Empirical Measures of Signed Serial Dependence in Categorical Time Series

Christian H. Weiß
Department of Mathematics,
Darmstadt University of Technology
This talk is based on the article

*Empirical measures of signed serial dependence in categorical time series.*

All references mentioned in this talk correspond to the references in this article.
Categorical Time Series Analysis

Brief Review
Categorical process:

$$(X_t)_\mathbb{N} \text{ with } \mathbb{N} = \{1, 2, \ldots \}, \text{ where each }$$

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

Christian H. Weiß — Darmstadt University of Technology
Notations for time-invariant probabilities:

If \((X_t)_N\) (strictly) stationary, then:

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

- bivariate probabilities \(p_{ij}(k) := P(X_t = i, X_{t-k} = j)\),
  conditional probabilities \(p_{i|j}(k) := P(X_t = i \mid X_{t-k} = j)\).
Let \((X_t)_N\) 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.**).
Most simple measure of dispersion: **Gini index** of $X_t$,

$$\nu_G(X_t) := \frac{m+1}{m} \cdot (1 - \sum_j p_j^2) = \frac{m+1}{m} \cdot (1 - s_2(p)).$$

- continuous and symmetric function of $p_1, \ldots, p_{m+1}$,
- range $[0; 1]$,
- maximal value 1 iff uniform distribution,
- minimal value 0 iff one-point distribution.

Categorical Time Series Analysis

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

Stationary categorical process \((X_t)_N\) said to be

- **serially independent** at lag \(k \in \mathbb{N}\)
  
  if \(p_i|j(k) = p_i\) (i.e., \(p_{ij}(k) = p_ip_j\)) for any \(i, j \in V\);

- **perfectly serially dependent** at lag \(k \in \mathbb{N}\)
  
  if for any \(j \in V\),
  
  conditional distribution \(p_{i|j}(k)\) is one-point distribution.

(\ldots)
(...)

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 $p_{i|i}(k) = 0$. 

Christian H. Weiß — Darmstadt University of Technology
Categorical Time Series Analysis

Measures of signed dependence:

(Weiβ, 2011; Weiβ & Göb, 2008)

Cohen’s $\kappa$:

$$\kappa(k) = 1 - \frac{1 - \sum_j p_{jj}(k)}{1 - s_2(p)}$$

with range \([-\frac{s_2(p)}{1-s_2(p)} ; 1]\).

Modified $\kappa$:

$$\kappa^*(k) = \frac{1}{m} \cdot (\sum_j p_{j|j}(k) - 1)$$

with range \([-\frac{1}{m} ; 1]\).

Some properties:

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

Cohen’s $\kappa$:

$$\kappa(k) = 1 - \frac{1 - \sum_j p_{jj}(k)}{1 - s_2(p)}$$

with range $[-\frac{s_2(p)}{1-s_2(p)}; 1]$.

Modified $\kappa$:

$$\kappa^*(k) = \frac{1}{m} \cdot (\sum_j p_{j|j}(k) - 1)$$

with range $[-\frac{1}{m}; 1]$.

Gini index $\nu_G$:

$$\nu_G(X_t) = \frac{m+1}{m} \cdot (1 - s_2(p))$$

with range $[0; 1]$. 

Christian H. Weiß — Darmstadt University of Technology
Let \((X_t)_N\) be stationary, we have segment \(X_1, \ldots, X_T\) of \((X_t)_N\).

\[ N_i(T) \text{ number of variables } X_t = i \text{ in segment,} \]
\[ N_{ij}(k, T) \text{ number of pairs } (X_t, X_{t-k}) = (i, j) \text{ in segment.} \]

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

\[ \hat{p}_i(T) := \frac{1}{T} \cdot N_i(T) \quad \text{and} \quad \hat{p}_{ij}(k, T) := \frac{1}{T-k} \cdot N_{ij}(k, T). \]
Lemma:
Let $X_1, \ldots, X_T$ be i.i.d.
Estimator $1 - \sum_j \hat{p}_j(T)^2$ of $1 - s_2(p)$ satisfies

$$E[1 - \sum_j \hat{p}_j(T)^2] = 1 - s_2(p) - \frac{1}{T} \cdot (1 - s_2(p)),$$

$$V[1 - \sum_j \hat{p}_j(T)^2] = \frac{4}{T} \cdot (s_3(p) - s_2^2(p)) + O(T^{-2}).$$

$\Rightarrow$ Define exactly unbiased empirical Gini index via

$$\hat{\nu}_G := \frac{m+1}{m} \cdot \frac{T}{T-1} \cdot (1 - \sum_j \hat{p}_j(T)^2).$$

Christian H. Weiß — Darmstadt University of Technology
**Lemma:**

Let $X_1, \ldots, X_T$ be \textbf{i.i.d.}

Enumerator $1 - \sum_i p_{ii}(k)$ of Cohen's $\kappa$:

\[
E\left[1 - \sum_i \hat{p}_{ii}(k, T)\right] = 1 - s_2(p),
\]

\[
V\left[1 - \sum_i \hat{p}_{ii}(k, T)\right] = \frac{1}{T-k} \cdot \left( s_2(p)(1 - s_2(p)) + 2(s_3(p) - s_2^2(p)) \right) + O(T^{-2}).
\]
Theorem: Define empirical Cohen’s $\kappa$ as

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

If $X_1, \ldots, X_T$ is i.i.d., then

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

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

$$V[\hat{\kappa}(k)] = \frac{1}{T} \cdot \left( 1 - \frac{1 + 2s_3(p) - 3s_2(p)}{(1-s_2(p))^2} \right) + O(T^{-2}).$$
Theorem: Define empirical modified $\kappa$ as

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

If $X_1, \ldots, X_T$ is i.i.d., then $\hat{\kappa}^*(k)$ asymptotically normally distributed with

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

$$V[\hat{\kappa}^*(k)] = \frac{1}{mT} + O(T^{-2}).$$
Empirical Measures of Signed Serial Dependence

An Application
Measured serial dependence at lag \( k \) called **significantly different** from 0 if

\[
|\hat{\kappa}(k)| > c \cdot \sqrt{\frac{1}{T} \cdot \left(1 - \frac{1+2s_3(\hat{p})-3s_2(\hat{p})}{(1-s_2(\hat{p}))^2}\right)}, \quad \text{or}
\]

\[
|\hat{\kappa}^*(k)| > c \cdot \sqrt{\frac{1}{m \cdot T}}.
\]

Common choice: \( c = 1.96 \quad (\approx \text{significance level } 5\%) \).

Concerning \( \hat{\kappa}(k) \), we used \( \hat{p} := \hat{p}(T) \) instead of true \( p \), since latter hardly known in practice.
Data Example:
Genome of Bovine Leukemia Virus, as in Weiß & Göb (2008).

Range of size 4 (⇒ m = 3),
coding a ⇔ 0, c ⇔ 1, g ⇔ 2, t ⇔ 3,
length $T = 8419$.

Estimated marginal probabilities:
$\hat{p}_0 = 0.220$, $\hat{p}_1 = 0.331$, $\hat{p}_2 = 0.210$, $\hat{p}_3 = 0.239$
⇒ Gini index $\hat{\nu}_G \approx 0.988$ (strong dispersion).
Data Example: (continued)

(Approximate) asymptotic standard error for $\hat{\kappa}(k)$: 0.00636.
Data Example: (continued)

Asymptotic standard error for $\hat{\kappa}^*(k)$: 0.00629.
Empirical Measures of Signed Serial Dependence

Finite-Sample Properties
Selected results of simulation study,  
detailed tables in Weiß (2011).  

Models with range of size 4 (i. e., $m = 3$).  

Study of true **significance level**:

‘i.i.d.-1’:  \[ p_1 = (0.20, 0.20, 0.25, 0.35)^\top, \nu_G(X) = 0.98 \]

‘i.i.d.-2’:  \[ p_2 = (0.05, 0.10, 0.15, 0.70)^\top, \nu_G(X) = 0.633. \]
Design of simulation study (continued):

Study of true power:

‘DAR(1)’: \( p_1 \) and \( \phi = 0.25 \) \( \Rightarrow \kappa(k) = \kappa^*(k) = 0.25^k \).

‘DMA(1)’: \( p_1 \) and \( \varphi = 0.25 \) \( \Rightarrow \kappa(1) = \kappa^*(1) = 0.1875, \kappa(k) = \kappa^*(k) = 0 \text{ for } k \geq 2 \).

‘NegMarkov’: \( p_1 \) and \( \alpha = 0.5 \) \( \Rightarrow \kappa(1) = -0.2678, \kappa(2) = 0.0818, \kappa(3) = -0.0276, \ldots \)
\kappa^*(1) = -0.2500, \kappa^*(2) = 0.0719, \kappa^*(3) = -0.0232, \ldots
Empirical rejection rates for $\hat{\kappa}(k)$:

<table>
<thead>
<tr>
<th>$k$ \ $T$</th>
<th>i.i.d.-1</th>
<th>i.i.d.-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>1</td>
<td>5.1</td>
<td>4.7</td>
</tr>
<tr>
<td>2</td>
<td>4.9</td>
<td>4.9</td>
</tr>
<tr>
<td>3</td>
<td>5.1</td>
<td>4.9</td>
</tr>
<tr>
<td>4</td>
<td>5.1</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>5.2</td>
<td>5.1</td>
</tr>
</tbody>
</table>

$\Rightarrow$ always close to nominal level of 5 %.
Empirical rejection rates for $\widehat{\kappa}^*(k)$:

<table>
<thead>
<tr>
<th>$k$ \ $T$</th>
<th>i.i.d.-1</th>
<th></th>
<th></th>
<th></th>
<th>i.i.d.-2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>5.1</td>
<td>4.6</td>
<td>4.8</td>
<td>4.9</td>
<td>3.7</td>
<td>4.4</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>4.8</td>
<td>4.9</td>
<td>5.0</td>
<td>5.5</td>
<td>3.9</td>
<td>4.2</td>
<td>4.6</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>4.8</td>
<td>5.0</td>
<td>5.0</td>
<td>4.9</td>
<td>3.8</td>
<td>4.0</td>
<td>4.7</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>5.1</td>
<td>4.9</td>
<td>5.0</td>
<td>5.2</td>
<td>3.9</td>
<td>4.3</td>
<td>5.1</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>5.2</td>
<td>4.9</td>
<td>4.9</td>
<td>5.0</td>
<td>3.5</td>
<td>4.0</td>
<td>4.6</td>
</tr>
</tbody>
</table>

$\Rightarrow$ for medium dispersion and $T \leq 200$, even below nominal level of 5 %.

Christian H. Weiß — Darmstadt University of Technology
Empirical rejection rates for DAR(1) model:

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\hat{\kappa}(k)$</th>
<th>$\hat{\kappa}^*(k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>1</td>
<td>97.4</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>18.3</td>
<td>31.5</td>
</tr>
<tr>
<td>3</td>
<td>7.1</td>
<td>7.9</td>
</tr>
<tr>
<td>4</td>
<td>6.3</td>
<td>6.5</td>
</tr>
<tr>
<td>5</td>
<td>6.4</td>
<td>6.6</td>
</tr>
</tbody>
</table>

⇒ similar performance for both measures,

at least 1$^{st}$ order dependence nearly always detected.
Empirical rejection rates for DMA(1) model:

<table>
<thead>
<tr>
<th>$k$ \ $T$</th>
<th>$\hat{\kappa}(k)$</th>
<th>$\hat{\kappa}^*(k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 200 500 1000</td>
<td>100 200 500 1000</td>
</tr>
<tr>
<td>1</td>
<td>86.7 99.3 100.0 100.0</td>
<td>87.4 99.3 100.0 100.0</td>
</tr>
<tr>
<td>2</td>
<td>5.4 5.7 5.6 5.7</td>
<td>5.3 5.5 5.6 5.6</td>
</tr>
<tr>
<td>3</td>
<td>5.9 5.8 5.9 5.6</td>
<td>5.7 5.5 5.9 5.5</td>
</tr>
<tr>
<td>4</td>
<td>5.9 5.8 5.7 5.4</td>
<td>5.5 5.7 6.0 5.6</td>
</tr>
<tr>
<td>5</td>
<td>6.3 6.1 5.6 5.7</td>
<td>5.9 6.0 5.7 5.8</td>
</tr>
</tbody>
</table>

⇒ similar performance for both measures.
For $T \geq 200$, $1^{st}$ order dependence nearly always detected.
For $k \geq 2$, slightly larger than 5%.

Christian H. Weiß — Darmstadt University of Technology
Empirical rejection rates for NegMarkov model:

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\hat{\kappa}(k)$ \ T</th>
<th>$\hat{\kappa}^*(k)$ \ T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.0 100.0 100.0 100.0</td>
<td>100.0 100.0 100.0 100.0</td>
</tr>
<tr>
<td>2</td>
<td>32.3 52.6 86.1 98.8</td>
<td>26.0 43.7 78.2 96.7</td>
</tr>
<tr>
<td>3</td>
<td>9.7 11.9 19.6 32.1</td>
<td>8.2 9.9 15.3 24.7</td>
</tr>
<tr>
<td>4</td>
<td>8.3 8.7 8.7 11.7</td>
<td>7.0 7.7 8.0 10.3</td>
</tr>
<tr>
<td>5</td>
<td>7.7 7.5 7.4 8.5</td>
<td>7.0 6.7 7.0 7.7</td>
</tr>
</tbody>
</table>

$\Rightarrow \hat{\kappa}^*(k)$ worse than $\hat{\kappa}(k)$,

at least 1$^{st}$ order dependence nearly always detected.
Conclusions

- Empirical measures of signed serial dependence, effective for identifying significant dependence.

- Finite-sample study shows that overall, $\hat{\kappa}(k)$ is best choice.
  For $T \geq 500$, both measures perform equivalently.

- **Work in progress:**
  $\hat{\kappa}(k)$, $\hat{\kappa}^*(k)$ and also empirical measures of unsigned dependence for NDARMA processes.
  Empirical dispersion measures for NDARMA processes.

Christian H. Weiβ — Darmstadt University of Technology
Thank You
for Your Interest!

Christian H. Weiß
Department of Mathematics
Darmstadt University of Technology
weiss@mathematik.tu-darmstadt.de