Serial Dependence in Categorical Time Series

MATH
STAT

Christian H. Weiß

Department of Mathematics & Statistics,
Helmut Schmidt University, Hamburg

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Categorical Time Series and Categorical Processes

Basic concepts

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 $(X_t)_{\mathbb{N}}$.

To simplify notations:

Range of $(X_t)_{\mathbb{N}}$ is coded as $\mathcal{V} = \{0, 1, \ldots, m\}$,

i. e., $P(X_t = 0) = 1 - \sum_{j=1}^{m} P(X_t = j)$.

Examples:

- Network traffic data;
- medical diagnoses of an examiner;
- text analysis, part-of-speech tagging;
- genetic or protein sequences;
- quality of produced items; . . .

Notations for time-invariant probabilities:

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

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

Models for stationary categorical processes

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

- **p$^{\text{th}}$ order Markov model**: $(m+1)^{\mathrm{p}} \cdot m$ parameters;
- **variable length M. m.** of Bühlmann & Wyner (1999): more parsimonious, but model choice difficult;
- **MTD(p) model** of Raftery (1985): still $m(m+1) + \mathrm{p} - 1$ parameters;
- **NDARMA(p, q) models** of Jacobs & Lewis (1983): $m + \mathrm{p} + \mathrm{q}$ 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 $\boldsymbol{p}$, 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 of NDARMA processes:

- marginal distribution $P(X_t = j) = p_j$;
- concerning the four measures of serial dependence discussed below, $\kappa(k)$, $\kappa^*(k)$, $\upsilon(k)$, $\tau(k)$, we have equality,

$$\kappa(k) \;=\; \kappa^*(k) \;=\; \upsilon(k) \;=\; \tau(k),$$

  and these measures satisfy **Yule-Walker-type equations**, i. e., **ARMA-like serial dependence structure**;

  (Weiß & Göb, 2008)

- ψ-mixing with exponent. decreasing weights (Weiß, 2013).

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Categorical Time Series Analysis

Basic concepts

Let $(X_t)_{\mathbb{N}}$ be stationary.

Measures of location:

only **mode** of X_t 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).

Measures of dispersion:

dispersion $\approx$ quantity of uncertainty, two extremes:

maximal dispersion if all p_j equal (**uniform distribution**),

minimal disp. if $p_j = 1$ for one $j \in \mathcal{V}$ (**one-point distrib.**).

Common standardized measures of dispersion:

Gini index: $\nu_{\mathrm{G}}(X) := \frac{m+1}{m}\left(1-\sum_j p_j^2\right).$

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

Properties:

- continuous and symmetric function of $p_0,\ldots,p_m$,
- range $[0;1]$,

  1 iff uniform distribution, 0 iff one-point distribution.

Alternative measures of dispersion: Weiß & Göb (2008).

Weiß & Göb (2008): **(signed) serial dependence**.

Stationary categorical process $(X_t)_{\mathbb{N}}$ said to be

- **serially independent** at lag $k \in \mathbb{N}$

  if $p_{i|j}(k) = p_i$ (i. e., $p_{ij}(k) = p_i p_j$) for any $i, j \in \mathcal{V}$;

- perfectly **serially dependent** at lag $k \in \mathbb{N}$

  if for any $j \in \mathcal{V}$,

  conditional distribution $p_{i|j}(k)$ is one-point distribution.

(. . .)

(. . .)

In case of perfect serial dependence at lag $k \in \mathbb{N}$:

- perfect **positive dependence**

  if $p_{i|j}(k) = 1$ iff $i = j$ for all $i, j \in \mathcal{V}$;

- perfect **negative dependence** if all $p_{i|i}(k) = 0$.

Measures of unsigned dependence:

(Weiß, 2011; Weiß & Göb, 2008)

Goodman and Kruskal's τ:

$$\tau(k) \;=\; \sqrt{\sum_{i,j} \frac{\big(p_{ij}(k) - p_i p_j\big)^2}{p_j\big(1 - s_2(\boldsymbol{p})\big)}} \qquad \text{with range } [0;1],$$

Cramer's v:

$$v(k) \;=\; \sqrt{\frac{1}{m} \cdot \sum_{i,j} \frac{\big(p_{ij}(k) - p_i p_j\big)^2}{p_i p_j}} \qquad \text{with range } [0;1].$$

Some properties:

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

Further measures of unsigned dependence, e. g.,

- (auto-)mutual information function (Dehnert et al., 2003; Biswas & Guha, 2009);
- Pearson measure (Weiß & Göb, 2008, Bagnato et al., 2012);
- Goodman and Kruskal's λ, uncertainty coefficient, Φ^2-measure, Sakoda measure (Weiß & Göb, 2008);
- auto-odds-ratio function (Biswas & Song, 2009); ...

Measures of signed dependence:

(Weiß, 2011; Weiß & Göb, 2008)

Cohen's κ:

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

Modified κ:

$$\kappa^*(k) \;=\; \tfrac{1}{m} \cdot \big(\textstyle\sum_j \;\; p_{j|j}(k) \;\; - 1 \big) \qquad \text{with range } [-\tfrac{1}{m}; 1].$$

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.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Empirical Measures of Serial Dependence

Asymptotic Properties

Empirical Measures of Serial Dependence

Let $(X_t)_{\mathbb{N}}$ be stationary,

we have segment $X_1, \ldots, X_T$ of $(X_t)_{\mathbb{N}}$.

$N_i(T)$ number of variables $X_t = i$ in segment,

$N_{ij}(k,T)$ number of pairs $(X_t, X_{t-k}) = (i,j)$ in segment.

Simple unbiased estimators for p_i and $p_{ij}(k)$:

$$\widehat{p}_i(T) := \tfrac{1}{T} \cdot N_i(T) \qquad \text{and} \qquad \widehat{p}_{ij}(k,T) := \tfrac{1}{T-k} \cdot N_{ij}(k,T).$$

Empirical Measures of Dispersion

Gini index: $\nu_{\mathrm{G}}(X_t) \;=\; \frac{m+1}{m} \cdot \big(1 - s_2(\boldsymbol{p})\big).$

Empirical Gini index:

$$\widehat{\nu}_{\mathrm{G}} \;:=\; \frac{m+1}{m} \cdot \frac{T}{T-1} \cdot \big(1 - s_2(\widehat{\boldsymbol{p}}(T))\big),$$

exactly unbiased if $X_1, \ldots, X_T$ be **i.i.d.** (Weiß, 2011).

Theorem: (Weiß, 2013)

Let $X_1, \ldots, X_T$ stem from stationary NDARMA(p, q) process,

let $c \;:=\; 1 + 2 \cdot \sum_{k=1}^{\infty} \kappa(k).$

Then $1 - s_2\big(\widehat{\boldsymbol{p}}(T)\big)$ asymptotically normal

with mean $1 - s_2(\boldsymbol{p})$ and variance $\frac{4c}{T} \cdot \big(s_3(\boldsymbol{p}) - s_2^2(\boldsymbol{p})\big)$.

Empirical Measures of Dispersion

Entropy: $\nu_{\mathrm{E}}(X_t) \;=\; -\frac{1}{\ln(m+1)} \sum_{j=0}^{m} p_j \ln p_j.$

Empirical entropy:

$$\widehat{\nu}_{\mathrm{E}}(X) \;:=\; -\frac{1}{\ln(m+1)} \sum_{j=0}^{m} \widehat{p}_j(T) \ln \widehat{p}_j(T).$$

Theorem: (Weiß, 2013)

Let $X_1, \ldots, X_T$ stem from stationary NDARMA(p, q) process,

let $c \;:=\; 1 + 2 \cdot \sum_{k=1}^{\infty} \kappa(k).$

Then $-\sum_{j=0}^{m} \widehat{p}_j(T) \ln \widehat{p}_j(T)$ asymptotically normal

with mean $-\sum_{j=0}^{m} p_j \ln p_j$

and variance $\frac{c}{T} \cdot \Big(\sum_{i=0}^{m} p_i \cdot (\ln p_i)^2 \;-\; \big(\sum_{i=0}^{m} p_i \cdot \ln p_i \big)^2 \Big).$

Empirical Measures of Serial Dependence

Cohen's κ: $$\kappa(k) \;=\; 1 - \frac{1 - \sum_j p_{jj}(k)}{1 - s_2(\boldsymbol{p})}.$$

Empirical Cohen's κ:

$$\widehat{\kappa}(k) \;:=\; 1 + \tfrac{1}{T} \;-\; \frac{1 - \sum_j \widehat{p}_{jj}(k,T)}{1 - \sum_j \widehat{p}_j(T)^2}.$$

Theorem: (Weiß, 2011)

If $X_1, \ldots, X_T$ are **i.i.d.**, then

$\widehat{\kappa}(k)$ asymptotically normally distributed with

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

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

Genome of Bovine Leukemia Virus: (Weiß & Göb, 2008)

Range of size 4 ($\Rightarrow m = 3$),

coding a $\mapsto$ 0, c $\mapsto$ 1, g $\mapsto$ 2, t $\mapsto$ 3,

length $T = 8419$.

Estimated marginal probabilities:

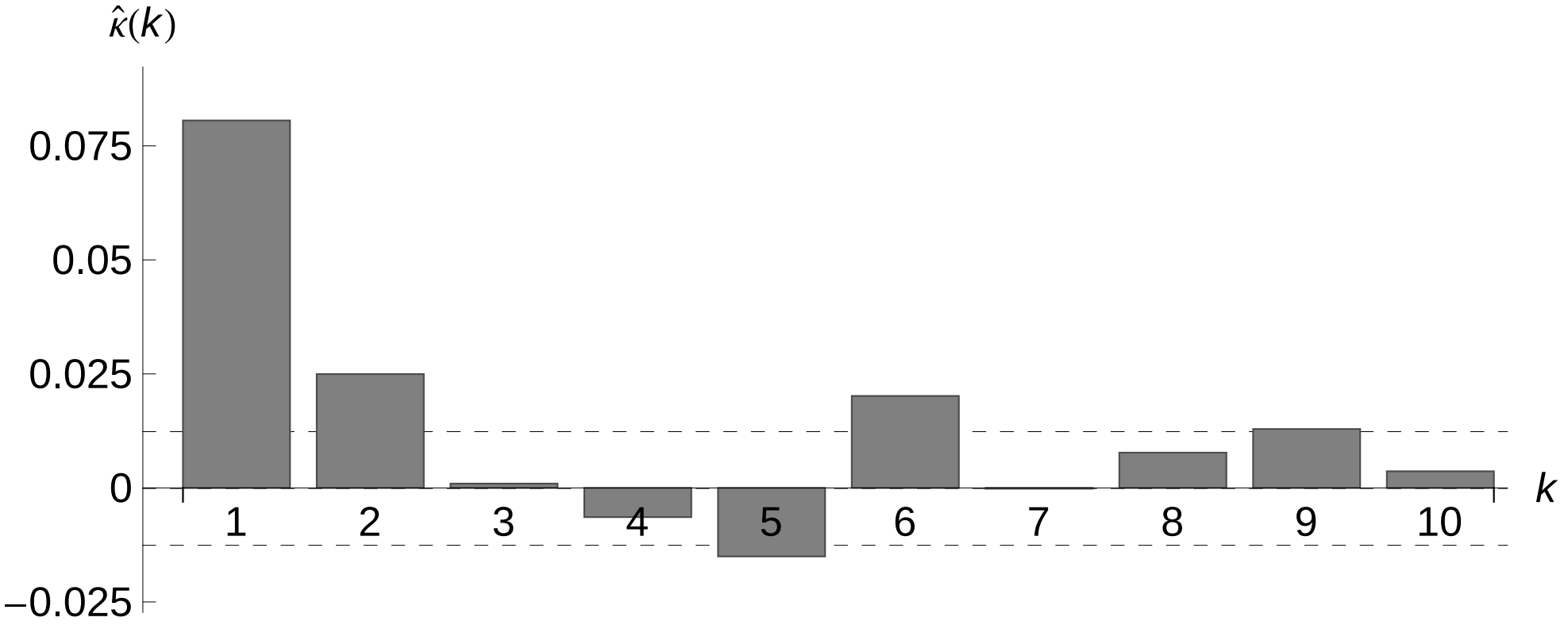
$\hat{p}_0 = 0.220$, $\hat{p}_1 = 0.331$, $\hat{p}_2 = 0.210$, $\hat{p}_3 = 0.239$

$\Rightarrow$ Gini index $\hat{\nu}_{\mathrm{G}} \approx 0.988$ (s.e. 0.0016),

entropy $\hat{\nu}_{\mathrm{E}} \approx 0.987$ (s.e. 0.0016) (strong dispersion).

Genome of Bovine Leukemia Virus: (continued)

(Approximate) asymptotic standard error for $\widehat{\kappa}(k)$: 0.00636.



Theorem 5.1.1 in Weiß (2013):

Asymptotics of $\widehat{\kappa}(k)$ for stationary NDARMA(p, q) process.

Corollary:

Let $X_1, \ldots, X_T$ stem from stationary DAR(1) process,

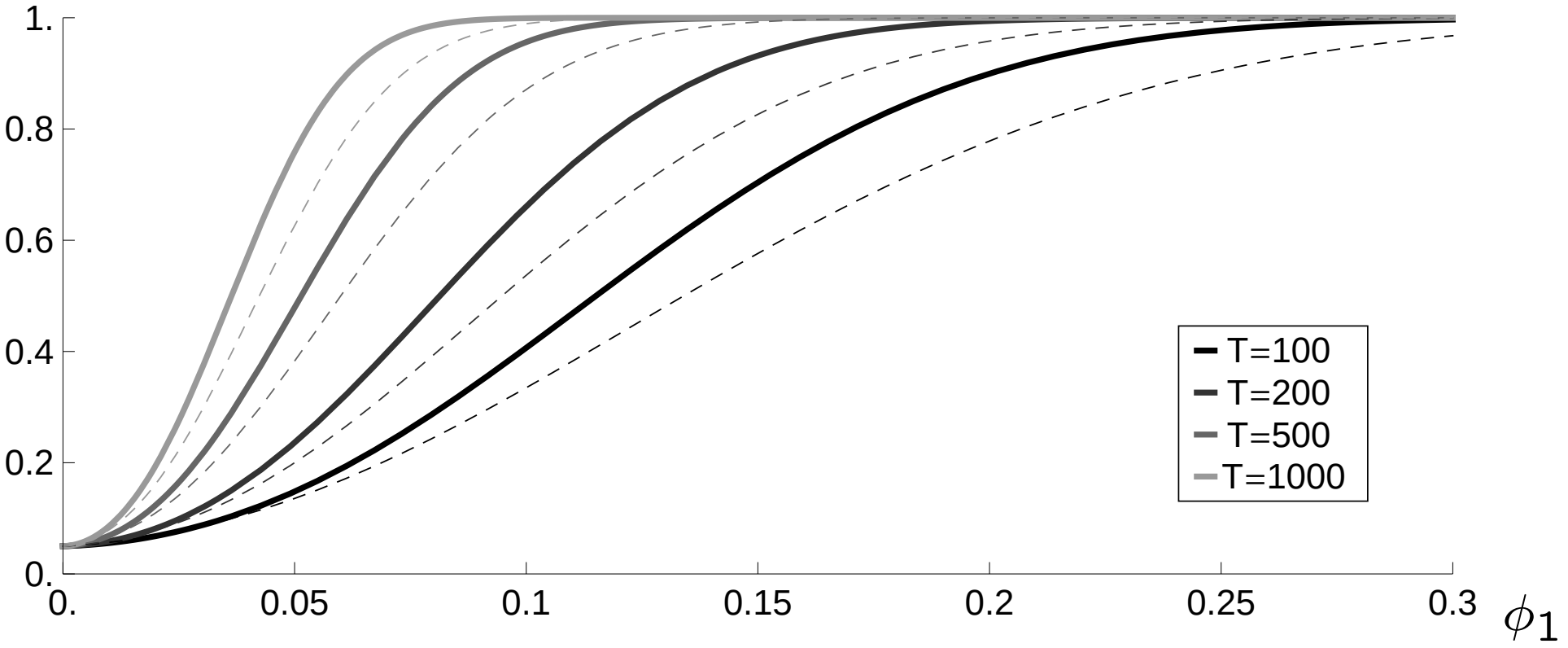
then $\widehat{\kappa}(1)$ asymptotically normal with mean ϕ_1 and variance

$$\frac{1}{T}\Big(2\cdot(2\phi_1-1)\cdot(1-\phi_1)\cdot\frac{s_3(\boldsymbol{p})-s_2^2(\boldsymbol{p})}{(1-s_2(\boldsymbol{p}))^2} + \frac{(1-\phi_1)\cdot s_2(\boldsymbol{p})}{1-s_2(\boldsymbol{p})} + \phi_1(1-\phi_1)\Big).$$

Applications: parameter estimation, power analysis ...

Power functions if testing $\widehat{\kappa}(1)$ against 0:

(marginals $(0.2, 0.2, 0.25, 0.35)^\top$ (solid) and $(0.05, 0.1, 0.15, 0.7)^\top$ (dashed))



Analogous results also available for **empirical modified** κ,

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

see Weiß (2011,2013).

But:

Practical issues like division by zero, finite-sample bias, ...

$\Rightarrow$ **empirical Cohen's** κ

preferable measure of *signed* serial dependence.

Empirical Measures of Serial Dependence

Cramer's v: $$v(k) = \sqrt{\frac{1}{m} \cdot \sum_{i,j=0}^{m} \frac{\big(p_{ij}(k) - p_i p_j\big)^2}{p_i p_j}}.$$

Empirical Cramer's v:

$$\widehat{v}^2(k) = \frac{1}{m} \cdot \sum_{i,j=0}^{m} \frac{\big(\widehat{p}_{ij}(k,T) - \widehat{p}_i(T)\widehat{p}_j(T)\big)^2}{\widehat{p}_i(T)\widehat{p}_j(T)}.$$

Theorem: (Weiß, 2013)

If $X_1, \ldots, X_T$ are **i.i.d.**, then

$\widehat{v}(k)$ asymptotically distributed via

$$T \cdot m \cdot \widehat{v}^2(k) \underset{\text{a.}}{\sim} \chi^2_{m^2}.$$

Asymptotic χ^2-distribution (confirmed through simulations) of empirical Cramer's v implies non-zero mean although $v(k) = 0$ for i.i.d. data!

$\Rightarrow \widehat{v}(k)$ problematic as estimator of $v(k)$;
similar problems with any unsigned measure.

Certainly useful for uncovering significant dependence.

Simulations: Less powerful than Cohen's κ.

Bovine data: $\widehat{v}(k)$ as 0.1134, 0.0445, 0.0281, 0.0222, ...
Crit. value 0.0259 (level 0.05) $\Rightarrow \widehat{v}(1)$, $\widehat{v}(2)$, $\widehat{v}(3)$ significant.

Analogous results also available for

empirical Goodman and Kruskal's τ,

$$\widehat{\tau}^2(k) \;=\; \sum_{i,j=0}^{m} \frac{\big(\widehat{p}_{ij}(k,T) - \widehat{p}_i(T)\widehat{p}_j(T)\big)^2}{\widehat{p}_j(T)\big(1 - s_2(\widehat{\boldsymbol{p}}(T))\big)},$$

see Weiß (2011,2013).

But asymptotic distribution now

quadratic form distribution instead of simple χ^2-distribution

$\Rightarrow$ $\widehat{\tau}(k)$ less attractive from practical point of view.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Conclusions

. . . and Future Research

Conclusions

- Both unsigned and signed empirical measures of serial dependence in categorical time series available, readily applicable to uncover significant dependence.
- The signed measure Cohen's κ appears most attractive, because also useful for parameter estimation, and good agreem. betw. asymptotic & finite-sample properties.
- Gini index useful measure of dispersion in categ. time series.

Future research:

- Effect of estimated parameters;
- unique framework for serial dependence.

Thank You

for Your Interest!

MATH
STAT

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

weissc@hsu-hh.de

Literature

Bagnato, Punzo, Nicolis, 2012. The autodependogram: a graphical device to investigate serial dependences. J. Time Ser. Anal. 33, 233–254.

Biswas, Guha, 2009. Time series analysis of categorical data using auto-mutual information. J. Statist. Plann. Inf. 139, 3076–3087.

Biswas, Song, 2009. Discrete-valued ARMA processes. Statist. Probab. Letters 79, 1884–1889.

Bühlmann, Wyner, 1999. Variable length M. chains. Ann. Stat. 27, 480–513.

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

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

Song et al., 2013. Statistical analysis of discrete-valued time series using categorical ARMA models. Comp. Statist. Data Anal. 57, 112–124.

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

Weiß, 2013. Serial Dependence of NDARMA Processes. Comp. Statist. Data Anal. 68, 213–238.

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