AStA (2008) 92: 71–89
DOI 10.1007/s10182-008-0055-4

ORIGINAL PAPER

Measuring serial dependence in categorical time series

Christian H. Weiß · Rainer Göb

Received: 30 May 2006 / Accepted: 13 August 2007 / Published online: 6 February 2008

Abstract The analysis of time-indexed categorical data is important in many fields, e.g., in telecommunication network monitoring, manufacturing process control, ecology, etc. Primary interest is in detecting and measuring serial associations and dependencies in such data. For cardinal time series analysis, autocorrelation is a convenient and informative measure of serial association. Yet, for categorical time series analysis an analogous convenient measure and corresponding concepts of weak stationarity have not been provided. For two categorical variables, several ways of measuring association have been suggested. This paper reviews such measures and investigates their properties in a serial context. We discuss concepts of weak stationarity of a categorical time series, in particular of stationarity in association measures. Serial association and weak stationarity are studied in the class of discrete ARMA processes introduced by Jacobs and Lewis (J. Time Ser. Anal. 4(1):19–36, 1983).

Keywords Categorical time series · Serial association · Weak stationarity · Discrete ARMA processes

An intrinsic feature of a time series is that, typically, adjacent observations are dependent. The nature of this dependence among observations of a time series is of considerable practical interest. Time series analysis is concerned with techniques for the analysis of this dependence. (Box et al. 1994, p. 1)

C.H. Weiß (✉) · R. Göb
Institute of Mathematics, Department of Statistics, University of Würzburg, Würzburg, Germany
e-mail: christian.weiss@mathematik.uni-wuerzburg.de

R. Göb
e-mail: goeb@mathematik.uni-wuerzburg.de

1 Introduction

A discrete-time categorical time series is a sequence $(X_t)_{\mathcal{T}}$ of random variables, where the time range can be assumed to be $\mathcal{T} = \mathbb{Z}$, and where the range $\mathcal{V}$ of the X_t consists of a finite number of *unordered* categories (symbols or letters), which may be coded by m pairwise different real numbers $x_1, \ldots, x_m$. As pointed out by the above remark of Box et al. (1994), the analysis of such a time series requires the analysis of serial dependence. In case of a cardinal or binary time series, a convenient tool of this analysis is the (partial) autocorrelation function.

Categorical time series occur in various fields of practice. Besides applications in biological sequence analysis, see Example 1.1, Göb (2006) used several examples from industrial process monitoring and control to illustrate the need for statistical modelling of categorical time series. Applications to language modeling, software usage and musical analysis are also known.

Example 1.1 (Bovine Leukemia Virus) The genome of an organism consists of sequences of the four DNA bases (nucleotides): the pyrimidines thymine ('t') and cytosine ('c'), and the purines adenine ('a') and guanine ('g'). It can be considered as a categorical time series with range $\mathcal{V} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. Consider the genome of the Bovine leukemia virus.[1] for instance, which is of length 8419. Even the first 120 symbols,

```
gaagcgttct cctcctgaga ccctagtgct cagctctcgg tcctgagctc
tcttgctccc gagaccttct ggtcggctat ccggcagcgg tcaggtaagg
caaaccacgg tttggagggt ...,
```

exhibit many runs, indicating the possible presence of serial dependence. The detection and analysis of patterns in genetic sequences is of great use for the recognition of genes, and the understanding of their structures, functionalities and evolutionary roots. Models for categorical time series play an important role in the analysis of biological sequences, with the aim of sequence identification, comparison, characterization and classification. Later, in Example 6.1, we will identify and fit a model to the genome of the Bovine leukemia virus.

At present, modelling of categorical time series is discussed mainly in the knowledge discovery from databases (KDD) community, see Weiss and Hirsh (1998). Models and inference in statistical time series analysis have concentrated on the paradigm of continuous numerical data whereas categorical time series analysis has received relatively little attention. In particular, standard measures for dispersion and serial dependence, or concepts of weak stationarity have not been defined.

In cardinal analysis, serial correlation is built from the bivariate correlation of each two random variables. It is natural to use an analogous approach for categorical analysis. Section 2 discusses general concepts of measures of association and dependence

[1] Source: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi??db=nucleotide\&val=NC_001414.

for categorical variables. Section 3 discusses the concept of weak stationarity and considers measures of serial association for categorical time series. For the particular class of discrete ARMA models introduced by Jacobs and Lewis (1983), measures of association and weak stationarity are investigated in Sects. 4 and 5. The results are used in Sect. 6 for the purpose of model identification and estimation in discrete ARMA processes. Section 7 contains our conclusions.

2 Concepts of dependence

Throughout this section, we consider the following *bivariate categorical scheme*.

Scheme 2.1 (Bivariate Categorical Scheme) Let X, Y be categorical random variables with range $\mathcal{V}_x = \{x_1, \ldots, x_{m_x}\}$ and $\mathcal{V}_y = \{y_1, \ldots, y_{m_y}\}$, respectively, where $x_1, \ldots, x_{m_x}$ and $y_1, \ldots, y_{m_y}$ are m_x and m_y pairwise different values, respectively. The distributions are given by $\mathrm{P}(X = x_i) = p_{x,i}$, $i = 1, \ldots, m_x$, and $\mathrm{P}(Y = y_j) = p_{y,j}$, $j = 1, \ldots, m_y$. The ranges are chosen in such a way that $p_{x,i}, p_{y,j} > 0$ for all $i = 1, \ldots, m_x$, $j = 1, \ldots, m_y$. Let $p_{ij} = \mathrm{P}(X = x_i, Y = y_j)$ be the joint probability, and let $p_{i|j} = \mathrm{P}(X = x_i \mid Y = y_j) = \frac{p_{ij}}{p_{y,j}}$ denote the corresponding conditional probability.

To motivate our approach for defining measures of dependence and association, we review briefly the approaches for measuring dispersion of a categorical random variable, see Appendix A. The definition starts from an intuitive point of view. Large dispersion is associated with high uncertainty about the outcome of X, and accordingly small dispersion is associated with a low uncertainty. In this informal understanding, the 'quantity of uncertainty' about the outcome of X measures the dispersion of X. So, the hallmarks for a measure of dispersion are the basic requirements that a uniform distribution represents maximal dispersion, and a one-point distribution represents minimal dispersion. In the third step, referring to Appendix A, axioms for a measure of dispersion can be defined by considering the exposed hallmarks.

We proceed in an analogous manner to define measures of dependence, considering the extreme cases of *stochastic independence* and *perfect stochastic dependence*.

Definition 2.2 (Dependence of Categorical Random Variables) Let X and Y be categorical random variables as in the model Scheme 2.1.

(i) X and Y are *(stochastically) independent*, iff $p_{ij} = p_{x,i} \cdot p_{y,j}$ (equivalently, $p_{i|j} = p_{x,i}$) for all $i = 1, \ldots, m_x$, $j = 1, \ldots, m_y$.

(ii) X *perfectly depends* on Y, iff for every $j = 1, \ldots, m_y$, the conditional distribution of X, conditioned on $Y = y_j$, is a one-point distribution, i.e., if for every $j = 1, \ldots, m_y$ there exists $\varphi(j) = i_j \in \{1, \ldots, m_x\}$ such that

$$p_{i|j} = \begin{cases} 1 & \text{if } i = i_j = \varphi(j), \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad p_{ij} = \begin{cases} p_{y,j} & \text{if } i = i_j = \varphi(j), \\ 0 & \text{otherwise.} \end{cases}$$

In Appendix B, we list desirable properties (A1)–(A5) of corresponding measures of association. Association measures A satisfying these requirements may also be called measures of dependence. Also several examples of such measures are briefly reviewed in Appendix B.

For given marginal distributions of X and Y, it is always possible to find a joint distribution, such that X and Y are independent. The same is not true for perfect dependence.

Lemma 2.3 (Perfect Dependence) *Let X and Y be categorical random variables as in the model Scheme 2.1.*

(a) *If X depends perfectly on Y, then $m_x \leq m_y$. More specifically, the mapping $\varphi: \{1, \ldots, m_y\} \to \{1, \ldots, m_x\}$ defined by part (ii) of Definition 2.2 is surjective.*
(b) *Let $m_x = m_y$. If X depends perfectly on Y, then φ is even bijective, $p_{y,j} = p_{x,\varphi(j)}$, and Y depends perfectly on X. The propositions (i) "X depends perfectly on Y", and (ii) "Y depends perfectly on X", are equivalent.*
(c) *If X depends perfectly on Y and $m_x < m_y$, then $p_{x,i} = \sum_{j \in \varphi^-(i)} p_{y,j}$, where $\varphi^-(i)$ denotes the preimage of i.*

See Appendix C.1 for the proof of Lemma 2.3. According to Lemma 2.3, perfect dependence is in general a *nonsymmetric* relation. Part (b) of Lemma 2.3 states that perfect dependence is a symmetric relation *between* X and Y only if the ranges of X and Y are of equal size. In the latter case the random variables X and $\varphi(Y)$ are identical with probability 1, and they show the same dispersion. In the situation of part (c), X is determined by Y by lumping some states together.

In some situations, it is desirable to further distinguish between *signed* (*orientated*) association and *unsigned* (*unorientated*) association. In the cardinal case, correlation is a signed measure, where positive and negative directions are distinguished and can be interpreted. The absolute value of the correlation is unsigned and measures only the strength of the relationship. In the general categorical case, the sign or orientation of association or dependence is meaningless. Accordingly, the unsigned measures considered in Appendix B are all nonnegative, where the value 0 is adopted if X and Y are independent, and where the maximum is adopted in case of perfect dependence.

An indispensable prerequisite for signed or orientated association to become meaningful is that the ranges of X and Y are identical, i.e., $m_x = m = m_y$, $\{x_1, \ldots, x_{m_x}\} = \{z_1, \ldots, z_m\} = \{y_1, \ldots, y_{m_y}\}$, also refer to part (b) of Lemma 2.3. In this case, we can distinguish *negative* and *positive* perfect dependence. Desirable properties of corresponding measures are listed in Appendix B, together with Cohen's κ as a concrete example.

Definition 2.4 (Perfect Positive, Negative Dependence) Let X and Y be categorical random variables as in the model Scheme 2.1 with an identical range $\{z_1, \ldots, z_m\}$.

X and Y are called *perfectly positively dependent*, if they are perfectly dependent and if $p_{i|j} = 1$ for $i = j$, i.e., X and Y necessarily adopt the same value. In contrast, if all $p_{i|i} = 0$, they show perfect *negative dependence*.

A related approach is known from the discipline of measuring mobility. The concept of perfect positive dependence includes that of immobility, while independence is interpreted as perfect mobility. For an elementary overview, see Shorrocks (1978).

3 Weak stationarity of categorical time series

We consider the time series $(X_t)_{\mathbb{Z}}$ of categorical variables X_t with identical ranges $\mathcal{V}_X = \{x_1, \ldots, x_{m_x}\}$. Concepts of stationarity are essential for time series analysis. The concept of *strict* (*strong*) stationarity, i.e., the identity of the joint distributions of $X_{t_1}, \ldots, X_{t_m}$ and of $X_{t_1+k}, \ldots, X_{t_m+k}$ for arbitrary $t_1, \ldots, t_m, k$, does not refer to the type of data considered, and is valid for time series of cardinal, ordinal or categorical observations. In contrast, the concept of weak stationarity from cardinal time series analysis, i.e., the invariance of the expectations $E[X_t]$ and all covariances $\mathrm{Cov}[X_t, X_{t+h}]$ in t, is essentially cardinal and inadequate for ordinal or categorical data.

The topic of weak stationarity of categorical time series seems to have received little interest in literature. Stationarity is commonly interpreted as strict stationarity, whereas concepts of weak stationarity are not used. A discussion of weak stationarity of categorical time series is necessary. We list some concepts of categorical stationarity, together with relations among them.

Definition and Theorem 3.1 (Stationarity of a Categorical Time Series) *Let $(X_t)_{\mathbb{Z}}$ be a time series of categorical variables with identical ranges $\mathcal{V}_X = \{x_1, \ldots, x_{m_x}\}$. The following concepts of weak stationarity of $(X_t)_{\mathbb{Z}}$ are defined.*

(i) $C(k)$ stationarity: *$C(k)$ stationarity holds if the k categories with largest probabilities are invariant in t. The probabilities of the categories with the k largest probabilities may vary.*
(ii) $D(k)$ stationarity: *$D(k)$ stationarity holds if the time series $(X_t)_{\mathbb{Z}}$ is $C(k)$ stationary and if, in addition, the k largest among the m_x probabilities $\mathrm{P}(X_t = x_1), \ldots, \mathrm{P}(X_t = x_{m_x})$ are invariant in t.*
(iii) Marginal stationarity: *Marginal stationarity holds if the m_x probabilities $\mathrm{P}(X_t = x_1), \ldots, \mathrm{P}(X_t = x_{m_x})$ are invariant in t.*
(iv) Measure A stationarity: *Measure A stationarity, i.e., stationarity with respect to an association measure A, holds if the associations $A(X_{t-k}, X_t)$ are invariant in t.*
(v) Bivariate stationarity: *Bivariate stationarity holds if the pairwise joint distribution of X_{t-k}, X_t is invariant in t.*

The following relations hold between the above defined concepts:

(a) *Bivariate stationarity $\Rightarrow$ measure A stationarity.*
(b) *Bivariate stationarity $\Rightarrow$ marginal stationarity $\Rightarrow$ $C(k)$, $D(k)$ stationarity.*
(c) *Marginal stationarity $\Longleftrightarrow$ $D(m_x)$ stationarity.*

The definition of $C(k)$ and $D(k)$ stationarity are generalizations of $C(1)$ and $D(1)$ stationarity, introduced by McGee and Harris (2005). $C(k)$, $D(k)$ and marginal sta-

Table 1 Measures of serial dependence for bivariate stationary time series

Measure	Definition	Range
Goodman and Kruskal's τ	$A_v^{(\tau)}(k) = \frac{\sum_{i,j=1}^m \frac{p_{ij}(k)^2}{p_j} - \sum_{i=1}^m p_i^2}{1-\sum_{i=1}^m p_i^2}$	$[0;\,1]$
Goodman and Kruskal's λ	$A_v^{(\lambda)}(k) = \frac{\sum_{j=1}^m \max_i p_{ij}(k) - \max_i p_i}{1-\max_i p_i}$	$[0;\,1]$
Uncertainty coefficient	$A_v^{(u)}(k) = -\frac{\sum_{i,j=1}^m p_{ij}(k) \ln(\frac{p_{ij}(k)}{p_i p_j})}{\sum_{i=0}^m p_i \ln p_i}$	$[0;\,1]$
Pearson measure	$\mathrm{X}_n^2(k) = n \sum_{i,j=1}^m \frac{(p_{ij}(k)-p_i p_j)^2}{p_i p_j}$	$[0;\,n(m-1)]$
Φ^2-measure	$\Phi^2(k) = \frac{\mathrm{X}_n^2(k)}{n}$	$[0;\,m-1]$
Sakoda measure	$p^*(k) = \sqrt{\frac{m\Phi^2(k)}{(m-1)(1+\Phi^2(k))}}$	$[0;\,1]$
Cramer's v	$v(k) = \frac{\Phi(k)}{\sqrt{m-1}}$	$[0;\,1]$
Cohen's κ	$\kappa(k) = \frac{\sum_{j=1}^m (p_{jj}(k)-p_j^2)}{1-\sum_{j=1}^m p_j^2}$	$[-\frac{\sum_{j=1}^m p_j^2}{1-\sum_{j=1}^m p_j^2};\;1]$

tionarity are based on the marginal distribution alone and ignore the aspect of serial association. As in the cardinal case, a reasonable definition of weak stationarity should account for serial association. This might be achieved by adding measure A stationarity to one of the marginal concepts. However, the problem is in choosing an equivalent to the correlation measure in the cardinal case. Appendix B presented a considerable variety of categorical association measures. It is not clear, which of these should be chosen to define a measure A stationarity to be added to the marginal stationarity.

Bivariate stationarity may be a reasonable alternative. It includes marginal and association measure stationarity of any kind, and may be used to characterize weak stationarity in the categorical case. Under bivariate stationarity, all association measures $A(X_{t-k}, X_t)$ are invariant in t and depend only on k. Hence, we can write the kth *order serial association* as $A(k) = A(X_{t-k}, X_t)$. In this notation, the serial versions of the measures of association are presented by Table 1.

All variables in the process $(X_t)_{\mathbb{Z}}$ are assumed to have the same range $\mathcal{V}$. Hence positive and negative kth-order serial association can be measured by Cohen's $\kappa(k)$. Distinguishing positive and negative serial association is important for time series analysis. Sequences with strong positive first-order serial association tend to show long runs of symbols from $\mathcal{V}$. Sequences with perfect positive sth-order serial dependence are s-periodic.

Models for purely categorical time series are rare, for an overview see McKenzie (2003). A sparsely parameterized class of such models are the discrete ARMA models of Jacobs and Lewis (1983). In the Sects. 4, 5 and 6, we shall study the specific class of categorical NDARMA processes. This class exhibits an important and useful property: association measure stationarity and bivariate stationarity are equivalent.

4 NDARMA models: definition and basic properties

We illustrate the usefulness of the measures of serial association introduced before in Sect. 3 by investigating association and dependence in a discrete variant of the familiar cardinal ARMA model. Under the name *discrete ARMA* (*DARMA*) models they have been introduced in several papers by Jacobs and Lewis (1978a, 1978b, 1978c). Jacobs and Lewis (1978a, 1978b, 1978c) discussed special cases. Jacobs and Lewis (1983) introduced the general version of DARMA$(p, q+1)$ models. However, DARMA processes appeared to be unnecessarily complex and difficult to interpret: They *"define the autoregression via an autoregressive tail"*, but *"the autoregression can be made explicit, as in the usual (normal theory) linear processes"* (Jacobs and Lewis 1983, p. 24). So they defined the NDARMA(p, q) *model* as a new discrete ARMA model, which "*is more reminiscent of the linear* ARMA(p, N) *process*" (Jacobs and Lewis 1983, p. 24). We shall concentrate on the latter type of models.

Definition 4.1 (NDARMA Model) Let $(X_t)_{\mathbb{Z}}$, $(\varepsilon_t)_{\mathbb{Z}}$ be categorical time series where all components have identical ranges $\mathcal{V} = \{x_1, \ldots, x_m\}$ with m pairwise different real numbers $x_1, \ldots, x_m$. Let $(\varepsilon_t)_{\mathbb{Z}}$ be i.i.d. with marginal distribution of each ε_t given by $\mathrm{P}(\varepsilon_t = x_j) = \pi_j$ for $j = 1, \ldots, m$, $t \in \mathbb{Z}$. For $t \in \mathbb{Z}$, ε_t is assumed to be independent of $(X_s)_{s<t}$. Let the i.i.d. decision variables $\boldsymbol{D}_t = (\alpha_{1,t}, \ldots, \alpha_{p,t}, \beta_{0,t}, \ldots, \beta_{q,t})$ have a multinomial distribution MULT$(1; \phi_1, \ldots, \phi_p, \varphi_0, \ldots, \varphi_q)$ and let $\boldsymbol{D}_t$ be independent of $(\varepsilon_s)_{\mathbb{Z}}$ and of $(X_s)_{s<t}$.

The process $(X_t)_{\mathbb{Z}}$ is called an NDARMA(p, q) process if it fulfills the recursion

$$X_t = \alpha_{1,t} X_{t-1} + \cdots + \alpha_{p,t} X_{t-p} + \beta_{0,t} \varepsilon_t + \cdots + \beta_{q,t} \varepsilon_{t-q}. \tag{4.1}$$

In case of $q = 0$, the process is called a DAR(p) *process*, and in case of $p = 0$, it is called a DMA(q) process.

Jacobs and Lewis (1983) considered two cases: initialized NDARMA process and strictly stationary NDARMA processes. The following lemma collects some elementary properties of NDARMA processes $(X_t)_{\mathbb{Z}}$ without assuming special cases.

Lemma 4.2 (NDARMA Model) *Let* $(X_t)_{\mathbb{Z}}$, $(\varepsilon_t)_{\mathbb{Z}}$ *be an* NDARMA(p, q) *process as characterized by the Definition* 4.1.

(a) *Equivalently to Definition* 4.1, $(X_t)_{\mathbb{Z}}$ *can be characterized as*

$$X_t = V_t X_{t-A_t} + (1 - V_t)\varepsilon_{t-D_t}, \tag{4.2}$$

where $(\varepsilon_t)_{\mathbb{Z}}$, $(A_t)_{\mathbb{Z}}$, $(D_t)_{\mathbb{Z}}$, $(V_t)_{\mathbb{Z}}$ *are mutually independent i.i.d. processes, and where* ε_t, A_t, D_t, V_t *are independent of* $(X_s)_{s<t}$. *Furthermore,* $\mathrm{P}(V_t = 1) = \phi_1 + \cdots + \phi_p$, $\mathrm{P}(V_t = 0) = \varphi_0 + \cdots + \varphi_q$, $\mathrm{P}(A_t = l) = \phi_l/(\phi_1 + \cdots + \phi_p)$ *for* $l = 1, \ldots, p$, $\mathrm{P}(D_t = m) = \varphi_m/(\varphi_0 + \cdots + \varphi_q)$ *for* $m = 0, \ldots, q$.

(b) *For $t \in \mathbb{Z}$, let $T_i(t)$, $i = 1, 2, \ldots$, be recursively defined by $T_1(t) = t$, $T_{i+1}(t) = T_i(t) - A_{T_i(t)}$, and let*

$$R_t = \begin{cases} \min\{i \geq 1 \mid V_{T_i(t)} = 0\}, & \text{if } \{i \geq 1 \mid V_{T_i(t)} = 0\} \neq \emptyset, \\ +\infty, & \text{if } \{i \geq 1 \mid V_{T_i(t)} = 0\} = \emptyset. \end{cases} \tag{4.3}$$

Then for each $t \in \mathbb{Z}$ we have $R_t < +\infty$ and $X_t = \varepsilon_{R_t}$ with probability 1. The processes $(R_t)_\mathbb{Z}$ and $(\varepsilon_t)_\mathbb{Z}$ are independent among each other.

(c) *For $t_1, \ldots, t_k \in \mathbb{Z}$, $x_{i_1}, \ldots, x_{i_k} \in \mathcal{V}$, we have*

$$\mathrm{P}(X_{t_1} = x_{i_1}, \ldots, X_{t_k} = x_{i_k}) = \pi_{i_1} \cdots \pi_{i_k} \sum_{\substack{s_1 \leq t_1, \ldots, s_k \leq t_k \\ i_l \neq i_m \Rightarrow s_l \neq s_m}} \mathrm{P}(R_{t_1} = s_1, \ldots, R_{t_k} = s_k). \tag{4.4}$$

(d) *For $t \in \mathbb{Z}$, $x_i \in \mathcal{V}$, we have $\mathrm{P}(X_t = x_i) = \pi_i = \mathrm{P}(\varepsilon_t = x_i)$, i.e., $(X_t)_\mathbb{Z}$ is marginally stationary and the distribution of each X_t equals the distribution of the i.i.d. variables ε_s, $s \in \mathbb{Z}$.*

(e) *For $t_1, t_2 \in \mathbb{Z}$, $x_{i_1}, x_{i_2} \in \mathcal{V}$, we have*

$$\mathrm{P}(X_{t_1} = x_{i_1}, X_{t_2} = x_{i_2}) = \pi_{i_1}\pi_{i_2}\mathrm{P}(R_{t_1} \neq R_{t_2}) + \delta_{i_1 i_2}\pi_{i_2}\mathrm{P}(R_{t_1} = R_{t_2}). \tag{4.5}$$

In particular, $(X_t)_\mathbb{Z}$ is bivariately stationary iff for $t_1, t_2 \in \mathbb{Z}$, the probabilities $\mathrm{P}(R_{t_1} = R_{t_2})$ depend on t_1, t_2 only through $|t_1 - t_2|$.

The proof of Lemma 4.2 is given in Appendix C.2.

5 Serial dependence in NDARMA processes

We shall study serial association and dependence in NDARMA processes. In Definition 4.1, the range $\mathcal{V} = \{x_1, \ldots, x_m\}$ was assumed to consist of real numbers. Hence, pro forma autocorrelations can be calculated. The subsequent Lemma 5.1 expresses autocorrelations in terms of probabilities.

Lemma 5.1 (Autocorrelation in NDARMA Processes) *Let $(X_t)_\mathbb{Z}$ be an* NDARMA(p, q) *series as introduced in Definition 4.1, and let $t_1, t_2 \in \mathbb{Z}$. Then we have*

$$\mathrm{Corr}[X_{t_1}, X_{t_2}] = \mathrm{P}(R_{t_1} = R_{t_2}), \tag{5.1}$$

$$\mathrm{P}(X_{t_1} = x_{i_1}, X_{t_2} = x_{i_2}) = \pi_{i_1}\pi_{i_2}\big(1 - \mathrm{Corr}[X_{t_1}, X_{t_2}]\big) + \delta_{i_1 i_2}\pi_{i_1}\,\mathrm{Corr}[X_{t_1}, X_{t_2}]. \tag{5.2}$$

In particular, $\mathrm{Corr}[X_{t_1}, X_{t_2}] \geq 0$.

Jacobs and Lewis (1983, p. 23) proved the validity of (5.1) for DARMA processes. The proof for the NDARMA case is completely analogous. Equation (5.2) follows from (4.5) and (5.1).

Based on Lemma 5.1, Theorem 5.2 shows that the autocorrelation is meaningful under the categorical interpretation, and can be used to express measure A stationarity of NDARMA processes.

Theorem 5.2 (Autocorrelation and Weak Stationarity of NDARMA Processes) *Let* $(X_t)_{\mathbb{Z}}$ *be an* NDARMA(p, q) *series as introduced in Definition* 4.1.

(a) *For* $t_1, t_2 \in \mathbb{Z}$, *the following formulae relate the correlation* $\mathrm{Corr}[X_{t_1}, X_{t_2}]$ *to measures of association, namely Cohen's* κ, *Cramér's* v, *Goodman and Kruskal's* $A_v{}^{(\tau)}$, Φ^2, *Pearson's* X^2, *and Sakoda's* p^*:

$$\begin{aligned}\mathrm{Corr}[X_{t_1}, X_{t_2}] &= \kappa(X_{t_1}, X_{t_2}) = v(X_{t_1}, X_{t_2}) \\ &= \sqrt{A_v^{(\tau)}(X_{t_1}, X_{t_2})} = \sqrt{\frac{\Phi^2(X_{t_1}, X_{t_2})}{m-1}} = d\sqrt{\frac{X_n^2(X_{t_1}, X_{t_2})}{n(m-1)}}, \quad (5.3)\end{aligned}$$

$$p^*(X_{t_1}, X_{t_2}) = \mathrm{Corr}[X_{t_1}, X_{t_2}]\sqrt{\frac{m}{1+(m-1)\,\mathrm{Corr}[X_{t_1}, X_{t_2}]^2}}. \quad (5.4)$$

(b) *The following concepts of weak stationarity of* $(X_t)_{\mathbb{Z}}$ *are equivalent*: (i) *association measure stationarity with respect to one of the association measures* A *from Appendix* B *and* (ii) *bivariate stationarity.*

The proof of Theorem 5.2 is provided in Appendix C.3.

Theorem 5.2 points out three strong and useful properties of NDARMA processes. (i) Except Goodman and Kruskal's λ, the uncertainty coefficient, and the deviance measure, all association measures A discussed in Appendix B are simple one-to-one functions of the correlation. (ii) Each single association measure A leads to the same concept of association measure stationarity in the sense of Definition 3.1. (iii) Association measure stationarity in the sense of (ii) is equivalent to bivariate stationarity, and also equivalent to correlation stationarity. These properties lead to a considerable simplification—both in theory and in modelling and data analysis—of NDARMA processes. The concept of correlation stationarity, which is familiar from cardinal time series analysis, is sufficient for the understanding and for the analysis of weak stationarity of NDARMA processes.

Lemma 5.1 relates autocorrelations and the probabilities $\mathrm{P}(R_{t_1} = R_{t_2})$, $\mathrm{P}(X_{t_1} = x_{i_1}, X_{t_2} = x_{i_2})$. These quantities can be expressed in terms of the model parameters $\phi_1, \ldots, \phi_p, \varphi_0, \ldots, \varphi_q$ used by Definition 4.1: Jacobs and Lewis (1983) showed that the autocorrelations of an NDARMA process satisfy a set of Yule–Walker equations, as in the case of cardinal ARMA models, see Brockwell and Davis (2002), for instance. Let $(X_t)_{\mathbb{Z}}$ be a bivariate stationary NDARMA(p, q) process, and let

$$\rho(k) = \mathrm{Corr}[X_t, X_{t+k}] \quad \text{for } k = 0, 1, 2, \ldots \quad (5.5)$$

be the stationary autocorrelation function. The Yule–Walker equations for the NDARMA(p, q) model established by Jacobs and Lewis (1983) are

$$\rho(k) = \sum_{j=1}^{p} \phi_j \rho\big(|k-j|\big) + \sum_{i=1}^{q-k} \varphi_{i+k} r(i) \quad \text{for } k \geq 1, \quad (5.6)$$

where the correlations $r(i) = \mathrm{Corr}[\epsilon_{t-i}, X_t]$ equal 0 for $i < 0$, and are determined by

$$r(0) = \varphi_0 \quad \text{and} \quad r(i) = \sum_{j=\max(0,i-p)}^{i-1} \phi_{i-j} r(j) + \sum_{j=1}^{q} \delta_{ij} \varphi_j. \tag{5.7}$$

In cardinal time series analysis, partial autocorrelation is a useful tool for model identification, see Box et al. (1994) or Brockwell and Davis (2002). Partial autocorrelations are determined by a system of linear equations: Let

$$\mathbf{R}_k = \big(\rho(|i-j|)\big)_{1 \le i,j \le k} \quad \text{for } k = 1, 2, \ldots \tag{5.8}$$

be the matrix of autocorrelations, and consider the solutions $(a_{k1}, \ldots, a_{kk})$ of the system

$$\mathbf{R}_k (a_{k1}, \ldots, a_{kk})^\top = \big(\rho(1), \ldots, \rho(k)\big)^\top. \tag{5.9}$$

Then $\rho_{\mathrm{p}}(k) = a_{kk}$ is the kth *order partial autocorrelation*. By Lemma 5.1, the autocorrelations are reasonable parameters of an NDARMA process, since they depend on the range only through the bivariate distributions of the X_t. In the same sense, the partial autocorrelations are also reasonable. From a modelling point of view, partial autocorrelations can be used in a way that is analogous to the familiar way of checking for an AR(p) model in cardinal time series analysis, see Brockwell and Davis (2002). Since the Yule–Walker equations of the DAR(p) process, see (5.6), are the same as for the cardinal AR(p) model, it follows immediately that

$$\rho_{\mathrm{p}}(k) = 0 \quad \text{for } k > p \text{ in case of a DAR}(p) \text{ model.}$$

Hence partial autocorrelations can be used to check whether a DAR(p) model is appropriate and of which order p it is.

Remark 5.3 (DAR(p) Process) Since the DAR(p) process is a Markov process of order p with very sparsely parameterized transition probabilities, given by

$$\mathrm{P}(X_t = i_0 \mid X_{t-1} = i_1, \ldots, X_{t-p} = i_p) = \varphi_0 \cdot \pi_{i_0} + \sum_{r=1}^{p} \delta_{i_0 i_r} \cdot \phi_r,$$

see (4.1), one can easily compute the likelihood function. Akaike's information criterion and the Bayesian information criterion (AIC and BIC, respectively), which are based on the maximized log-likelihood function (see Katz 1981) can then be applied to determine the model order p.

6 NDARMA model: estimation and model identification

Consider a categorical time series $(X_t)_{\mathbb{Z}}$. Assume that $(X_t)_{\mathbb{Z}}$ satisfies bivariate stationarity in the sense of Definition 3.1. All serial association measures $A(k) = A(X_t, X_{t+k})$ in Table 1 depend on the range size m_x, the marginal probabilities

$\mathrm{P}(X_s = x_i) = p_i$, and on the bivariate probabilities $\mathrm{P}(X_t = x_i, X_{t+k} = x_j) = p_{ij}(k)$. Estimators $\widehat{A}(k)$ for $A(k)$ are obtained by replacing the probabilities p_i and $p_{ij}(k)$ by the respective estimators $\widehat{p}_i$, $\widehat{p}_{ij}(k)$ in the formulae for the measures $A(k)$ in Table 1. Let $X_1, \ldots, X_T$ be a segment of observations from $(X_t)_{\mathbb{Z}}$, and let

$$N_{ij}(k) = \text{number of pairs } (X_t, X_{t-k}) = (i, j) \text{ in } X_1, \ldots, X_T, \tag{6.1}$$

$$N_i = \text{number of variables } X_t \text{ in the segment } X_1, \ldots, X_T \text{ equal to } i. \tag{6.2}$$

Then unbiased estimators for p_i and $p_{ij}(k)$ are

$$\widehat{p}_i = \frac{N_i}{T} \quad \text{and} \quad \widehat{p}_{ij}(k) := \frac{N_{ij}(k)}{T-k}. \tag{6.3}$$

Model identification is very important in time series analysis. We can use Theorem 5.2 to establish an empirical check for the adequacy of an NDARMA(p, q) model. If an NDARMA(p, q) model is adequate, then by Theorem 5.2 the estimates $\widehat{\kappa}(k)$ for Cohen's κ, $\widehat{v}(k)$ for Cramér's v, and the square root of the estimate $\widehat{A}_v{}^{(\tau)}$ for Goodman and Kruskal's τ should be approximately equal. In case of large differences among the estimates, an NDARMA(p, q) model is inappropriate.

If an NDARMA model seems to be appropriate, the stationary correlation $\rho(k) = \mathrm{Corr}[X_t, X_{t+k}]$ is estimated by $\widehat{\rho}(k) = \widehat{\kappa}(k)$. The model parameters $\phi_1, \ldots, \phi_p, \varphi_0, \ldots, \varphi_q$ used by Definition 4.1 can be estimated by inserting $\widehat{\rho}(k)$ into the Yule–Walker equations of (5.6). By solving the resulting equations, one obtains estimates for $\phi_1, \ldots, \phi_p, \varphi_0, \ldots, \varphi_q$.

Under an NDARMA model, by Theorem 5.2, the stationary correlation $\rho(k)$ may be estimated by $\widehat{\kappa}(k)$, $\widehat{v}(k)$, or by the square root of $\widehat{A}_v^{(\tau)}$. Other measures from Table 1 should not be used for estimating the serial association. Particular caution is required with respect to Goodman and Kruskal's λ. In cases with a dominant state $r \in \mathcal{V}$, dominant in the sense

$$\pi_r\big(1 - \rho(k)\big) \geq \pi_i\big(1 - \rho(k)\big) + \rho(k) \quad \text{for all } i \neq r, \tag{6.4}$$

it follows that

$$\sum_{j=0}^{m} \pi_j \cdot \max_i p_{i|j}(k) = \pi_r\big(1 - \rho(k)\big) \sum_{j=0}^{m} \pi_j + \pi_r \rho(k) = \pi_r, \quad \text{so } A_v^{(\lambda)}(k) = 0.$$

Hence $A_v^{(\lambda)}(k)$ indicates serial independence, independent of $\rho(k)$.

Example 6.1 (Bovine Leukemia Virus) In the following, we analyze the genome of the Bovine leukemia virus, see Example 1.1. The marginal distribution can be estimated by the relative frequencies $\widehat{\pi}_a = 0.220$, $\widehat{\pi}_c = 0.331$, $\widehat{\pi}_g = 0.210$ and $\widehat{\pi}_t = 0.239$. Estimates of some of the measures of serial dependence of Theorem 5.2

Table 2 Parameter estimates for varying model order p

p	Method	$\widehat{\pi}_a$	$\widehat{\pi}_c$	$\widehat{\pi}_g$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	AIC	BIC
0	YW	0.220	0.331	0.210				
	ML	0.220	0.331	0.210			23053	23074
1	YW	0.220	0.331	0.210	0.080			
	ML	0.220	0.331	0.208	0.081		22900	22928
2	YW	0.220	0.331	0.210	0.079	0.018		
	ML	0.219	0.331	0.209	0.079	0.020	22892	22927

up to lag 5 are summarized in the first columns of (6.5).

Lag k	$\widehat{\kappa}(k)$	$\widehat{v}(k)$	$\sqrt{\widehat{A}_v^{(\tau)}(k)}$	$\widehat{\rho}_{\mathrm{p}}(k)$
1	0.0804	0.1134	0.1118	0.0804
2	0.0248	0.0445	0.0447	0.0185
3	0.0008	0.0281	0.0299	−0.0026
4	−0.0065	0.0222	0.0232	−0.0069
5	−0.0151	0.0294	0.0300	−0.0141

(6.5)

Since the estimates for Cohen's $\kappa(k)$, Cramér's $v(k)$, and the square root of Goodman and Kruskal's $A_v^{(\tau)}(k)$ are roughly equal to each other, it is plausible to model the data by an NDARMA model. Based on the estimates $\widehat{\kappa}(k)$ of Cohen's κ, estimates for the partial autocorrelations $\widehat{\rho}_{\mathrm{p}}(k)$ are determined as the solution of (5.9); the results are summarized in the last column of (6.5). The partial autocorrelations $\widehat{\rho}_{\mathrm{p}}(k)$ of (6.5) are about 0 for $k \geq 3$, which implies modelling the genetic sequence by a DAR(p) model with $p \leq 2$.

Assuming an underlying DAR(p) model of order $p \leq 2$, we used two approaches to estimate the corresponding model parameters. First, we computed simple Yule–Walker estimates (YW), i.e., $\boldsymbol{\pi}$ is estimated by the relative frequencies above, and $\phi_1, \ldots, \phi_p$ are estimated by solving the Yule–Walker equations (5.6), with inserted estimates $\widehat{\kappa}(k)$. Second, we numerically maximized the log-likelihood function, leading to ML estimates, see Remark 5.3. This was done using Mathematica, with the YW estimates as initial values.

The results are summarized in Table 2, together with the respective values of the information criteria AIC and BIC (see Katz 1981). It becomes clear that the model assuming serial independence ($p = 0$) performs worst. Both criteria prefer the DAR(2) model, although its BIC is only slightly below that of the DAR(1) model.

7 Conclusion

We reviewed measures of association among categorical variables and applied these quantities for measuring serial association in categorical time series. We discussed

several approaches to defining weak stationarity in a categorical time series. The rationale of these measures and concepts has been demonstrated for the class of NDARMA(p, q) models introduced by Jacobs and Lewis (1983). The following results are important: (i) For NDARMA processes, the correlation is meaningful under a strictly categorical interpretation. (ii) Cohen's κ, Cramér's v, Goodman and Kruskal's ${A_v}^{(\tau)}$, Φ^2, Pearson's X^2, and Sakoda's p^* are equivalent measures. (iii) All of the latter measures are one-to-one functions of the correlation. (iv) Stationarity in one of the latter measures is equivalent to the stationarity of the bivariate distributions. We have demonstrated the usefulness of these results for estimation and model identification.

Future research should concentrate on generalizing the results from NDARMA processes to a wider class of processes. If analogous results can be established for a large class of categorical processes, it may be possible to define a simple convenient measure of serial categorical association in analogy to the autocorrelation familiar from the cardinal case.

Acknowledgements The authors thank the referees for useful comments on an earlier draft, and Prof. Dr. M. Biewen, Department of Statistics, University of Mainz, for drawing their attention to the work on mobility indices.

Appendix A: Measures of dispersion—a brief review

The literature suggested various approaches to an axiomatic definition of a measure of dispersion for a categorical random variable; for an overview see Uschner (1987). Several authors, e.g., Uschner (1987) and Vogel and Kiesl (1999), collected desirable properties of such a measure. The following properties are essential:

(D1) The dispersion measure should have range $[0; u]$ for some $u > 0$. For a *standardized* measure, we should have $u = 1$ independent of the number m of categories.

(D2) The dispersion measure should adopt its minimum value 0 in case of a one-point distribution, and its maximum value u in case of a uniform distribution.

(D3) The dispersion measure should depend on the range $\{x_1, \ldots, x_m\}$ only through the size m of the range and the probability distribution $p_1 = \mathrm{P}(X = x_1), \ldots, \mathrm{P}(X = x_m)$.

Popular standardized measures of dispersion and some elementary properties are summarized in Table 3. For estimators of these measures, see Lehmann and Casella (1998) and Johnson et al. (1997).

Table 3 Standardized measures of dispersion

Standardized measures of dispersion:
• *Gini index*: $v_G(X) := \frac{m}{m-1}(1 - \sum_{j=1}^{m} p_j^2)$.
• *Entropy*: $v_E(X) := -\frac{1}{\ln m} \sum_{j=1}^{m} p_j \ln p_j, 0 \cdot \ln 0 := 0$.
• *Chebycheff dispersion*: $v_C(X) := \frac{m}{m-1}(1 - \max_j p_j)$.
Properties: (D1), (D2), (D3), with range [0; 1].

Appendix B: Measures of association and dependence

Stochastic *dependence* is a concept defined by exclusion of stochastic independence, see Definition 2.2. In contrast, there is no unique definition of *association*. The term 'association' is used if interest is in establishing measures of mutual relationship of variables, see Liebetrau (1983), Goodman and Kruskal (1979), Agresti (1990) or Gibbons (1993). In each particular case, the relation of such measures and the concept of stochastic independence has to be investigated. In the following, we collect essential properties of categorical measures $A(X, Y)$ of association relating association to independence and perfect dependence. These properties are motivated by the arguments of Sect. 2.

B.1 Measures of unsigned association

The following properties are important for unsigned measures, whereas additional properties of signed measures are discussed afterwards.

- (A1) The measure $A(X, Y)$ depends on the ranges $\{x_1, \ldots, x_{m_x}\}$, $\{y_1, \ldots, y_{m_y}\}$ only through m_x, m_y and the distribution parameters $p_{x,i}$, $p_{y,j}$, p_{ij} and is a continuous function thereof.
- (A2a) If X, Y are independent, the association measure $A(X, Y)$ adopts the value 0.
- (A2b) If $A(X, Y) = 0$, then X, Y are independent.
- (A3) Under fixed m_x, m_y and given marginal distributions of X and Y, the measure $A(X, Y)$ has the range $[0; a]$ where $a > 0$.
- (A4a) If X depends perfectly on Y, then the association measure $A(X, Y)$ adopts the maximum of its range.
- (A4b) If Y depends perfectly on X, then the association measure $A(X, Y)$ adopts the maximum of its range.
- (A4c) If the association measure $A(X, Y)$ adopts the maximum of its range, then X depends perfectly on Y, or Y depends perfectly on X.
- (A5) The measure is symmetric in X and Y.

The unsigned measures suggested in Tables 4 and 5 are organized into two categories: measures based on proportional reduction of variation, and measures based on Pearson's χ^2-statistic. Measures of the second group are summarized in Table 5. The proof of the properties given there is simple and based on elementary algebra. Measures of the first group make use of a parametric measure of dispersion ν, refer to Appendix A. Then consider the conditional dispersion $\nu(X|Y = y_k)$, which is obtained by replacing p_j by $p_{j|k} := \mathrm{P}(X = x_j \mid Y = y_k)$ in the definition of ν. Then $E[\nu(X|Y)] = \sum_{k=1}^{m_y} \nu(X|Y = y_k) \cdot \mathrm{P}(Y = y_k)$, and a parametric measure of association may be defined as

$$A_\nu(X|Y) := \frac{\nu(X) - E[\nu(X|Y)]}{\nu(X)} = 1 - \frac{E[\nu(X|Y)]}{\nu(X)}. \tag{B.1}$$

By inserting the dispersion measures introduced in Table 3 into the scheme of formula (B.1), we obtain the specific measures of association summarized in Table 4. For the proof of the properties given there, see Goodman and Kruskal (1979) on the τ and the λ measure, and Theil (1972) on the uncertainty coefficient.

Table 4 Measures of association by proportional reduction of variation

Measuring association by proportional reduction of variation:

• *Goodman and Kruskal's* τ based on the *Gini index*:

$$A_v^{(\tau)}(X|Y) = \frac{\sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\frac{p_{ij}^2}{p_{y,j}} - \sum_{i=1}^{m_x} p_{x,i}^2}{1-\sum_{i=1}^{m_x} p_{x,i}^2} = \frac{\sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\frac{(p_{ij}-p_{x,i}\ p_{y,j})^2}{p_{y,j}}}{1-\sum_{i=1}^{m_x} p_{x,i}^2}.$$

• *Goodman and Kruskal's* λ based on the *Chebycheff dispersion*:

$$A_v^{(\lambda)}(X|Y) = \frac{\sum_{j=1}^{m_y}\max_i p_{ij} - \max_i p_{x,i}}{1-\max_i p_{x,i}}.$$

• The *uncertainty coefficient* based on the *entropy*:

$$A_v^{(u)}(X|Y) = -\frac{\sum_{i=1}^{m_x}\sum_{j=1}^{m_y} p_{ij}\ln(\frac{p_{ij}}{p_{x,i}p_{y,j}})}{\sum_{i=1}^{m_x} p_{x,i}\ln p_{x,i}}.$$

Properties: Goodman and Kruskal's τ and uncertainty coefficient satisfy (A1), (A2a), (A2b), (A3), (A4a), (A4c). Goodman and Kruskal's λ satisfies (A1), (A2a), (A3), (A4a), (A4c). All three fail to satisfy the symmetry required by (A5). The range of these measures is [0; 1].

Table 5 Measures of association derived from sample statistics

Measuring association derived from sample statistics:

• *Pearsons's* X^2 derived from the χ^2*-statistic*:

$$\mathrm{X}_n^2(X,Y) = n\sum_{i=1}^{m_x}\sum_{j=1}^{m_y}\frac{(p_{ij}-p_{x,i}p_{y,j})^2}{p_{x,i}p_{y,j}}.$$

• The Φ^2 *measure* based on Pearson's X^2:

$$\Phi^2(X,Y) = \frac{1}{n}\cdot\mathrm{X}_n^2(X,Y).$$

• *Sakoda's standardized measure*, where $m := \min(m_x, m_y)$:

$$p^*(X,Y) = \sqrt{\frac{m}{m-1}\cdot\Big(1-\frac{1}{1+\Phi^2(X,Y)}\Big)}.$$

• *Cramér's* v:

$$v(X,Y) := \Phi(X,Y)/\sqrt{m-1}, \quad \text{where } m = \min\{m_x, m_y\}.$$

Properties: The measures satisfy all requirements (A1), (A2a), (A2b), (A3), (A4a), (A4b), (A4c), (A5). In particular, defining $m = \min\{m_x, m_y\}$, Pearson's $X^2(X,Y)$ has range $[0; n\cdot(m-1)]$, the Φ^2 measure has range $[0; m-1]$, and Sakoda's measure and Cramér's $v(X,Y)$ have range [0; 1].

B.2 Measures of signed association

For signed measures, see Definition 2.4, the following requirements replace the requirements (A3) and (A4) established above:

Table 6 Measuring signed association

Measuring signed association:
Cohen's κ: $\kappa(X,Y) := \frac{\sum_{j=1}^{m}(p_{jj}-p_{x,j}\,p_{y,j})}{1-\sum_{j=1}^{m}p_{x,j}\,p_{y,j}}$.
Properties: *Cohen's* κ has range $[-\frac{\sum_{j=1}^{m}p_{x,j}\,p_{y,j}}{1-\sum_{j=1}^{m}p_{x,j}\,p_{y,j}};\,1]$ and satisfies (A1), (A2a), (AS3), (AS4a), (AS4b) and (A5). In general, the properties (A2b) and (AS4c) do not hold.

(AS3) The measure has the range $[l;u]$ where $l<0<u$.

(AS4a) In case of perfect negative dependence the measure adopts the minimum l, and in case of perfect positive dependence the measure adopts the maximum u.

(AS4b) If the measure adopts the maximum of its range, then X and Y exhibit perfect positive dependence.

(AS4c) If the measure adopts the minimum of its range, then X and Y exhibit perfect negative dependence.

An example of a signed measure of association is *Cohen's* κ, see Table 6. The proofs of the assertions in Table 6 are evident.

Appendix C: Proofs of theorems and lemmata

C.1 Proof of Lemma 2.3

Let X perfectly depend on Y. Since all $p_{x,i}, p_{y,j}>0$, we have $|\{p_{ij}>0\}| = |\{p_{y,j}>0\}| = |\mathcal{V}_y|$, and since $p_{x,i} = \sum_{j\in\mathcal{V}_y} p_{ij}$, it follows that $|\mathcal{V}_x| = |\{p_{x,i}>0\}| \leq |\mathcal{V}_y|$.

If for a given $i\in\mathcal{V}_x$ we have $p_{ij}=0$ for all $j\in\mathcal{V}_y$, then also $p_{x,i} = \sum_{j\in\mathcal{V}_y} p_{ij} = 0$, which is a contradiction. Hence, there exists a $j\in\mathcal{V}_y$ for all $i\in\mathcal{V}_x$ such that $\varphi(j)=i$, so φ is surjective.

If $|\mathcal{V}_x| = |\mathcal{V}_y| < \infty$, then any surjective mapping from $\mathcal{V}_y$ onto $\mathcal{V}_x$ is also injective. Hence, there is a one-to-one correspondence between $\mathcal{V}_x$ and $\mathcal{V}_y$ such that

$$p_{x,\varphi(j)} = p_{y,j} \quad \text{and} \quad p_{x,i} = p_{y,\varphi^{-1}(i)}.$$

So Y also perfectly depends on X. In contrast, if $|\mathcal{V}_x| < |\mathcal{V}_y|$, then $p_{ij}=0$ for all $j\notin\varphi^{-}(i)$. So $p_{x,i} = \sum_{j\in\mathcal{V}_y} p_{ij} = \sum_{j\in\varphi^{-}(i)} p_{ij} = \sum_{j\in\varphi^{-}(i)} p_{y,j}$. This completes the proof.

C.2 Proof of Lemma 4.2

Part (a) is obvious. From Definition 4.1, we obtain the recursion

$$\mathrm{P}(R_{t-j}=t-n) = \sum_{i=0}^{q}\varphi_i\cdot\delta_{n-j,i} + \sum_{l=1}^{p}\phi_l\cdot\mathrm{P}(R_{t-j-l}=t-n), \quad j,n\geq 0.$$

Therefore, it follows for $m \geq q + k \cdot p$, $k \geq 1$, that

$$\mathrm{P}(R_t \geq t - m) \geq \varphi_\bullet\big(1 + \phi_\bullet + \cdots + \phi_\bullet^k\big) = 1 - \phi_\bullet^{k+1} \underset{k\to\infty}{\longrightarrow} 1,$$

where $\varphi_\bullet = \sum_{i=0}^{q} \varphi_i$ and $\phi_\bullet = \sum_{j=1}^{p} \phi_j$, so part (b) follows. For a proof of parts (d) and (e), we refer to Jacobs and Lewis (1983), and part (c) can be shown in analogy to part (e).

C.3 Proof of Theorem 5.2

From (5.2) we obtain

$$\begin{aligned}\mathrm{P}(X_{t_1} = X_{t_2}) &= \sum_{j=1}^{m} \pi_j^2\big(1 - \mathrm{Corr}[X_{t_1}, X_{t_2}]\big) + \sum_{j=1}^{m} \pi_j \cdot \mathrm{Corr}[X_{t_1}, X_{t_2}] \\ &= \sum_{j=1}^{m} \pi_j^2 + \mathrm{Corr}[X_{t_1}, X_{t_2}]\left(1 - \sum_{j=1}^{m} \pi_j^2\right),\end{aligned}$$

hence

$$\mathrm{Corr}[X_{t_1}, X_{t_2}] = \frac{\mathrm{P}(X_{t_1} = X_{t_2}) - \sum_{j=1}^{m} \pi_j^2}{1 - \sum_{j=1}^{m} \pi_j^2} = \kappa(X_{t_1}, X_{t_2}).$$

Since

$$\mathrm{P}(X_{t_1} = x_i, X_{t_2} = x_j) - \pi_i \pi_j = \pi_j \,\mathrm{Corr}[X_{t_1}, X_{t_2}](\delta_{ij} - \pi_i),$$

we obtain for Pearson's X^2 of Table 5

$$\begin{aligned}\mathrm{X}_n^2(X_{t_1}, X_{t_2}) &= n \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{(\mathrm{P}(X_{t_1} = x_i, X_{t_2} = x_j) - \pi_i \pi_j)^2}{\pi_j \pi_j} \\ &= {}_{(5.2)}\, n \cdot \mathrm{Corr}[X_{t_1}, X_{t_2}]^2 \sum_{i,j=1}^{m} \frac{\pi_j}{\pi_i}(\delta_{ij} - \pi_i)^2 \\ &= n \cdot \mathrm{Corr}[X_{t_1}, X_{t_2}]^2 \left(\sum_{\substack{1 \leq i,j \leq m \\ i \neq j}} \pi_i \pi_j + \sum_{i=1}^{m} (1 - \pi_i)^2 \right) \\ &= n \cdot \mathrm{Corr}[X_{t_1}, X_{t_2}]^2 \left(\sum_{1 \leq i,j \leq m} \pi_i \pi_j + \sum_{i=1}^{m} (1 - 2\pi_i) \right) \\ &= n(m-1)\,\mathrm{Corr}[X_{t_1}, X_{t_2}]^2.\end{aligned}$$

The latter result provides

$$\mathrm{Corr}[X_{t_1}, X_{t_2}] = \sqrt{\frac{\Phi^2(X_{t_1}, X_{t_2})}{m-1}} = \sqrt{\frac{\mathrm{X}_n^2(X_{t_1}, X_{t_2})}{n(m-1)}} = v(X_{t_1}, X_{t_2}),$$

see Table 5. For Goodman and Kruskal's τ, we obtain from Table 4 and (5.2):

$$\begin{aligned} A_{\nu}^{(\tau)}(X_{t_1}, X_{t_2}) &= \frac{\sum_{i,j=1}^{m} \frac{(\mathrm{P}(X_{t_1}=x_i, X_{t_2}=x_j)-\pi_i\pi_j)^2}{\pi_j}}{1-\sum_{i=1}^{m}\pi_i^2} \\ &= \frac{\mathrm{Corr}[X_{t_1}, X_{t_2}]^2}{1-\sum_{i=1}^{m}\pi_i^2}\left(\sum_{i,j\neq i}\pi_i^2\pi_j + \sum_{i=1}^{m}\pi_i(1-\pi_i)^2\right) \\ &= \frac{\mathrm{Corr}[X_{t_1}, X_{t_2}]^2}{1-\sum_{i=1}^{m}\pi_i^2}\sum_{i=1}^{m}\left(\pi_i^2(1-\pi_i)+\pi_i(1-\pi_i)^2\right) \\ &= \frac{\mathrm{Corr}[X_{t_1}, X_{t_2}]^2}{1-\sum_{i=1}^{m}\pi_i^2}\sum_{i=1}^{m}\pi_i(1-\pi_i) = \mathrm{Corr}[X_{t_1}, X_{t_2}]^2. \end{aligned}$$

Equation (5.4) follows from Table 5 and (5.3). This completes the proof of assertion (a) of Theorem 5.2. Assertion (b) follows from assertion (a), Definition 3.1 and Lemma 5.1.

References

Agresti, A.: Categorical Data Analysis. Wiley, New York (1990)

Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis—Forecasting and Control, 3rd edn. Prentice Hall, Englewood Cliffs (1994)

Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting, 2nd edn. Springer, New York (2002)

Gibbons, J.D.: Nonparametric Measures of Association. Quantitative Applications in the Social Sciences. Sage, Thousand Oaks (1993)

Göb, R.: Data mining and statistical control—A review and some links. In: Lenz, Wilrich (eds.) Frontiers in Stat. Qual. Control, vol. 8, pp. 285–308. Physica, Heidelberg (2006)

Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classifications. Springer, New York (1979)

Jacobs, P.A., Lewis, P.A.W.: Discrete time series generated by mixtures. I: Correlational and runs properties. J. R. Stat. Soc. B **40**(1), 94–105 (1978a)

Jacobs, P.A., Lewis, P.A.W.: Discrete time series generated by mixtures. II: Asymptotic properties. J. R. Stat. Soc. B **40**(2), 222–228 (1978b)

Jacobs, P.A., Lewis, P.A.W.: Discrete time series generated by mixtures. III: Autoregressive processes (DAR(p)). Naval Postgraduate School Tech. Report NPS55-78-022 (1978c)

Jacobs, P.A., Lewis, P.A.W.: Stationary discrete autoregressive-moving average time series generated by mixtures. J. Time Ser. Anal. **4**(1), 19–36 (1983)

Johnson, N.L., Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions. Wiley, New York (1997)

Katz, R.W.: On some criteria for estimating the order of a Markov chain. Technometrics **23**(3), 243–249 (1981)

Lehmann, E.L., Casella, G.: Theory of Point Estimation, 2nd edn. Springer, New York (1998)

Liebetrau, A.M.: Measures of Association. Sage, Thousand Oaks (1983)

McGee, M., Harris, I.R.: Coping with nonstationarity in categorical time series. Technical Report SMU-TR-319, 2nd version, Southern Methodist University, Dallas. Download: http://www.smu.edu/statistics/TechReports/TR319.pdf

McKenzie, E.: Discrete variate time series. In: Rao, C.R., Shanbhag, D.N. (eds.) Handbook of Statistics, pp. 573–606. Elsevier, Amsterdam (2003)

Shorrocks, A.F.: The measurement of mobility. Econometrica **46**(5), 1013–1024 (1978)

Theil, H.: Statistical Decomposition Analysis. North-Holland, Amsterdam (1972)

Uschner, H.: Streuungsmessung nominaler Merkmale mit Hilfe von Paarvergleichen. Doctoral Dissertation, Friedrich-Alexander-University Erlangen-Nürnberg(1987)
Vogel, F., Kiesl, H.: Deskriptive und induktive Eigenschaften zweier Streuungsmaße für nominale Merkmale. In: Vogel (ed.): Arbeiten aus der Statistik, Otto-Friedrich-University Bamberg (1999)
Weiss, G.M., Hirsh, H.: Learning to predict rare events in event sequences. In: Proc. of the 4th Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD-98), pp. 359–363 (1998)