Empirical Measures of Signed Serial Dependence in Categorical Time Series

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich **Mathematik**

Christian H. Weiß

Department of Mathematics,

Darmstadt University of Technology

This talk is based on the article

Weiß, C.H. (2011).
Empirical measures of signed serial dependence in categorical time series.
Journal of Statistical Computation and Simulation 81(4), 411–429.

All references mentioned in this talk correspond to the references in this article.

Categorical Time Series Analysis

Brief Review

Categorical process:

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

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

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.**).

Most simple measure of dispersion: **Gini index** of X_t,

$$\nu_{\mathsf{G}}(X_t) \; := \; \tfrac{m+1}{m}\cdot\Big(1-\textstyle\sum_j\; p_j^2\Big) \;=\; \tfrac{m+1}{m}\cdot\Big(1-s_2(\boldsymbol{p})\Big).$$

- 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.

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 signed dependence:

(Weiß, 2011; Weiß & Göb, 2008)

Cohen's κ:

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

Modified κ:

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

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.

Empirical Measures of Signed Serial Dependence

Asymptotic Properties

Just to remember . . .

Cohen's κ:

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

Modified κ:

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

Gini index ν_{G}:

$$\nu_{\mathrm{G}}(X_t) \;=\; \tfrac{m+1}{m}\cdot\Big(1-s_2(\boldsymbol{p})\Big) \qquad \text{with range } [0;1].$$

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

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

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

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

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

Lemma:

Let $X_1, \ldots, X_T$ be **i.i.d.**

Estimator $1 - \sum_j \widehat{p}_j(T)^2$ of $1 - s_2(\boldsymbol{p})$ satisfies

$$E\Big[1 - \textstyle\sum_j \widehat{p}_j(T)^2\Big] \;=\; 1 - s_2(\boldsymbol{p}) \;-\; \tfrac{1}{T} \cdot \Big(1 - s_2(\boldsymbol{p})\Big),$$

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

$\Rightarrow$ Define exactly unbiased **empirical Gini index** via

$$\widehat{\nu}_{\mathrm{G}} \;:=\; \tfrac{m+1}{m} \cdot \tfrac{T}{T-1} \cdot \Big(1 - \textstyle\sum_j \widehat{p}_j(T)^2\Big).$$

Lemma:

Let $X_1, \ldots, X_T$ be **i.i.d.**

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

$$E\Big[1 - \textstyle\sum_i\ \widehat{p}_{ii}(k,T)\Big]\ =\ 1 - s_2(\boldsymbol{p}),$$

$$V\Big[1 - \textstyle\sum_i\ \widehat{p}_{ii}(k,T)\Big]\ =$$
$$\tfrac{1}{T-k}\cdot\Big(s_2(\boldsymbol{p})\big(1 - s_2(\boldsymbol{p})\big)\ +\ 2\big(s_3(\boldsymbol{p}) - s_2^2(\boldsymbol{p})\big)\Big)\ +\ O(T^{-2}).$$

Theorem: Define **empirical Cohen's** κ as

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

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

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

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

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

Theorem: Define **empirical modified** κ as

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

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

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

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

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

Empirical Measures of Signed Serial Dependence

An Application

Measured serial dependence at lag k called

significantly different from 0 if

$$\big|\hat{\kappa}(k)\big| \;>\; c\cdot\sqrt{\tfrac{1}{T}\cdot\Big(1\;-\;\tfrac{1+2s_3(\hat{\boldsymbol{p}})-3s_2(\hat{\boldsymbol{p}})}{(1-s_2(\hat{\boldsymbol{p}}))^2}\Big)},\qquad \text{or}$$

$$\big|\hat{\kappa}^*(k)\big| \;>\; c\cdot\sqrt{\tfrac{1}{m\cdot T}}.$$

Common choice: $c = 1.96$ ($\approx$ significance level 5 %).

Concerning $\hat{\kappa}(k)$, we used $\hat{\boldsymbol{p}} := \hat{\boldsymbol{p}}(T)$ instead of true $\boldsymbol{p}$, since latter hardly known in practice.

Data Example:

Genome of Bovine Leukemia Virus, as in Weiß & Göb (2008).

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

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

length $T = 8419$.

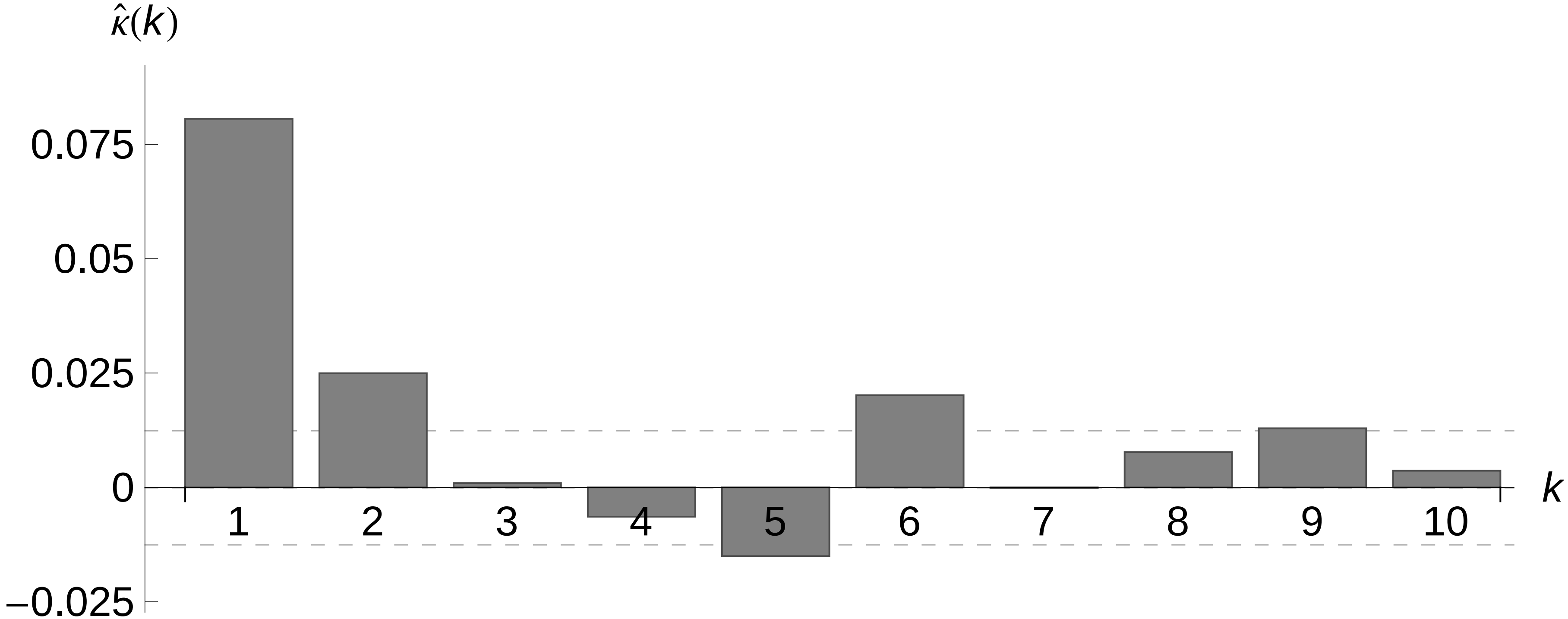
Estimated marginal probabilities:

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

$\Rightarrow$ Gini index $\widehat{\nu}_{\mathsf{G}} \approx 0.988$ (strong dispersion).

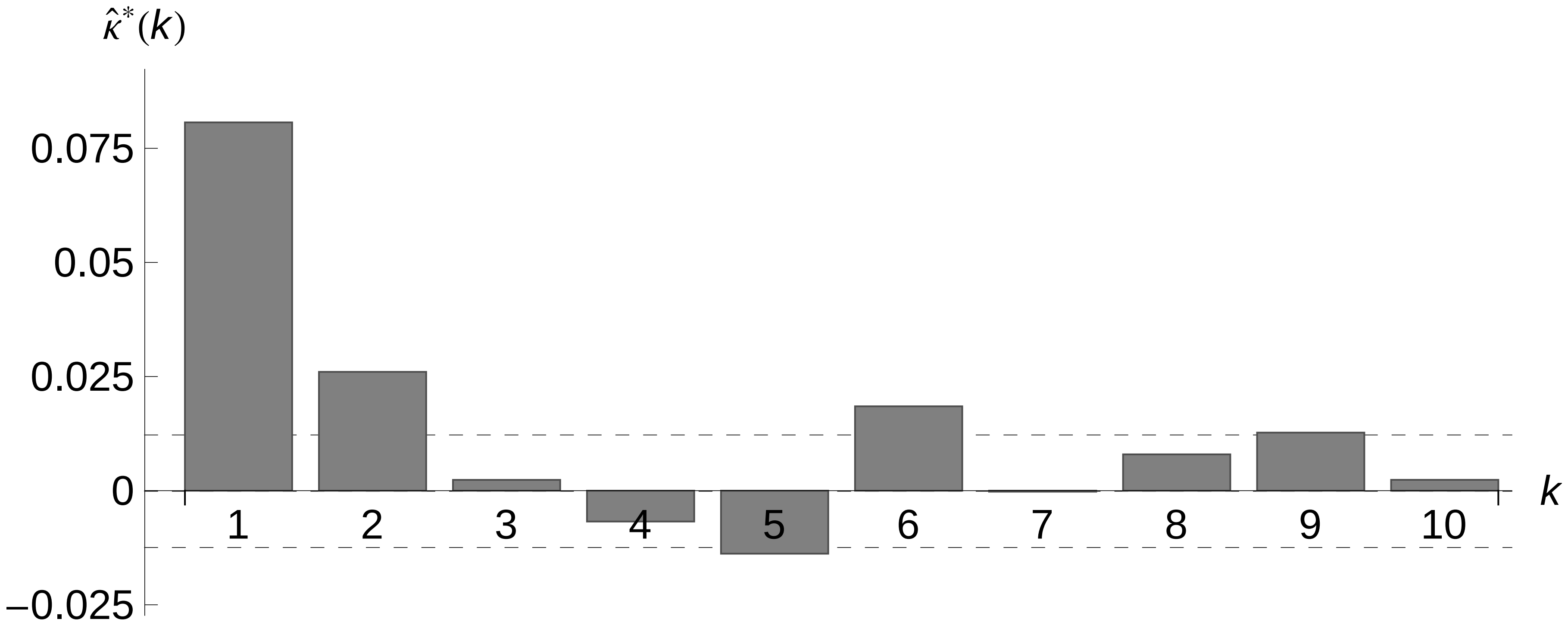
Data Example: (continued)

(Approximate) asymptotic standard error for $\widehat{\kappa}(k)$: 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': $\boldsymbol{p}_1 = (0.20, 0.20, 0.25, 0.35)^\top$, $\nu_{\mathsf{G}}(X) = 0.98$

'i.i.d.-2': $\boldsymbol{p}_2 = (0.05, 0.10, 0.15, 0.70)^\top$, $\nu_{\mathsf{G}}(X) = 0.633$.

Design of simulation study (continued):

Study of true **power**:

'DAR(1)': $\boldsymbol{p}_1$ and $\phi = 0.25 \Rightarrow \kappa(k) = \kappa^*(k) = 0.25^k$.

'DMA(1)': $\boldsymbol{p}_1$ and $\varphi = 0.25 \Rightarrow$

$\kappa(1) = \kappa^*(1) = 0.1875$, $\kappa(k) = \kappa^*(k) = 0$ for $k \geq 2$.

'NegMarkov': $\boldsymbol{p}_1$ and $\alpha = 0.5 \Rightarrow$

$\kappa(1) = -0.2678$, $\kappa(2) = 0.0818$, $\kappa(3) = -0.0276$, ...

$\kappa^*(1) = -0.2500$, $\kappa^*(2) = 0.0719$, $\kappa^*(3) = -0.0232$, ...

Empirical rejection rates for $\widehat{\kappa}(k)$:

	i.i.d.-1				i.i.d.-2			
$k \setminus T$	100	200	500	1000	100	200	500	1000
1	5.1	4.7	4.8	4.9	5.0	4.6	4.7	4.9
2	4.9	4.9	4.8	5.5	5.5	4.9	4.9	5.4
3	5.1	4.9	5.0	5.1	5.2	5.2	5.4	5.0
4	5.1	5.0	5.0	5.2	5.6	5.6	5.2	4.7
5	5.2	5.1	5.0	5.1	5.7	5.2	5.1	5.1

$\Rightarrow$ always close to nominal level of 5 %.

Empirical rejection rates for $\widehat{\kappa}^*(k)$:

	i.i.d.-1				i.i.d.-2			
$k \setminus T$	100	200	500	1000	100	200	500	1000
1	5.1	4.6	4.8	4.9	3.7	4.4	4.5	5.2
2	4.8	4.9	5.0	5.5	3.9	4.2	4.6	5.0
3	4.8	5.0	5.0	4.9	3.8	4.0	4.7	5.0
4	5.1	4.9	5.0	5.2	3.9	4.3	5.1	4.4
5	5.2	4.9	4.9	5.0	3.5	4.0	4.6	5.0

$\Rightarrow$ for medium dispersion and $T \leq 200$,

even below nominal level of 5 %.

Empirical rejection rates for DAR(1) model:

	$\widehat{\kappa}(k)$				$\widehat{\kappa}^*(k)$			
$k \quad \backslash \quad T$	100	200	500	1000	100	200	500	1000
1	97.4	100.0	100.0	100.0	97.2	100.0	100.0	100.0
2	18.3	31.5	63.0	88.9	17.9	31.5	63.6	89.4
3	7.1	7.9	10.8	14.1	6.6	7.9	10.7	14.6
4	6.3	6.5	6.6	6.7	6.0	6.3	6.6	6.8
5	6.4	6.6	6.6	6.4	6.0	6.2	6.4	6.5

$\Rightarrow$ similar performance for both measures,

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

Empirical rejection rates for DMA(1) model:

	$\widehat{\kappa}(k)$				$\widehat{\kappa}^*(k)$			
$k \setminus T$	100	200	500	1000	100	200	500	1000
1	86.7	99.3	100.0	100.0	87.4	99.3	100.0	100.0
2	5.4	5.7	5.6	5.7	5.3	5.5	5.6	5.6
3	5.9	5.8	5.9	5.6	5.7	5.5	5.9	5.5
4	5.9	5.8	5.7	5.4	5.5	5.7	6.0	5.6
5	6.3	6.1	5.6	5.7	5.9	6.0	5.7	5.8

$\Rightarrow$ similar performance for both measures.

For $T \geq 200$, 1$^{\text{st}}$ order dependence nearly always detected.

For $k \geq 2$, slightly larger than 5 %.

Empirical rejection rates for NegMarkov model:

	$\widehat{\kappa}(k)$				$\widehat{\kappa}^*(k)$			
$k \quad \backslash \quad T$	100	200	500	1000	100	200	500	1000
1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	32.3	52.6	86.1	98.8	26.0	43.7	78.2	96.7
3	9.7	11.9	19.6	32.1	8.2	9.9	15.3	24.7
4	8.3	8.7	8.7	11.7	7.0	7.7	8.0	10.3
5	7.7	7.5	7.4	8.5	7.0	6.7	7.0	7.7

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

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 overally, $\widehat{\kappa}(k)$ is best choice. For $T \geq 500$, both measures perform equivalently.

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

Thank You

for Your Interest!

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich **Mathematik**

Christian H. Weiß

Department of Mathematics

Darmstadt University of Technology

weiss@mathematik.tu-darmstadt.de