MATH STAT

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

Analysis and Modeling of Categorical Time Series

Christian H. Weiß
Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany; **weissc@hsu-hh.de**.

Introduction

During the last years, there has been growing interest in time series $x_1, \dots, x_T$ with a *categorical* range, i.e., with a discrete non-metric range consisting of a finite number $m+1$ of categories with $m \in \mathbb{N}$ (*state space*). If the range exhibits a natural ordering, it is referred to as an *ordinal* range, and as a *nominal* range otherwise. Here, we shall consider this latter, most general case, i.e., even if there would be some ordering, we would not make use of it but assume a finite number of unordered categories. To simplify notations, the range is coded as $\mathcal{S} = \{0, \dots, m\}$. But this does not imply that there is any natural order between the states in $\mathcal{S}$, except a lexicographic order.

Possible applications:

- Manufacturing processes: quality inspection of produced items, which are classified as either $i \in \{1, \dots, m\}$ if item was non-conforming of type i, or as 0 for a conforming item;
- biological sequence analysis: genetic sequences of $m+1=4$ DNA bases, or of $m+1=20$ amino acids;
- part-of-speech tagging: each word of a text is assigned its part of speech (out of finitely many options);
- network monitoring: time series of alarm messages signaling one out of finitely many error types.

Analyzing Categorical Processes

If being concerned with stationary *real-valued* time series, then a huge toolbox for analyzing such time series is readily available and well-known to a broad audience. To mention a few basic approaches, the time series may be visualized by simply plotting the observed values against time, marginal properties such as location and dispersion may be measured in terms of mean/median and variance/quantiles, and serial dependence may be quantified in terms of autocorrelation.

Things change if the available time series is *categorical*. In the ordinal case, a time series plot is still feasible by arranging the possible outcomes in their natural ordering along the Y axis, and location could still be measured by the median. In the purely nominal case as considered here, not even these basic analytic tools are applicable. Therefore, tailor-made solutions are required when analyzing a (stationary) *categorical process* $(X_t)_{\mathbb{Z}}$ with range $\mathcal{S} = \{0, \dots, m\}$. In the sequel, we denote the time-invariant marginal probabilities by $\boldsymbol{\pi} := (\pi_0, \dots, \pi_m)^{\mathrm{T}}$ with $\pi_i := P(X_t = i) \in (0;1)$ and $\pi_0 = 1 - \pi_1 - \dots - \pi_m$. As their sample counterpart, we consider the vector $\hat{\boldsymbol{\pi}}$ of relative frequencies computed from $X_1, \dots, X_T$. The lagged bivariate (conditional) probabilities are denoted by $p_{ij}(k) := P(X_t = i,\ X_{t-k} = j)$ and $p_{i|j}(k) := P(X_t = i \mid X_{t-k} = j)$, respectively, with the empirical counterparts $\hat{p}_{ij}(k)$, $\hat{p}_{i|j}(k)$ being the corresponding relative frequencies of (i,j) within the pairs $(X_{k+1}, X_1), \dots, (X_T, X_{T-k})$.

Although there are a few proposals for a visual analysis of a categorical time series [W08], a reasonable substitute of the simple time series plot is still missing. But a number of non-visual tools are now available.

Location:

The (empirical) *mode* seems to be the only established solution.

Dispersion:

Categorical *dispersion measures* compare the actual marginal distribution with the two possible extremes of a one-point distribution (no dispersion; maximal concentration) and a uniform distribution (maximal dispersion; no concentration).

Several measures have been proposed for this purpose, see the survey in Appendix A of [WG08], among others:

- (empirical) *Gini index*,
  $\nu_{\mathrm{G}} := \frac{m+1}{m}\left(1 - \sum_{i=0}^{m} \pi_i^2\right)$ and $\hat{\nu}_{\mathrm{G}} := \frac{m+1}{m}\frac{T}{T-1}\left(1 - \sum_{i=0}^{m} \hat{\pi}_i^2\right)$;
- (empirical) *entropy*,
  $\nu_{\mathrm{E}} := \frac{-1}{\ln(m+1)} \sum_{i=0}^{m} \pi_i \ln(\pi_i)$ and $\hat{\nu}_{\mathrm{E}} := \frac{-1}{\ln(m+1)} \sum_{i=0}^{m} \hat{\pi}_i \ln(\hat{\pi}_i)$.

Both measures ν_{G} and ν_{E} have the range $[0;1]$, with the value 0 iff being concerned with a one-point distribution, and the value 1 iff having a uniform distribution.

If $(X_t)_{\mathbb{Z}}$ is i.i.d., then $\hat{\nu}_{\mathrm{G}}$ is an exactly unbiased estimator for ν_{G}, which is also asymptotically normally distributed with variance $\frac{4}{T}\left(\frac{m+1}{m}\right)^2 \left(\sum_{j=0}^{m} \pi_j^3 - \left(\sum_{j=0}^{m} \pi_j^2\right)^2\right)$. For serially dependent data stemming from an NDARMA model (see below), at least bias corrections are available. In contrast, $\hat{\nu}_{\mathrm{E}}$ is biased even in the i.i.d. case, and the bias becomes larger if serial dependence is present. Therefore, $\hat{\nu}_{\mathrm{G}}$ appears to be the most preferable measure of categorical dispersion. [W11,W13]

Serial Dependence:

The process $(X_t)_{\mathbb{Z}}$ is said to be *serially independent* at lag $k \in \mathbb{N}$ iff all bivariate probabilities satisfy $p_{ij}(k) = \pi_i\, \pi_j$.

$(X_t)_{\mathbb{Z}}$ is said to be *perfectly serially dependent* at lag $k \in \mathbb{N}$ iff, for any $j \in \mathcal{S}$, the conditional distribution $p_{\cdot|j}(k)$ is a one-point distribution. More precisely, we speak of perfect *positive dependence* iff $p_{i|j}(k) = 1$ exactly for $i = j$ (i.e., we remain in the state reached k times before), while perfect *negative dependence* requires all $p_{i|i}(k) = 0$ (change of state). [WG08] So like positive autocorrelation implies that large values tend to be followed by large values, for instance, positive dependence implies that the process tends to stay in the state it has reached (and vice versa).

Several measures of serial dependence have been proposed, e.g., by [WG08,BS09,BPN12]. If such measures are not able to distinguish between positive and negative dependence, they are referred to as unsigned measures, and as signed measures otherwise.

Possible measures of *unsigned serial dependence* [WG08,W13]:

- (empirical) *Cramer's* v,
  $v(k) := \sqrt{\frac{1}{m}\sum_{i,j=0}^{m} \frac{\left(p_{ij}(k) - \pi_i\pi_j\right)^2}{\pi_i\pi_j}}$, $\hat{v}(k) := \sqrt{\frac{1}{m}\sum_{i,j=0}^{m} \frac{\left(\hat{p}_{ij}(k) - \hat{\pi}_i\hat{\pi}_j\right)^2}{\hat{\pi}_i\hat{\pi}_j}}$;
- (empirical) *Goodman and Kruskal's* τ,
  $\tau(k) := \sqrt{\sum_{i,j=0}^{m} \frac{\left(p_{ij}(k) - \pi_i\pi_j\right)^2}{\pi_j\left(1 - \sum_{r=0}^{m} \pi_r^2\right)}}$, $\hat{\tau}(k) := \sqrt{\sum_{i,j=0}^{m} \frac{\left(\hat{p}_{ij}(k) - \hat{\pi}_i\hat{\pi}_j\right)^2}{\hat{\pi}_j\left(1 - \sum_{r=0}^{m} \hat{\pi}_r^2\right)}}$.

Both $v(k)$ and $\tau(k)$ have range $[0;1]$, with 0 indicating serial independence, and 1 for perfect serial dependence at lag k.

Possible measure of *signed serial dependence* [WG08,W11]:

- (empirical) *Cohen's* κ,
  $\kappa(k) := \frac{\sum_{j=0}^{m}\left(p_{jj}(k) - \pi_j^2\right)}{1 - \sum_{r=0}^{m} \pi_r^2}$, $\hat{\kappa}(k) := \frac{1}{T} + \frac{\sum_{j=0}^{m}\left(\hat{p}_{jj}(k) - \hat{\pi}_j^2\right)}{1 - \sum_{r=0}^{m} \hat{\pi}_r^2}$.

The range of $\kappa(k)$ also includes negative values: $\left[\frac{-\sum_{j=0}^{m} \pi_j^2}{1 - \sum_{r=0}^{m} \pi_r^2}; 1\right]$.
The sign of $\kappa(k)$ goes along with the sign of serial dependence.

The performance of the empirical measures $\hat{v}(k), \hat{\tau}(k), \hat{\kappa}(k)$ for uncovering significant serial dependence was investigated in [W11,W13]. If $(X_t)_{\mathbb{Z}}$ is i.i.d., then $\hat{\kappa}(k)$ is asymptotically normally distributed with mean and variance given by

$$0 + O(T^{-2}) \text{ and } \frac{1}{T}\left(1 - \frac{1+2\sum_{j=0}^{m} \pi_j^3 - 3\sum_{j=0}^{m} \pi_j^2}{\left(1 - \sum_{j=0}^{m} \pi_j^2\right)^2}\right) + O(T^{-2}), \text{ resp.}$$

For $\hat{v}(k)$, the asymptotics $T \cdot m \cdot \hat{v}^2(k) \sim \chi^2_{m^2}$ hold, while $T \cdot \hat{\tau}^2(k)$ follows a quadratic form distribution [W13]. Therefore, $\hat{v}(k)$ is more preferable than $\hat{\tau}(k)$ from a practical point of view. But overall, Cohen's $\hat{\kappa}(k)$ seems to be the measure of choice, having the best performance for uncovering significant serial dependence (w.r.t. both size and power), being well-suited for parameter estimation (see below), and showing many analogies to the well-known (signed) autocorrelation function.

Modeling Categorical Processes

A possible application of the above tools for analyzing categorical time series is the identification and fitting of appropriate models. Perhaps the most obvious approach is to use a p^{th} *order Markov model* for $(X_t)_{\mathbb{Z}}$. In the special case $\mathrm{p} = 1$ (*Markov chain*), the stochastic properties are solely determined by the (1-step) transition probabilities $p_{i|j}$ or the corresponding transition matrix $\mathbf{P} = (p_{i|j})_{i,j}$, respectively. General p^{th} order Markov models, however, have the practical disadvantage of a huge number of parameters, $m \cdot (m+1)^{\mathrm{p}}$.

For this reason, more parsimonious Markov-type models for categorical processes have been proposed in the literature, e.g., the *variable length Markov model* by [BW99], *hidden Markov models* as discussed in [HM09], or the *mixture transition distribution (MTD) model* by [R85]. The latter extends a Markov chain with transition matrix $\mathbf{Q}$ to the MTD(p) model by assuming that $P(X_t = i \mid X_{t-1} = j_1, \dots, X_{t-\mathrm{p}} = j_{\mathrm{p}}) = \sum_{r=1}^{\mathrm{p}} \lambda_r \cdot q_{i|j_r}$ with $\sum_{r=1}^{\mathrm{p}} \lambda_r = 1$ and $\lambda_r \geq 0$ (only $m(m+1) + \mathrm{p} - 1$ parameters).

An even more parsimonious model class, which also allows for non-Markovian forms of serial dependence, are the *new discrete ARMA (NDARMA) models* by [JL83], which are motivated by the standard ARMA models for real-valued processes (and which are equivalent to the ARMA model discussed by [BS09]).

Definition. [WG08] Let $(\epsilon_t)_{\mathbb{Z}}$ be i.i.d. with marginal distribution $\boldsymbol{\pi}$ and, independently, let $\boldsymbol{D}_t = \left(\alpha_{t,1}, \dots, \alpha_{t,\mathrm{p}}, \beta_{t,0}, \dots, \beta_{t,\mathrm{q}}\right)^{\mathrm{T}}$ be a $(\mathrm{p}+\mathrm{q}+1)$-dimensional vector, where exactly one of the components takes the value 1 (either an $\alpha_{t,i}$ with probability ϕ_i or a $\beta_{t,j}$ with probability φ_j; $\phi_1 + \dots + \varphi_{\mathrm{q}} = 1$) and all others are equal to 0. Both ϵ_t and $\boldsymbol{D}_t$ are assumed to be independent of $(X_s)_{s<t}$. Then $(X_t)_{\mathbb{Z}}$ defined by the random mixture

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

is said to be an NDARMA process of order (p, q).

Although written down in an ARMA-like manner, X_t simply chooses either one of $X_{t-1}, \dots, X_{t-\mathrm{p}}$ or $\epsilon_t, \dots, \epsilon_{t-\mathrm{q}}$. Therefore, this approach is applicable to categorical processes. If $\mathrm{q} > 0$, then $(X_t)_{\mathbb{Z}}$ is not Markovian, while the model order $(\mathrm{p}, 0)$ leads to a special type of p^{th} order Markov process, the DAR(p) process.

Generally, the NDARMA process is stationary with marginal distribution $\boldsymbol{\pi}$, and if serial dependence is measured in terms of Cohen's κ, then $\kappa(k)$ satisfies a set of *Yule-Walker-type equations* in analogy to the standard ARMA case [WG08; same relations also hold for $v(k), \tau(k)$]:

$$\kappa(k) = \sum_{j=1}^{\mathrm{p}} \phi_j \cdot \kappa(|k-j|) + \sum_{i=0}^{\mathrm{q}-k} \varphi_{i+k} \cdot r(i) \quad \text{for } k \geq 1,$$

where the $r(i)$ are determined by $r(i) = 0$ for $i < 0$, $r(0) = \varphi_0$, and $r(i) = \sum_{j=\max\{0,i-\mathrm{p}\}}^{i-1} \phi_{i-j} \cdot r(j) + \sum_{j=0}^{\mathrm{q}} \delta_{i,j} \cdot \varphi_j$ for $i > 0$.

This implies to use the empirical version, $\hat{\kappa}(k)$, not only for uncovering significant serial dependence, but also for identifying the model order of an NDARMA process, and for estimating the model parameters in analogy to the method of moments. The asymptotic distribution of $\hat{\kappa}(k)$ for general NDARMA processes was derived in [W13].

References:

[BPN12] Bagnato L, Punzo A, Nicolis O (2012) The autodependogram: a graphical device to investigate serial dependences. *J. Time Ser. Anal*. 33, 233–254.

[BS09] Biswas A, Song PXK (2009) Discrete-valued ARMA processes. *Stat. Probab. Lett.* 79, 1884–1889.

[BW99] Bühlmann P, Wyner AJ (1999) Variable length Markov chains. *Ann. Stat*. 27, 480–513.

[JL83] Jacobs PA, Lewis PAW (1983) Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal*. 4, 19–36.

[R85] Raftery AE (1985) A model for high-order Markov chains. *J. Roy. Stat. Soc. B* 47, 528–539.

[W08] Weiß CH (2008) Visual analysis of categorical time series. *Stat. Methodol.* 5, 56–71.

[W11] Weiß CH (2011) Empirical measures of signed serial dependence in categorical time series. *J. Stat. Comput. Simul*. 81, 411–429.

[W13] Weiß CH (2013) Serial dependence of NDARMA processes. *Comput. Statist. Data Anal*. 68, 213–238.

[WG08] Weiß CH, Göb R (2008) Measuring serial dependence in categorical time series. *Adv. Stat. Anal*. 92, 71–89.

[ZM09] Zucchini W, MacDonald IL (2009) *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC, London.