Control Charts for Time-Dependent Categorical Processes

Christian H. Weiß

Department of Mathematics & Statistics,

Helmut Schmidt University, Hamburg

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH
STAT

Monitoring of Categorical Processes

Introduction

Particular type of attributes data processes $(X_t)_{\mathbb{N}}$:

range of X_t of **categorical** nature, i. e., :

X_t has discrete and non-metric range (**state space**)

consisting of finite number $m+1$ of categories with $m \in \mathbb{N}$.

Sometimes range exhibits natural ordering, then **ordinal** range.

Otherwise, without inherent order, **nominal** range.

Here, we assume that

X_t takes one of finite number of **unordered** categories.

To simplify notations:

range coded as $\mathcal{S} = \{0, \ldots, m\}$, just lexicographic order.

Quality-related applications:

X_t describes result of inspection of item,

leads to classification $X_t = i$ for an $i = 1, \ldots, m$ iff

t^{th} item was non-conforming of type 'i',

or $X_t = 0$ for conforming item.

Examples:

- Mukhopadhyay (2008): non-conforming ceiling fan cover according to most predominant paint defect, e. g., 'poor covering' or 'bubbles'.
- Ye et al. (2002): monitoring of network traffic data with different types of audit events.

During last few years, increasing research interest
in monitoring of categorical processes,
e. g., Chen et al. (2011), Ryan et al. (2011), Weiß (2012).

Restriction with these works:
underlying process assumed **serially independent**
in its in-control state, i. e., $X_1, X_2, \ldots$ i. i. d.

Researchers and practitioners often ill at ease when being
concerned with **time-dependent** categorical data:
concepts for categ. serial dependence not well communicated,
simple (ARMA-like) models not known to broader audience.

Outline:

- Brief survey of approaches for modeling and analyzing categorical processes.

  Then **two strategies** for monitoring categorical process:

- If process evolves too fast to be monitored continuously, take segments from process at selected times, plot sample statistic on control chart.
  Here, carefully consider serial dependence **within** sample.
- If possible to continuously monitor the process, then serial dependence taken into account **between** plotted statistics.

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

MATH STAT

Modeling and Analyzing Categorical Processes

Survey

- Stationary **real-valued** time series:

  huge toolbox for analyzing and modeling such time series readily available and well-known to broad audience.

  E. g., time series

  visualized by simply plotting observations against time,

  marginal location/dispersion by mean/variance,

  serial dependence quantified in terms of autocorrelation.

  Enumerable models, basic ARMA or extensions.

- **Categorical but ordinal** time series:
  time series plot still feasible by arranging
  possible outcomes in natural ordering along Y axis,
  location could be measured by median.
- **Nominal** time series: tailor-made solutions required.

Notations concerning $(X_t)_{\mathbb{Z}}$ with range $\mathcal{S} = \{0, \ldots, m\}$, $m > 1$:

time-invariant marginal probabilities $\boldsymbol{\pi} := (\pi_0, \ldots, \pi_m)^\top$
with $\pi_i := P(X_t = i) \in (0; 1)$ and $\pi_0 = 1 - \pi_1 - \ldots - \pi_m$.
Sample counterpart: $\hat{\boldsymbol{\pi}}$ with relative frequencies from $X_1, \ldots, X_T$.

- Visual analysis: few proposals (Weiß, 2008), reasonable substitute of time series plot still missing.
- Location: (empirical) mode.
- Dispersion: two possible extremes, one-point distribution (no dispersion) and uniform distribution (maximal dispersion). Several measures available (Weiß & Göb, 2008), recommendation: (empirical) **Gini index**,

$$\nu_G = \frac{m+1}{m}\left(1-\sum_{j=0}^{m}\pi_j^2\right) \quad \text{and} \quad \hat{\nu}_G = \frac{m+1}{m}\,\frac{T}{T-1}\left(1-\sum_{j=0}^{m}\hat{\pi}_j^2\right).$$

(Empirical) **Gini index**,

$$\nu_{\mathsf{G}} \;=\; \tfrac{m+1}{m}\,\big(1-\textstyle\sum_{j=0}^{m}\pi_j^2\big) \quad \text{and} \quad \hat{\nu}_{\mathsf{G}} \;=\; \tfrac{m+1}{m}\,\tfrac{T}{T-1}\,\big(1-\textstyle\sum_{j=0}^{m}\hat{\pi}_j^2\big).$$

Theoretical Gini index ν_{G} has range $[0;1]$,
where increasing values indicate increasing dispersion,
with extremes $\nu_{\mathsf{G}}=0$ iff X_t has one-point distribution,
and $\nu_{\mathsf{G}}=1$ iff X_t has uniform distribution.

Empirical Gini index $\hat{\nu}_{\mathsf{G}}$ unbiased in i. i. d. case (Weiß, 2013).

- Serial dependence:

  several measures available (Weiß & Göb, 2008; Weiß, 2013), relying on lagged

  bivariate probabilites, $p_{ij}(k) := P(X_t = i, X_{t-k} = j)$,

  with empirical counterpart $\widehat{p}_{ij}(k)$ being relative

  frequency of (i,j) within pairs $(X_{k+1}, X_1), \ldots, (X_T, X_{T-k})$.

  Recommendation: (empirical) **Cohen's** κ,

$$\kappa(k) = \frac{\sum_{j=0}^{m} \left(p_{jj}(k) - \pi_j^2\right)}{1 - \sum_{j=0}^{m} \pi_j^2}, \qquad \widehat{\kappa}(k) := \frac{1}{T} + \frac{\sum_{j=0}^{m} \left(\widehat{p}_{jj}(k) - \widehat{\pi}_j^2\right)}{1 - \sum_{j=0}^{m} \widehat{\pi}_j^2}.$$

(Empirical) **Cohen's** κ,

$$\kappa(k) \;=\; \frac{\sum_{j=0}^{m}\big(p_{jj}(k)-\pi_j^2\big)}{1-\sum_{j=0}^{m}\pi_j^2}, \qquad \hat{\kappa}(k) \;:=\; \frac{1}{T}+\frac{\sum_{j=0}^{m}\big(\hat{p}_{jj}(k)-\hat{\pi}_j^2\big)}{1-\sum_{j=0}^{m}\hat{\pi}_j^2}.$$

Theoretical $\kappa(k)$ has range $[-\frac{\sum_{j=0}^{m}\pi_j^2}{1-\sum_{j=0}^{m}\pi_j^2}\ ;\ 1]$,
where 0 corresponds to serial independence.

Signed serial dependence: (Weiß & Göb, 2008)

- perfect (**unsigned**) serial dependence at lag $k \in \mathbb{N}$ iff for any j, $p_{\cdot|j}(k)$ one-point distribution,
- perfect **positive** (**negative**) dependence iff all $p_{i|i}(k) = 1$ (all $p_{i|i}(k) = 0$).

(Empirical) **Cohen's** κ,

$$\kappa(k) \;=\; \frac{\sum_{j=0}^{m}\big(p_{jj}(k)-\pi_j^2\big)}{1-\sum_{j=0}^{m}\pi_j^2}, \qquad \hat{\kappa}(k) \;:=\; \frac{1}{T} + \frac{\sum_{j=0}^{m}\big(\hat{p}_{jj}(k)-\hat{\pi}_j^2\big)}{1-\sum_{j=0}^{m}\hat{\pi}_j^2}.$$

Empirical $\hat{\kappa}(k)$ nearly unbiased in i. i. d. case,

distribution well approximated by normal distribution $\mathrm{N}(0,\sigma^2)$

with (Weiß, 2011)

$$T\,\sigma^2 \;=\; 1-\frac{1+2\,\sum_{j=0}^{m}\pi_j^3-3\,\sum_{j=0}^{m}\pi_j^2}{\big(1-\sum_{j=0}^{m}\pi_j^2\big)^2}.$$

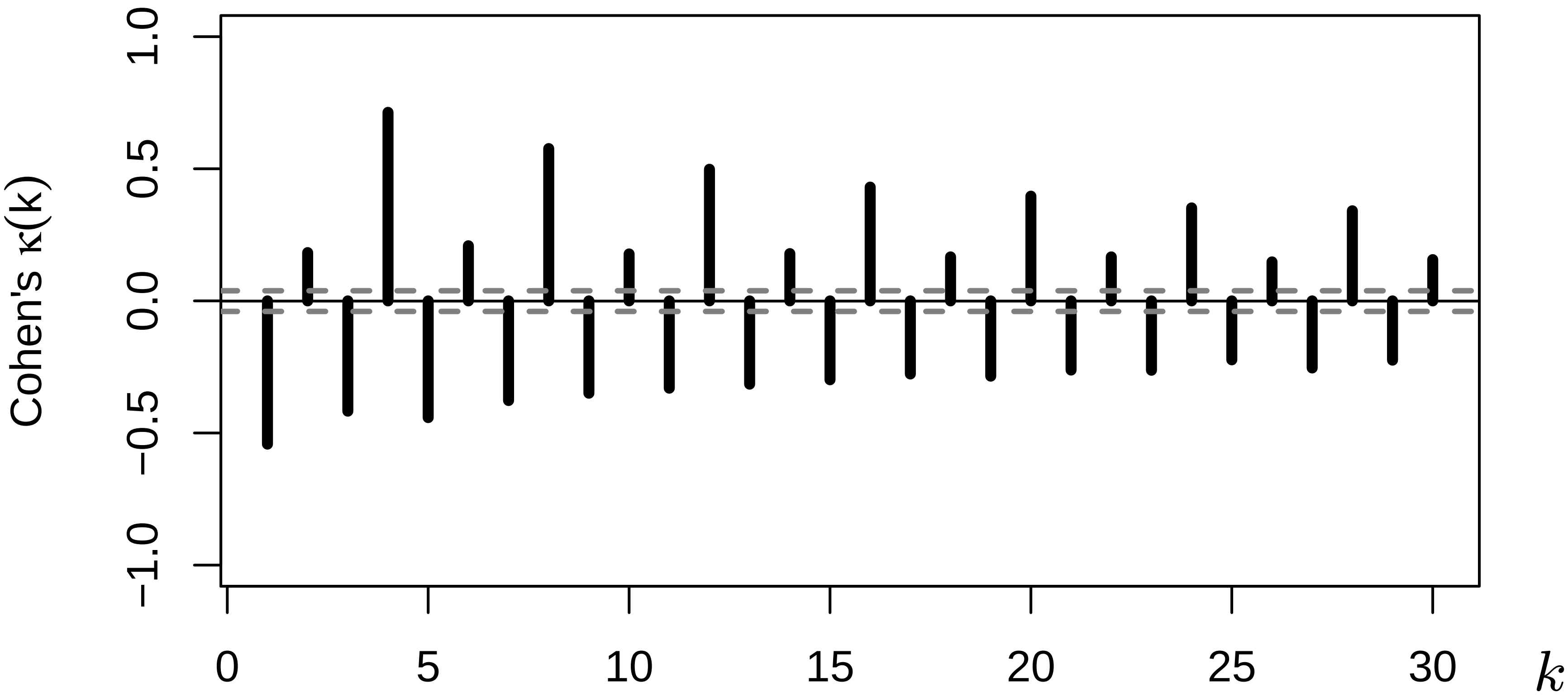
$\Rightarrow$ testing for significant serial dependence in categorical t. s.,

serial dependence plot.

(Non-industrial) Example for **serial dependence plot**:

morning twilight song of Wood Pewee,

composed of three different phrases, length $T = 1327$:



Several **models for categorical processes**, e. g.,

(Hidden) Markov models, regression models, . . . , **here:**

NDARMA(p, q) model by Jacobs & Lewis (1983).

$(X_t)_{\mathbb{Z}}$, $(\epsilon_t)_{\mathbb{Z}}$: categorical processes with state space $\mathcal{S}$;

$(\epsilon_t)_{\mathbb{Z}}$: i. i. d. with marginal $\boldsymbol{\pi}$, 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

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

Properties of NDARMA processes: (Weiß & Göb, 2008)

- Marginal distribution $P(X_t = j) = \pi_j$;
- $\kappa(k) \geq 0$, equality $\kappa(k) = v(k)$;
- **Yule-Walker-type equations**

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

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

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

$\Rightarrow$ Model identification as in ARMA case!

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Sample-based Monitoring of Categorical Processes

Some Results

Take (non-overlapping) segments of length $n > 1$

at times $t_1, t_2, \ldots$ with $t_k - t_{k-1} > n$ sufficiently large,

i. e., samples $X_{t_k}, \ldots, X_{t_k+n-1}$.

Proceedings paper: detailed survey about

- Sample-based monitoring for **binary case**,
- and for **categorical but i. i. d. case**.
- Brief discussion about approaches for **ordinal data** and **compositional data**.

Let $\boldsymbol{N}_k{}^{(n)} = (N_{k;\,0}{}^{(n)}, \ldots, N_{k;\,m}{}^{(n)})^\top$ with $N_{k;\,i}{}^{(n)}$ being the absolute frequency of state 'i' in sample $X_{t_k}, \ldots, X_{t_k+n-1}$ such that $N_{k;\,0}{}^{(n)} + \ldots + N_{k;\,m}{}^{(n)} = n$.

- If $(X_t)_{\mathbb{N}}$ i. i. d., then $\boldsymbol{N}_k \underset{i.\,i.\,d.}{\sim} \mathrm{MULT}(n;\ \pi_0, \ldots, \pi_m)$ with covariance matrix $n\,\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = (\sigma_{ij}) \quad \text{is given by } \sigma_{ij} \;=\; \begin{cases} \pi_i(1-\pi_i) & \text{if } i = j, \\ -\pi_i\pi_j & \text{if } i \neq j. \end{cases}$$

Let $\boldsymbol{N}_k{}^{(n)} = (N_{k;\,0}{}^{(n)}, \ldots, N_{k;\,m}{}^{(n)})^\top$ with $N_{k;\,i}{}^{(n)}$ being the absolute frequency of state 'i' in sample $X_{t_k}, \ldots, X_{t_k+n-1}$ such that $N_{k;\,0}{}^{(n)} + \ldots + N_{k;\,m}{}^{(n)} = n$.

- If $(X_t)_{\mathbb{N}}$ DAR(1) process with autoregressive parameter ρ, then $\kappa(k) = \rho^k$ and

  $\boldsymbol{N}_k \underset{i.\,i.\,d.}{\sim}$ MM($n;\ \pi_0, \ldots, \pi_m;\ \rho$) (Wang & Yang, 1995)

  with asymptotic covariance matrix $c \cdot n\,\boldsymbol{\Sigma}$, where

$$c \ := \ 1 + 2 \cdot \sum_{i=1}^{\infty} \kappa(i) \ = \ \frac{1+\rho}{1-\rho}.$$

$\Rightarrow$ effect of serial dependence **within** sample.

Sample statistics to be monitored:

- **Pearson's χ^2-statistic** (Duncan, 1950)

$$C_k^{(n)} \;=\; \sum_{j=0}^{m} \frac{(N_{k;\,j} - n\,\pi_{0;\,j})^2}{n\,\pi_{0;\,j}},$$

where $\boldsymbol{\pi}_0 := (\pi_{0;\,0}, \ldots, \pi_{0;\,m})^\top$ in-control categ. probabilities.

- **Gini statistic** (Weiß, 2012)

$$G_k^{(n)} \;=\; \frac{1 - n^{-2}\,\sum_{j=0}^{m} N_{k;\,j}^2}{1 - \sum_{j=0}^{m} \pi_{0;\,j}^2},$$

motivated by conforming probability

$\pi_{0;\,0} \gg \pi_{0;\,1}, \ldots, \pi_{0;\,m}$ (i. e., low categorical dispersion).

Proceedings paper: simulation study for diverse scenarios, asymptotic distributions for $C_k^{(n)}, G_k^{(n)}$ available for arbitrary NDARMA processes (Weiß, 2013), but too imprecise for chart design.

Illustrative example: (Mukhopadhyay, 2008)

$\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$

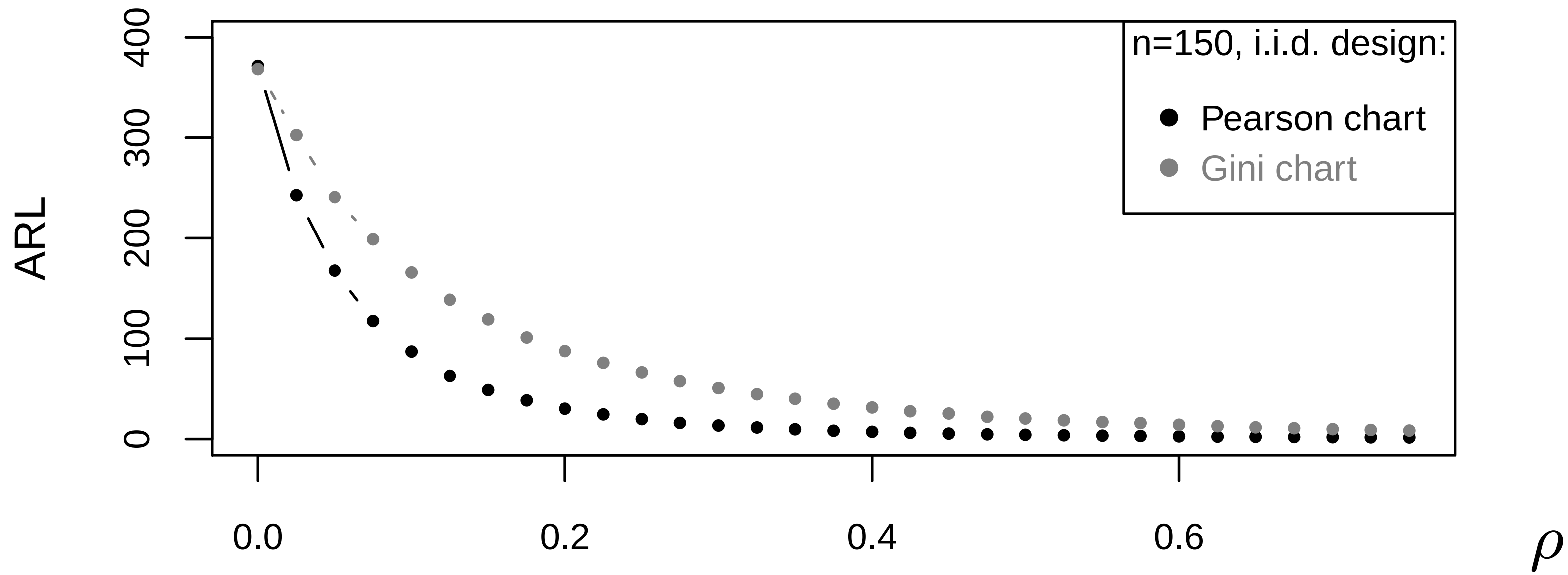
with Gini dispersion 0.463, sample size $n = 150$.

DAR(1) dependence, where $\rho = 0$ expresses i. i. d. case.

$n = 150$, $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$:

In-control ARL performance if **i. i. d. design**, i. e.,

Pearson with $u_P = 22.41094$, Gini with $u_G = 1.327252$.

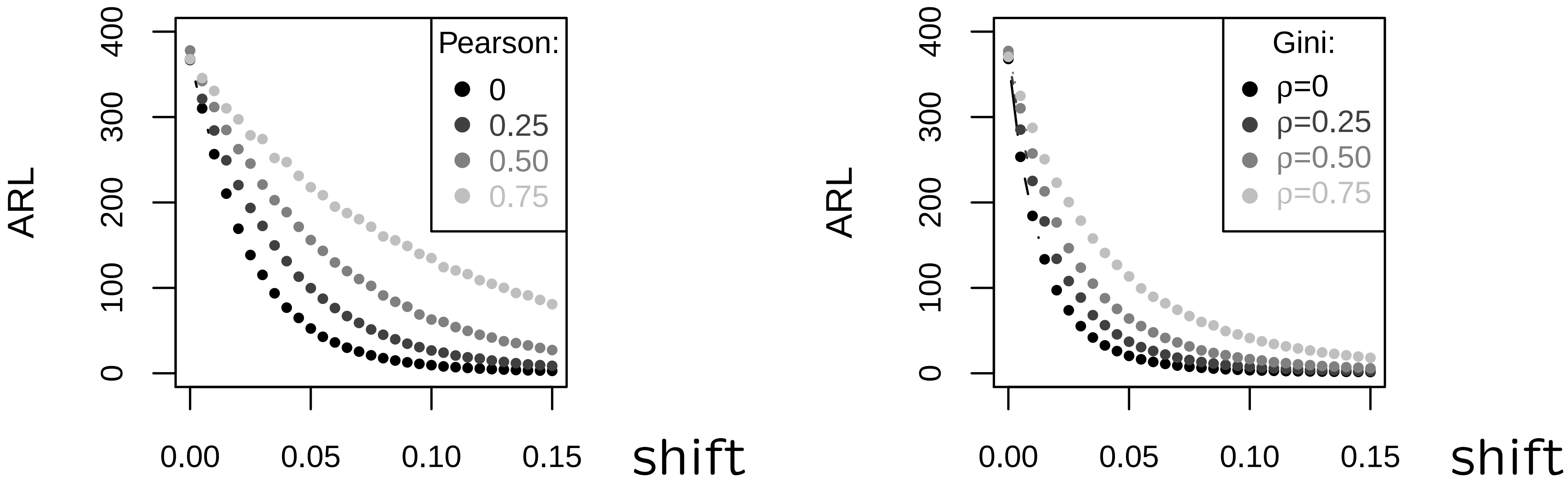


$\Rightarrow$ Strong influence of serial dependence on chart design.

$n = 150$, $\boldsymbol{\pi}_0 = (0.769, 0.081, 0.059, 0.022, 0.023, 0.022, 0.025)^\top$:

Out-of-control ARL performance

for $\pi_{1;0} = (1 - \text{shift})\,\pi_{0;0}$, $\pi_{1;k} = \frac{1-\text{shift}\cdot\pi_{0;0}}{1-\pi_{0;0}}\,\pi_{0;k}$ otherwise.



$\Rightarrow$ Decaying power with increasing serial dependence.

Continuous Monitoring of Categorical Processes

Some Results

Log-likelihood ratio CUSUM constitutes feasible approach.

Ryan et al. (2011): **i. i. d. case**, where $\boldsymbol{\pi}_1$ denotes relevant out-of-control distribution. Then

$$S_t \;=\; \max\{0,\, S_{t-1} + L_t\} \quad \text{with } S_0 := 0,$$

where $L_t = \ln\big(P_{\boldsymbol{\pi}_1}(X_t)/P_{\boldsymbol{\pi}_0}(X_t)\big)$ with $P_{\boldsymbol{\pi}}(X_t = i) = \pi_i$.

Following Mousavi & Reynolds (2009) ($\rightarrow$ binary Markov chain), one may define

$$L_t \;=\; \ln\left(\frac{P_{\boldsymbol{\pi}_1}(X_t|X_{t-1},\ldots,X_{t-\mathrm{p}})}{P_{\boldsymbol{\pi}_0}(X_t|X_{t-1},\ldots,X_{t-\mathrm{p}})}\right)$$

for Markov-dependent categorical process.

Illustrative example:

DAR(1) process $(X_t)_{\mathbb{N}}$ with autoregressive parameter ρ.

Approaches for $S_t \;=\; \max\{0,\, S_{t-1} + L_t\}$:

- i. i. d.-CUSUM statistic

$$L_t \;=\; \ln \frac{P_{\pi_1}(X_t)}{P_{\pi_0}(X_t)} \;=\; \ln \frac{\pi_{1;\,X_t}}{\pi_{0;\,X_t}},$$

  but with adjusted limits; or

- adjusted CUSUM statistic:

$$L_t \;=\; \ln \frac{(1-\rho)\,\pi_{1;\,X_t} \;+\; \delta_{X_t,X_{t-1}}\,\rho}{(1-\rho)\,\pi_{0;\,X_t} \;+\; \delta_{X_t,X_{t-1}}\,\rho}.$$

Proceedings paper: scenarios from Ryan et al. (2011), i. e.,

Case 1: $\pi_0 = (0.65, 0.25, 0.10)^\top$, $\pi_1 = (0.4517, 0.2999, 0.2484)^\top$;
Case 2: $\pi_0 = (0.94, 0.05, 0.01)^\top$, $\pi_1 = (0.8495, 0.0992, 0.0513)^\top$;
Case 3: $\pi_0 = (0.994, 0.005, 0.001)^\top$, $\pi_1 = (0.9848, 0.0099, 0.0053)^\top$;
Case 4: $\pi_0 = (0.65, 0.20, 0.10, 0.05)^\top$, $\pi_1 = (0.3960, 0.3283, 0.1734, 0.1023)^\top$.

with dispersion $\nu_G \approx 0.758, 0.171, 0.018$ and 0.7.

Illustration:

		i. i. d.-CUSUM			i. i. d.-CUSUM, adj.			adj. CUSUM		
Case	ρ	h	ARL_0	ARL_1	h	ARL_0	ARL_1	h	ARL_0	ARL_1
2	0	2.8	501.8	36.3						
	0.25	2.8	245.7	37.2	3.85	509.8	52.4	2.55	503.4	45.6
	0.5	2.8	170.8	39.3	5.2	500.2	72.6	2.25	508.4	58.8
	0.75	2.8	155.2	48.3	7.6	500.7	107.8	1.7	514.7	86.0

$\Rightarrow$ adjusted CUSUM shows best power.

Conclusions and Future Research

- Monitoring of serially dependent categorical processes: Shewhart charts for sample-based monit. (Pearson, Gini), LR-CUSUM for continuous monitoring; simulations required for chart design and performance evaluation.

Future research:

- Sample-based CUSUM $L_k = \ln\left(P_{\pi_1}(\boldsymbol{N}_k{}^{(n)})/P_{\pi_0}(\boldsymbol{N}_k{}^{(n)})\right)$, but distribution of $\boldsymbol{N}_k{}^{(n)}$ difficult.
- Unique charts for categorical & compositional data.
- Phase I application of categorical control charts.
- Process capability indices for categorical data.

Thank You

for Your Interest!

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

weissc@hsu-hh.de

Literature

Chen et al. (2011) The application of multinomial . . . Int. J. Indust. Eng. 18, 244–253.

Duncan (1950) A chi-square chart for . . . Indust. Qual. Control 7, 11–15.

Jacobs & Lewis (1983) Stationary discrete autoregressive . . . J. Time Ser. Anal. 4, 19–36.

Mousavi & Reynolds (2009) A CUSUM chart for monitoring . . . JQT 41, 401–414.

Mukhopadhyay (2008) Multivariate attribute control . . . J. Appl. Statist. 35, 421–429.

Ryan et al. (2011) Methods for monitoring multiple proportions . . . JQT 43, 237–248.

Wang & Yang (1995) On a Markov multinomial distribution. Math. Scien. 20, 40–49.

Weiß (2008) Visual analysis of categorical time series. Stat. Meth. 5, 56–71.

Weiß (2011) Empirical measures of . . . J. Stat. Comp. Simul. 81, 411–429.

Weiß (2012) Continuously monitoring categorical processes. QTQM 9, 171–188.

Weiß (2013) Serial dependence of NDARMA . . . Comp. Stat. Data Anal. 68, 213–238.

Weiß & Göb (2008) Measuring serial dependence . . . Adv. Stat. Anal. 92, 71–89.

Ye et al. (2002) Multivariate statistical analysis . . . IEEE Trans. Computers 51, 810–820.