)_{\mathbb{Z}}$.

For the RS-DARMA(p, q) model considered here, we compute

$$P(\boldsymbol{Z}_t = \boldsymbol{a} \mid \boldsymbol{Z}_{t-(\mathrm{p+q+1})} = \boldsymbol{b}) \;=\; \sum_{\substack{a_{\max\{\mathrm{p},1\}},\ldots,a_{\mathrm{p+q}},\\ b_{\mathrm{q}},\ldots,b_{\mathrm{p+q}}}} \prod_{j=0}^{\mathrm{p+q}} \Big(P(X_{t-j} = a_j \mid X_{t-j-1} = a_{j+1},\ldots,\epsilon_{t-j} = b_j,\ldots) \cdot P(\epsilon_{t-j} = b_j \mid X_{t-j-1} = a_{j+1},\ldots,\epsilon_{t-j-1} = b_{j+1},\ldots)\Big)$$

for all $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{S}^{\max\{\mathrm{p},1\}+\mathrm{q}}$. The first factor follows from (5) as

$$P(X_{t-j} = a_j \mid X_{t-j-1} = a_{j+1},\ldots,\epsilon_{t-j} = b_j,\ldots) \;=\; p_{(\mathrm{p,q})}\big(a_j \mid a_{j+1},\ldots,a_{j+\mathrm{p}},\, b_j,\ldots,b_{j+\mathrm{q}};\, \boldsymbol{\rho}^{(\pi(a_{j+1}))}\big),$$

the second one is given by

$$P(\epsilon_{t-j} = b_j \mid X_{t-j-1} = a_{j+1},\ldots,\epsilon_{t-j-1} = b_{j+1},\ldots) \;=\; p_{b_j}^{(\pi(a_{j+1}))}.$$

The rest of the proof is done in the same way as in Appendix A.1 of Weiß [17], by using that all $\boldsymbol{p}^{(k)}$ and $\boldsymbol{\rho}^{(k)}$ have only non-zero entries.

References

1. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, 1st ed.; Holden-Day: San Francisco, CA, USA, 1970.
2. Holan, S.H.; Lund, R.; Davis, G. The ARMA alphabet soup: A tour of ARMA model variants. *Stat. Surv.* **2010**, *4*, 232–274. [CrossRef]
3. Tong, H.; Lim, K.S. Threshold autoregression, limit cycles and cyclical data. *J. R. Stat. Soc. Ser. B* **1980**, *42*, 245–292. [CrossRef]
4. Tong, H. Threshold models in time series analysis–30 years on. *Stat. Its Interface* **2011**, *4*, 107–118 [CrossRef]
5. Weiß, C.H. *An Introduction to Discrete-Valued Time Series*; John Wiley & Sons, Inc.: Chichester, UK, 2018.
6. Möller, T.A. Self-exciting threshold models for time series of counts with a finite range. *Stoch. Model.* **2016**, *32*, 77–98. [CrossRef]
7. Möller, T.A.; Silva, M.E.; Weiß, C.H.; Scotto, M.G.; Pereira, I. Self-exciting threshold binomial autoregressive processes. *AStA Adv. Stat. Anal.* **2016**, *100*, 369–400. [CrossRef]
8. Monteiro, M.; Scotto, M.G.; Pereira, I. Integer-valued self-exciting threshold autoregressive processes. *Commun. Stat. Methods* **2012**, *41*, 2717–2737. [CrossRef]
9. Thyregod, P.; Carstensen, J.; Madsen, H.; Arnbjerg-Nielsen, K. Integer valued autoregressive models for tipping bucket rainfall measurements. *Environmetrics* **1999**, *10*, 395–411 [CrossRef]
10. Wang, C.; Liu, H.; Yao, J.-F.; Davis, R.A.; Li, W.K. Self-excited threshold Poisson autoregression. *J. Am. Stat. Assoc.* **2014**, *109*, 777–787. [CrossRef]
11. Stevens, S.S. Measurement, psychophysics and utility. In *Measurement: Definitions and Theories*; Churchman, C.W., Ratoosh, P., Eds.; John Wiley & Sons, Inc.: New York, NY, USA, 1959; pp. 18–63.
12. Klein, I.; Mangold, B.; Doll, M. Cumulative paired ϕ-entropy. *Entropy* **2016**, *18*, 248. [CrossRef]
13. Weiß, C.H. Measures of dispersion and serial dependence in categorical time series. *Econometrics* **2019**, *7*, 17. [CrossRef]
14. Weiß, C.H. Distance-based analysis of ordinal data and ordinal time series. *J. Am. Stat. Assoc.* **2019**. [CrossRef]
15. Jacobs, P.A.; Lewis, P.A.W. Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal.* **1983**, *4*, 19–36. [CrossRef]
16. Weiß, C.H.; Göb, R. Measuring serial dependence in categorical time series. *AStA Adv. Stat. Anal.* **2008**, *92*, 71–89. [CrossRef]
17. Weiß, C.H. Serial dependence of NDARMA processes. *Comput. Stat. Data Anal.* **2013**, *68*, 213–238. [CrossRef]
18. Billingsley, P. *Convergence of Probability Measures*, 2nd ed.; John Wiley & Sons, Inc.: New York, NY, USA, 1999.
19. Möller, T.A.; Weiß, C.H. Generalized discrete ARMA models. *Appl. Stoch. Models Bus. Ind.* **2020**. [CrossRef]
20. Taylor, W.R. The classification of amino acid conservation. *J. Theor. Biol.* **1986**, *119*, 205–218. [CrossRef]

21. Zucchini, W.; MacDonald, I.L.; Langrock, R. *Hidden Markov Models for Time Series: An Introduction Using R*, 2nd ed.; Chapman & Hall/CRC Press: London, UK, 2016.
22. Mansour, T. *Combinatorics of Set Partitions*; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2013.
23. Raftery, A.E. A model for high-order Markov chains. *J. R. Stat. Soc. Ser. B* **1985**, *47*, 528–539. [CrossRef]
24. Billingsley, P. *Statistical Inference for Markov Processes*; University of Chicago Press: Chicago, IL, USA, 1961.
25. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: New York, NY, USA, 2002.
26. Katz, R.W. On some criteria for estimating the order of a Markov chain. *Technometrics* **1981**, *23*, 243–249. [CrossRef]
27. Churchill, G.A. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **1989**, *51*, 79–94. [CrossRef]
28. Dehnert, M.; Helm, W.E.; Hütt, M.-T. A discrete autoregressive process as a model for short-range correlations in DNA sequences. *Physica A* **2003**, *327*, 535–553. [CrossRef]