Categorical Time Series: Analysis, Modelling, Monitoring?

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich **Mathematik**

Christian H. Weiß

Department of Mathematics,

Darmstadt University of Technology

Categorical Time Series

Motivation

Log data of a web server:

Several types of categorical features (IP addresses, web addresses, state codes, etc.).

(→ ENBIS talk in 2006)

Possible applications:

- Measuring success of a web site,
- intrusion detection.

Medical Diagnoses

of an examiner for a certain type of examination.

(→ ENBIS talk in 2009)

Possible applications:

- Compare diagnosis behavior of different examiners,

- compare examiner with a norm profile.

Text:

Sequence of letters, words, word classes, grammatical entities, etc.

Possible applications:

- Language recognition,
- text recognition,
- part-of-speech tagging, etc.

Genetic or protein sequences.

Possible applications:

- Structure analysis to understand functionality of segments of the sequence,
- recognition of similar sequences,
- recognition of common evolutionary roots,
- identification of sequences through comparison to consensus model.

SPC-related examples.

X_t = result of **inspection of item**, with

$X_t = i$ for $i = 1, \ldots, m$ iff item has nonconformity type i,

$X_t = 0$ iff conforming.

Mukhopadhyay (2008): $m = 6$ paint defects of ceiling fan cover ('poor covering', 'bubbles', etc.).

Overall defect category = most predominant defect.

Ye et al. (2002) monitor network traffic data (284 different types of audit events) for intrusion detection.

Categorical Random Variables

Properties

Categorical random variable:

X takes one of **finite** number of **unordered** categories, say $b_0, \ldots, b_m$. E. g.,

- diagnoses $b_0, \ldots, b_m$, or
- parts of speech $b_0, \ldots, b_m$, or
- types of defects $b_0, \ldots, b_m$, or ...

To simplify notations:

Range of X is coded as $\mathcal{V} = \{0, \ldots, m\}$.

X categorical r.v. with range $\mathcal{V} = \{0, \ldots, m\}$

$\Rightarrow P(X = 0) = 1 - \sum_{j=1}^{m} P(X = j)$.

Abbreviate marginal probabilities: $p_i := P(X = i) \in (0; 1)$,
whole distribution determined by m parameters $p_1, \ldots, p_m$.

1$^{\text{st}}$ problem: Number m of parameters might be very large,
in contrast to real-valued or count-data random variables,
no sparse parametric models available yet!

... except trivial cases:
one-point distribution, uniform distribution.

Typical for **cardinal** random variables:

quantify basic properties, e. g.,

- location via mean or median,
- dispersion via variance or quartiles.

But:

- no arithmetic operations for categorical range
  $\Rightarrow$ no mean, variance, etc.
- unordered range $\Rightarrow$ no quantiles.

So how can we quantify

location and dispersion

of a categorical random variable?

Measures of location:

only **mode** of X in use, i. e.,

value $i \in \mathcal{V}$ such that $p_i \geq p_j$ for all $j \in \mathcal{V}$.

Often not uniquely determined (e. g., uniform distribution).

Intuitive understanding of dispersion:

$$X \text{ shows large dispersion}$$

$$\approx$$

$$\text{High uncertainty about the outcome of } X$$

$\Rightarrow$ Uncertainty of categorical random variable?

Two extreme cases:

Uniform distribution:

Maximal uncertainty about the outcome of X.

One-point distribution:

Perfect certainty about the outcome of X.

$\Rightarrow$ Hallmarks for definition of any measure of dispersion!

Contributions in literature (desirable properties, measures), e. g., by Uschner (1987), Vogel & Kiesl (1999).

Common standardized measures of dispersion:

Gini index: $\nu_{\mathsf{G}}(X) \ := \ \dfrac{m+1}{m}\,(1 \ - \ \sum_j p_j^2).$

Entropy: $\nu_{\mathsf{E}}(X) \ := \ -\dfrac{1}{\ln{(m+1)}}\,\sum_j p_j \ln p_j.$

Chebycheff dispersion: $\nu_{\mathsf{C}}(X) \ := \ \dfrac{m+1}{m}\,(1 \ - \ \max_j p_j).$

Important properties of these measures:

- continuous and symmetric functions of $\boldsymbol{p}$,
- range $[0;1]$,
- maximum value 1 in case of uniform distribution,
- minimal value 0 in case of one-point distribution,
- inequality: $\frac{m+1}{m}\,(1\ -\ \min_j p_j)\ \geq\ \nu_{\mathsf{G}}(X)\ \geq\ \nu_{\mathsf{C}}(X)$.

Empirical measures of dispersion

based on sample $X_1, \ldots, X_T$.

Binarization $\boldsymbol{Y}_t \in \{0, 1\}^{m+1}$ of X_t via $Y_{t,i} := \delta_{i,X_t}$.

Unbiased estimator for p_i: $\hat{p}_i(T) \ := \ \frac{1}{T} \cdot \sum_{t=1}^{T} \ Y_{t,i}$.

Abbreviate $\boldsymbol{p} := (p_0, \ldots, p_m)^\top$,

$\hat{\boldsymbol{p}}(T) := \big(\hat{p}_0(T), \ldots, \hat{p}_m(T)\big)^\top \ = \ \frac{1}{T} \cdot \sum_t \ \boldsymbol{Y}_t$,

and $s_k(\boldsymbol{p}) := \sum_j p_j^k$ for $k \in \mathbb{N}$.

Weiß (2011): **Empirical Gini index** defined by

$$\widehat{\nu}_{\mathsf{G}}(X) \;:=\; \tfrac{m+1}{m}\cdot\tfrac{T}{T-1}\cdot\Big(1-s_2\big(\widehat{\boldsymbol{p}}(T)\big)\Big).$$

If computed from i.i.d. data, then

$\widehat{\nu}_{\mathsf{G}}(X)$ is exactly unbiased and asymptotically normally distributed with variance determined by

$$V\Big[1-s_2\big(\widehat{\boldsymbol{p}}(T)\big)\Big] \;=\; \tfrac{4}{T}\cdot\big(s_3(\boldsymbol{p})-s_2^2(\boldsymbol{p})\big) \;+\; O(T^{-2}).$$

Current research:

$\widehat{\nu}_{\mathsf{G}}(X)$ and $\widehat{\nu}_{\mathsf{E}}(X)$ for NDARMA processes (see below).

Categorical Time Series

Terms & Notations

Categorical process:

$(X_t)_{\mathbb{N}}$ with $\mathbb{N} = \{1, 2, \ldots\}$, where each X_t takes one of **finite** number of **unordered** categories.

Categorical time series:

Realizations $(x_t)_{t=1,\ldots,T}$ from $(X_t)_{\mathbb{N}}$.

$(X_t)_{\mathbb{N}}$ said to be **stationary**
if joint distribution of $(X_t, \ldots, X_{t+k})$
independent of t for all $k \in \mathbb{N}_0$.

Notations for time-invariant probabilities:

If $(X_t)_{\mathbb{N}}$ stationary:

- marginal probabilities $p_i := P(X_t = i) \in (0;1)$.
  $\boldsymbol{p} := (p_0, \ldots, p_m)^\top$, and
  $s_k(\boldsymbol{p}) := \sum_j p_j^k$ for $k \in \mathbb{N}$; obviously $s_1(\boldsymbol{p}) = 1$.

- bivariate probabilities $p_{ij}(k) := P(X_t = i, X_{t-k} = j)$,
  conditional probabilities $p_{i|j}(k) := P(X_t = i \mid X_{t-k} = j)$.

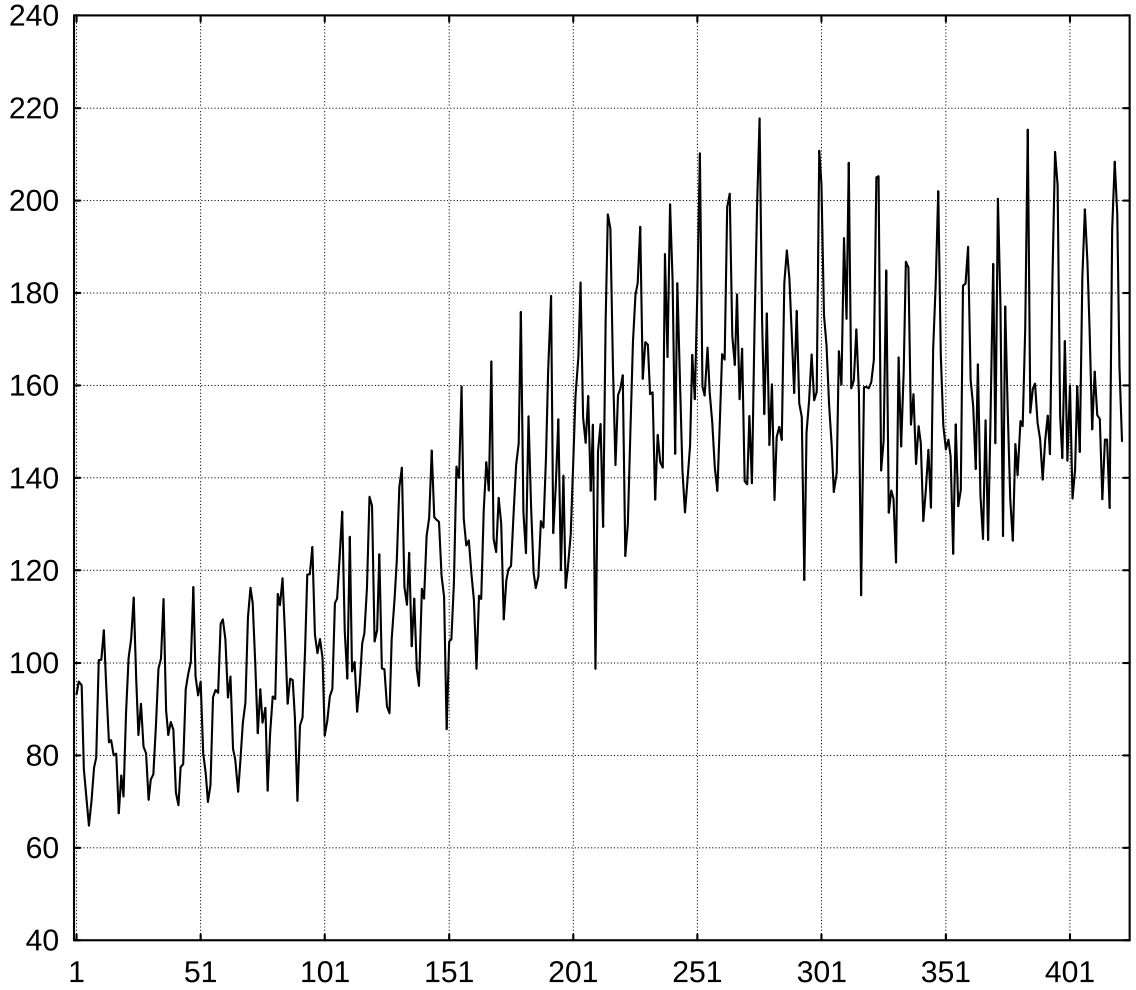
Categorical Time Series

Visual Analysis

"Little existing work deals directly with categorical time series analysis, and much less deals with the visualization of categorical time series." (Ribler, 1997, p. 11)

Several visual tools for real-valued time series:
line plot, periodogram, correlogram, etc.

At least line plot simple and universal tool!


240
220
200
180
160
140
120
100
80
60
40
1
51
101
151
201
251
301
351
401

Visual tools for categorical time series:

only few proposals,

often from computer science and biology,

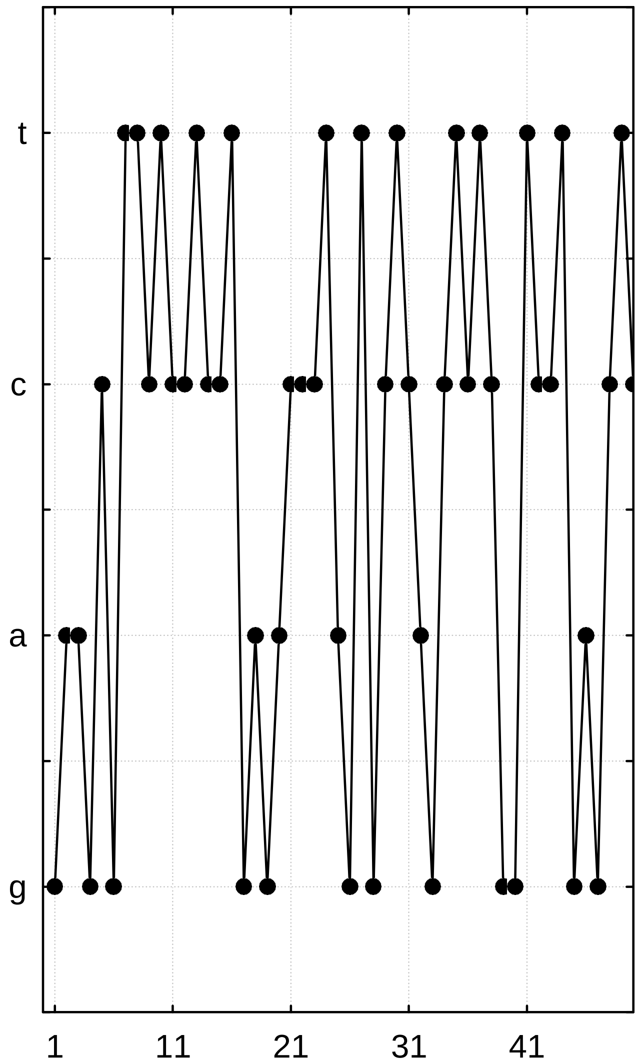
see survey by Weiß (2008).

In fact problematic: Analogue of line plot?

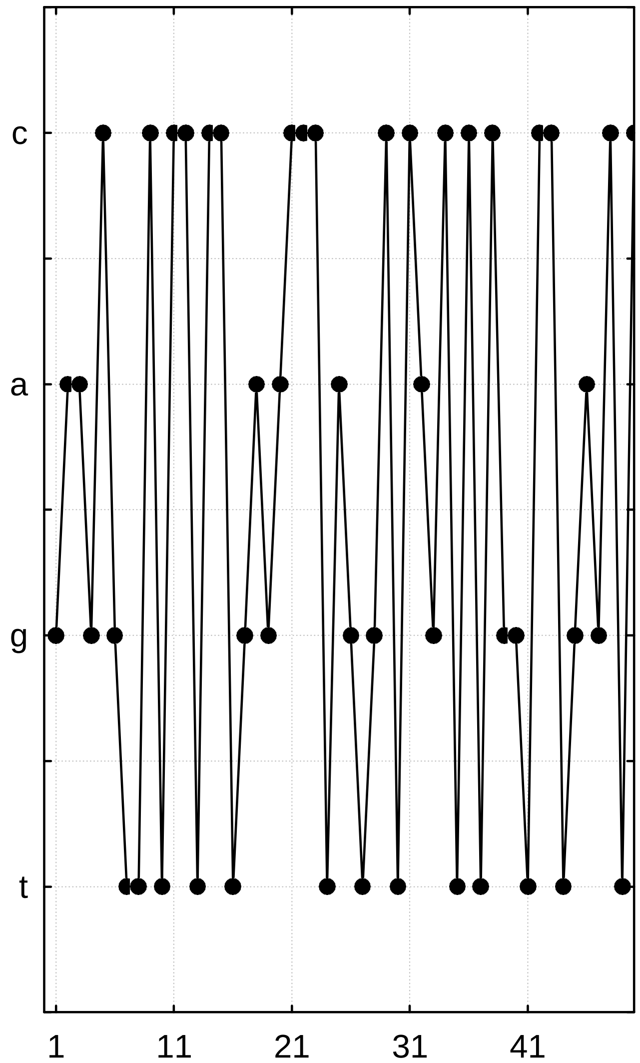
Lack of a natural order within purely categorical range

$\Rightarrow$ arrangement of range along ordinate

arbitrary and misleading.



or



Alternative (Keim & Kriegel, 1996): Map range onto set of colors or symbols, plot $x_1, x_2, \ldots$ successively on space-filling curve (line-by-line, column-by-column, Peano-Hilbert curves, spiral, etc.).

However: These graphs perform poorly, characteristic features of time series difficult to recognize, interpretation of resulting plot is problematic, appearance depends heavily on choice of underlying curve.

Genome of Bovine leukemia virus:

So how to analyze categorical time series visually?

Proposal by Ribler (1997): **Rate evolution graph**.

Categorical process $(X_t)_{\mathbb{N}}$, range coded as $\mathcal{V} = \{0, \ldots, m\}$.

Binarization $\boldsymbol{Y}_t \in \{0, 1\}^{m+1}$ with $Y_{t,i} = \delta_{i,X_t}$, $i = 0, \ldots, m$.

Define the cumulated sums $\boldsymbol{C}_t := \sum_{s=1}^{t} \boldsymbol{Y}_s$, i. e.,

$$C_{t,i} \ = \text{number of } X_s,\ s = 1, \ldots, t, \text{ equal to } i.$$

Rate evolution graph of $(X_t)_{\mathbb{N}}$: (Ribler, 1997)

Multiple line plot of all component series $C_{t,i}$, $i = 0, \ldots, m$, i. e., all $C_{t,i}$ are plotted simultaneously into one chart.

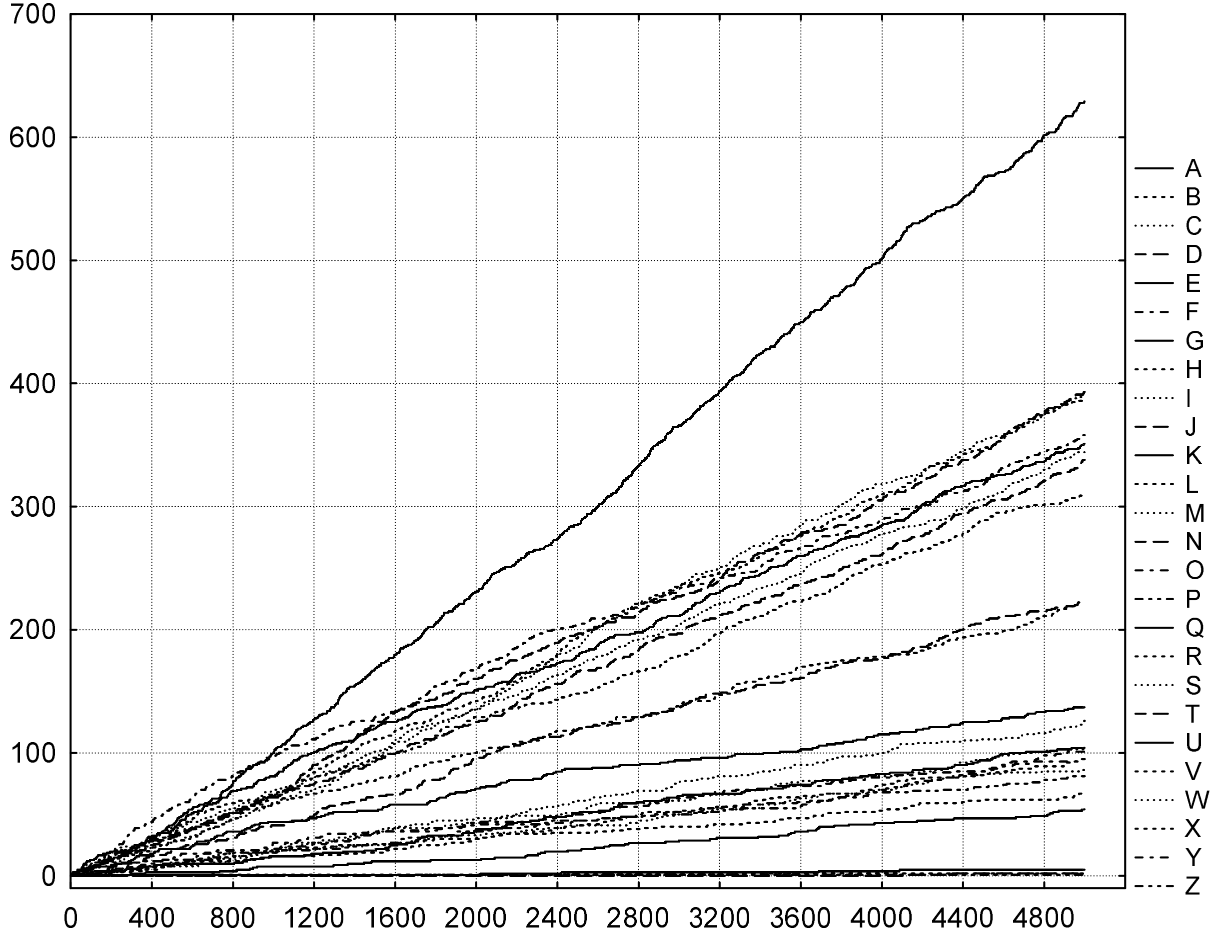
Interpretation:

Slope of graphs is estimate for corresp. marginal probability.

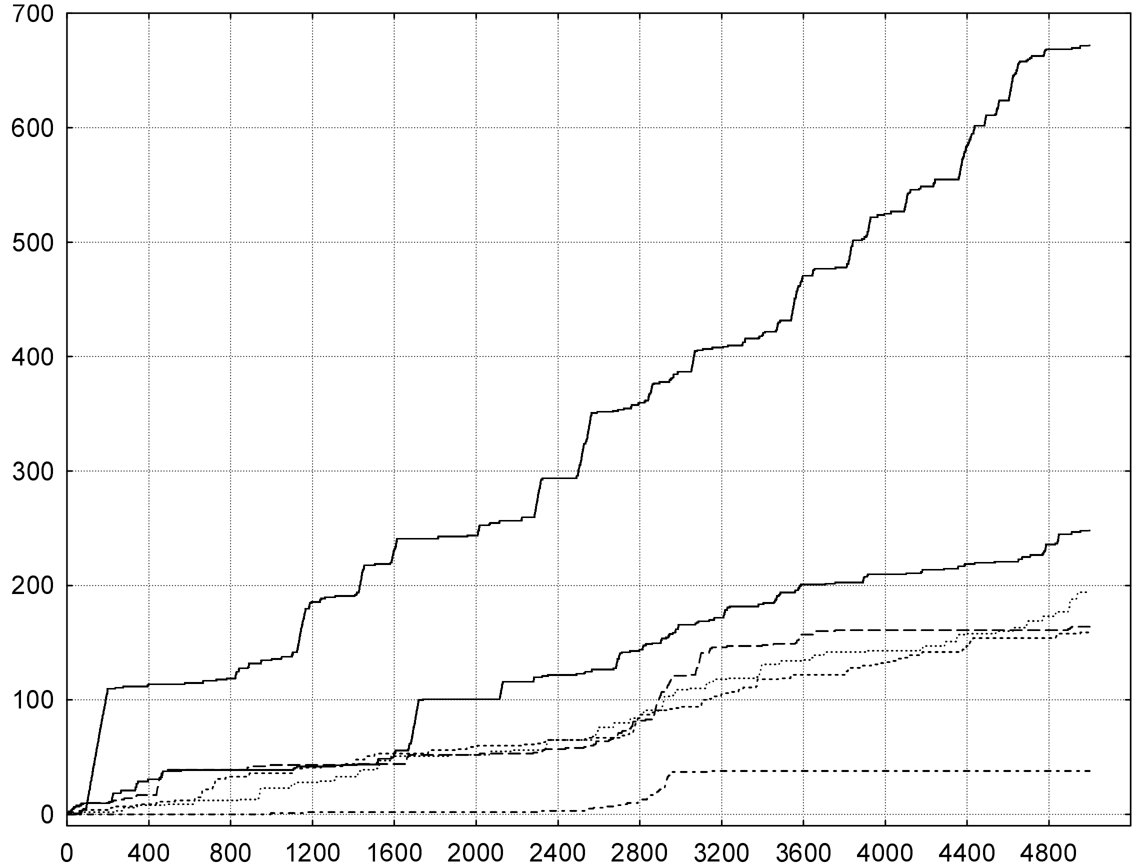
If $(X_t)_{\mathbb{N}}$ stationary and at most moderately serially dependent, then graphs approximately linear in t

$\Rightarrow$ Simple visual tool for checking stationarity.

Shakespeare's (1593) poem "Venus and Adonis":



Log data (2005) of Statistics server at Univ. of Würzburg: Access to home directory of five members.



Some further tools: (see Weiß (2008) for references)

- **VizTree** or **IFS circle transformation** for visualizing string frequencies; (→ ENBIS talk in 2005)
- **pattern histograms** (e. g., based on runs or cycles);
- **categorical control charts** (also see below).

However: These tools are very specialized.

Universal instrument (like line plot), providing multiple types of information at once, still missing.

Categorical Time Series

Serial Dependence

Cardinal time series:

Convenient measure of serial dependence:

(Partial) Autocorrelation.

Categorical time series:

Measures of serial dependence?

Weiß & Göb (2008): same strategy as for dispersion measures, i. e., we start with

Extreme cases:

- X_t, X_{t-k} *(stochastically) independent* iff $p_{ij}(k) = p_i \cdot p_j$.

- X_t *perfectly depends* on X_{t-k} iff for every $j = 0, \ldots, m$: conditional distribution of X_t, conditioned on $X_{t-k} = j$, is a one-point distribution.

Weiß & Göb (2008) proposed, among others,

Goodman and Kruskal's τ: $$\tau(k) = \sqrt{\sum_{i,j} \frac{\big(p_{ij}(k) - p_i p_j\big)^2}{p_j\big(1 - s_2(\boldsymbol{p})\big)}},$$

Cramer's v: $$v(k) = \sqrt{\frac{1}{m} \cdot \sum_{i,j} \frac{\big(p_{ij}(k) - p_i p_j\big)^2}{p_i p_j}}.$$

Some properties: $\tau(k), v(k)$ have range $[0;1]$ with

- X_t, X_{t-k} independent $\Leftrightarrow$ $\tau(k) = v(k) = 0$.
- X_t depends perfectly on X_{t-k} $\Leftrightarrow$ $\tau(k) = v(k) = 1$.

Cardinal case: Positive and negative autocorrelation.

$\Rightarrow$

Concept of **signed dependence:** (Weiß & Göb, 2008)

- X_t, X_{t-k} **perfectly positively dependent**,
  if perfectly dependent and
  if $p_{i|i}(k) = 1$ for all i.
- X_t, X_{t-k} **perfectly negatively dependent**
  if perfectly dependent and
  if $p_{i|i}(k) = 0$ for all i.

Measures of signed dependence:

(Weiß, 2011; Weiß & Göb, 2008)

Cohen's κ:

$$\kappa(k) \ = \ \frac{\sum_j p_{jj}(k) - s_2(\boldsymbol{p})}{1 - s_2(\boldsymbol{p})}$$ with range $[-\frac{s_2(\boldsymbol{p})}{1-s_2(\boldsymbol{p})}\ ;\ 1]$,

Modified κ:

$$\kappa^*(k) \ = \ \frac{1}{m} \cdot \Big(\sum_j \ p_{j|j}(k) \ - 1 \Big)$$ with range $[-\frac{1}{m}; 1]$.

Some properties:

- X_t, X_{t-k} independent $\quad \Rightarrow \quad \kappa(k) = \kappa^*(k) = 0$.
- X_t, X_{t-k} perf. positively dep. $\quad \Leftrightarrow \quad \kappa(k) = \kappa^*(k) = 1$.
- X_t, X_{t-k} perf. negatively dep. $\quad \Rightarrow \quad \kappa(k), \kappa^*(k)$ minimal.

Empirical measures of (signed) serial dependence

based on time series $X_1, \ldots, X_T$.

Weiß (2011): **empirical Cohen's κ and modified κ**:

$$\hat{\kappa}(k) \ := \ 1 + \tfrac{1}{T} \ - \ \frac{1 - \sum_j \hat{p}_{jj}(k,T)}{1 - s_2(\hat{\boldsymbol{p}}(T))},$$

$$\hat{\kappa}^*(k) \ := \ \tfrac{1}{T} \ + \ \tfrac{1}{m} \cdot \Big(\sum_j \ \frac{\hat{p}_{jj}(k,T)}{\hat{p}_j(T)} \ - 1 \Big).$$

Empirical measures of (signed) serial dependence

If computed from i.i.d. data, then

$\widehat{\kappa}(k)$, $\widehat{\kappa}^*(k)$ are asymptotically normally distributed with

$$E\big[\widehat{\kappa}(k)\big] \;=\; E\big[\widehat{\kappa}^*(k)\big] \;=\; 0 \;+\; O(T^{-2}),$$

and variances given by

$$V\big[\widehat{\kappa}(k)\big] \;=\; \tfrac{1}{T}\cdot\Big(1 \;-\; \tfrac{1+2s_3(\boldsymbol{p})-3s_2(\boldsymbol{p})}{(1-s_2(\boldsymbol{p}))^2}\Big) \;+\; O(T^{-2});$$

$$V\big[\widehat{\kappa}^*(k)\big] \;=\; \tfrac{1}{m\cdot T} \;+\; O(T^{-2}).$$

Current and future research:

- $\widehat{\kappa}(k)$, $\widehat{\kappa}^*(k)$ and also $\widehat{\tau}(k)$, $\widehat{\upsilon}(k)$ for NDARMA processes;

- goodness-of-fit tests for NDARMA processes.

Categorical Time Series

Modelling

Models for stationary categorical processes

with range $\mathcal{V} = \{0, \ldots, m\}$:

- **p^th order Markov model**: $(m+1)^{\mathsf{p}} \cdot m$ parameters;

- **variable length M. m.** of Bühlmann & Wyner (1999): more parsimonious, but model choice difficult;

- **MTD(p) model** of Raftery (1985): still $m(m+1) + \mathsf{p} - 1$ parameters;

- **NDARMA(p, q) models** of Jacobs & Lewis (1983): $m + \mathsf{p} + \mathsf{q}$ parameters, also non-Markovian dependence.

$(X_t)_{\mathbb{Z}}$, $(\epsilon_t)_{\mathbb{Z}}$: categorical processes with range $\mathcal{V} = \{0, \ldots, m\}$;
$(\epsilon_t)_{\mathbb{Z}}$: i.i.d. with marginal distribution $P(\epsilon_t = j) = p_j > 0$,
ϵ_t independent of $(X_s)_{s<t}$.

For $\varphi_{\mathsf{q}} > 0$, with $\phi_{\mathsf{p}} > 0$ if $\mathsf{p} \geq 1$, let

$$\boldsymbol{D}_t = (\alpha_{t,1}, \ldots, \alpha_{t,\mathsf{p}}, \beta_{t,0}, \ldots, \beta_{t,\mathsf{q}}) \sim MULT(1; \phi_1, \ldots, \phi_{\mathsf{p}}, \varphi_0, \ldots, \varphi_{\mathsf{q}})$$

be i.i.d. and independent of $(\epsilon_t)_{\mathbb{Z}}$, $(X_s)_{s<t}$.

$(X_t)_{\mathbb{Z}}$ is **NDARMA(p, q) process** if

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

Some properties:

- marginal distribution $P(X_t = j) = p_j$;
- only shows positive serial dependence, and

$$\kappa(k) \;=\; \kappa^*(k) \;=\; v(k) \;=\; \tau(k);$$

- Yule-Walker-type equations

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

where $r(i) = 0$ for $i < 0$, $r(0) = \varphi_0$, and

$$r(i) \;=\; \textstyle\sum_{j=\max\{0,i-\mathsf{p}\}}^{i-1} \phi_{i-j} \cdot r(j) + \sum_{j=0}^{\mathsf{q}} \delta_{ij} \cdot \varphi_j \qquad \text{for } i > 0.$$

Disadvantage of NDARMA models:

Only describe positive dependence, i. e., categorical time series with long runs of symbols. (e. g., genetic sequence vs. letters of text)

Issues for future research:

- Find simple and sparsely parametrized models that also allow for negative dependence!
- Definition of trend or seasonality?
- And: How to remove such trend or seasonality?

Categorical Processes

Monitoring

Only very few approaches for monitoring processes of **unordered** and **mutually exclusive** categories.

Monitoring **samples** from an **i.i.d.** categorical process: Duncan (1950), Marcucci (1985), Nelson (1987), Mukhopadhyay (2008) plot Pearson's χ^2-statistic for goodness of fit on a control chart.

Continuously monitoring cat. process (100 % inspect.): Weiß (2010) proposes

- two moving-average-type charts, based either on Pearson statistic or Gini index,
- a (k, r)-runs chart.

(k, r)-**run:** finished after k successive observations of either '1' or '2' or ... or 'r', i. e., after observing one of $(1, \ldots, 1), (2, \ldots, 2), \ldots, (r, \ldots, r)$ of length k each.

(k, r)**th run lengths** $(Y_n^{(k,r)})_{\mathbb{N}}$ determined as

$Y_1^{(k,r)}$:= No. obs. until first occurrence of k-tuple of '1's or ... 'r's,

$Y_n^{(k,r)}$:= No. obs. after $(n-1)^{\text{th}}$ occurrence of k-tuple of '1's or ... 'r's until n^{th} occurrence of k-tuple of '1's or ... 'r's, for $n \geq 2$.

Example: $m = 3$ (i. e., $\mathcal{V} = \{0, 1, 2, 3\}$)

and $(k, r) = (2, 3)$, (fictive) time series:

$$\underbrace{1\ 2\ 0\ 0\ 0\ 3\ 0\ \mathbf{2}\ \mathbf{2}}_{9}\ \underbrace{0\ 3\ 0\ 0\ 1\ 0\ \mathbf{1}\ \mathbf{1}}_{8}\ \underbrace{\mathbf{1}\ \mathbf{1}}_{2}\ \underbrace{3\ 0\ 2\ 0\ \mathbf{3}\ \mathbf{3}}_{6}\ 2\ 3\ \ldots$$

If $(X_t)_{\mathbb{N}}$ Markov chain, then

$(Y_n^{(k,r)})_{\mathbb{N}}$ i.i.d. process, range $\mathbb{N}_k := \{k, k+1, \ldots\}$.

Properties: (Chryssaphinou et al., 1994)

Let $c_{k,r}(z) \;:=\; \sum_{i=1}^{r} \frac{(1-p_i z)(p_i z)^k}{1-(p_i z)^k}$, then

$$E[Y^{(k,r)}] \;=\; \frac{1}{c_{k,r}(1)}, \qquad V[Y^{(k,r)}] \;=\; \frac{1+c_{k,r}(1)-2c'_{k,r}(1)}{c^2_{k,r}(1)}.$$

(k, r)-**runs chart:** (Weiß, 2010)

$(Y_n^{(k,r)})_{\mathbb{N}}$ plotted on chart with $k \leq LCL < UCL$.

Exact ANE **computation** with Markov chain approach.

Issues for future research:

- charts based on different patterns,
  e. g., cycles instead of runs;

- CUSUM or EWMA methods for categorical processes,
  e. g., related to patterns or certain statistics; ...

Conclusions

In a nutshell,

the field of categorical time series . . .

- is relevant for practice, and

- offers a lot of topics for future research,

  in any of the disciplines

analysis, modelling and monitoring!

Thank You

for Your Interest!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich
Mathematik

Christian H. Weiß

Department of Mathematics

Darmstadt University of Technology

Literature

Box et al. (1994): *Time series analysis – forecasting and control*. 3rd edition, Prentice Hall, Inc., New Jersey.

Bühlmann & Wyner (1999): Variable length Markov chains. *Ann. Stat.*, **27**, 480–513.

Chryssaphinou et al. (1994): On the waiting time of appearance of given patterns. In: Godbole & Papastavridis (Eds.): *Runs and Patterns in Probability*, Kluwer Academic Publishers, 231–241.

Duncan (1950): A chi-square chart for controlling a set of percentages. *Industrial Quality Control*, **7**, 11–15.

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

Keim & Kriegel (1996): Visualization techniques for mining large databases: A comparison. *IEEE Trans. Knowl. Data Eng.*, **8**(6), 923–938.

Marcucci (1985): Monitoring multinomial processes. *J. Qual. Tech.*, **17**(2), 86–91.

Mukhopadhyay (2008): Multivariate attribute control chart using Mahalanobis D^2 statistic. *J. Appl. Statist.*, **35**(4), 421–429.

Nelson (1987): A chi-square control chart for several proportions. *J. Qual. Tech.*, **19**(4), 229–231.

Raftery (1985): A model for high-order Markov chains. *J. Royal Stat. Soc.*, B, **47**(3), 528–539.

(. . .)

Literature

(. . .)

Ribler (1997): *Visualizing categorical time series data with applications to computer and communications network traces*. PhD thesis, Virginia State University.

Uschner (1987): *Streuungsmessung nominaler Merkmale mit Hilfe von Paarvergleichen*. Doctoral dissertation, University Erlangen-Nürnberg.

Vogel & Kiesl (1999): Deskriptive und induktive Eigenschaften zweier Streuungsmaße für nominale Merkmale. In: Vogel (Edt.): *Arbeiten aus der Statistik*, Univ. Bamberg.

Weiß (2008): Visual analysis of categorical time series. *Statist. Meth.*, **5**(1), 56–71.

Weiß (2010): Continuously monitoring categorical processes. *Qual. Tech. Quant. Manag.*, to appear.

Weiß (2011): Empirical measures of signed serial dependence in categorical time series. *J. Statist. Comp. Simulation*, **81**(4), 411–429.

Weiß & Göb (2008): Measuring serial dependence in categorical time series. *Adv. Statist. Anal.*, **92**(1), 71–89.

Ye et al. (2002): Multivariate statistical analysis of audit trails for host-based intrusion detection. *IEEE Trans. Computer*, **51**(7), 810–820.