Ordinal Compositional Data and Time Series

Christian H. Weiß

Department of Mathematics & Statistics,

Helmut Schmidt University, Hamburg

MATH
STAT

Funded by DFG – Projektnr. 516522977.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Ordinal Compositional Data and Time Series

Introduction

CoDa = "proportions of some whole" (Aitchison, 1986).

$\boldsymbol{x} = (x_0, \ldots, x_d)^\top$ normalized $(d+1)$-part composition

with $d \in \mathbb{N} = \{1, 2, \ldots\}$ if $\boldsymbol{x}$ from $(d+1)$-part unit simplex

$$\mathbb{S} := \left\{ \boldsymbol{x} \in (0;1)^{d+1} \;\middle|\; x_0 + \ldots + x_d = 1 \right\}.$$

Connection to **categorical data** (Agresti, 2002):

CoDa vectors $\boldsymbol{p} \in \mathbb{S}$ might serve as PMF of

categorical RV Q with qualitative range $\mathcal{S} = \{s_0, \ldots, s_d\}$.

In many applications, categories behind CoDa unordered,

so $\mathcal{S}$ is nominal range and Q nominal RV,

notation "$s_0, \ldots, s_d$" just uses some lexicographic order.

For such nominal CoDa, reasonable to require that "applying a compositional analysis, the information due to the order of the different classes, plays no role", see Pawlowsky-Glahn & Buccianti (2011).

But as conceded by Pawlowsky-Glahn & Buccianti (2011), "in some cases the parts can be assumed to be ordered", so $\mathcal{S}$ exhibits natural order $s_0 < \ldots < s_d$ and Q is ordinal RV (Agresti, 2010). Then, $\boldsymbol{x} \in \mathbb{S}$ called **ordinal composition**.

Example: $\boldsymbol{x} = (x_0, x_1, x_2)^\top$ expressing proportions of people in ordered age categories < 15 (x_0), 15–60 (x_1), and > 60 (x_2).

Introduction

If concerned with **ordinal CoDa**, certainly still justified to apply well-established CoDa approaches, although ignoring "information due to the order of the different classes".

However, gain additional insights if existing CoDa approaches complemented by new ones accounting for order within $\mathcal{S}$.

Such **new ordinal CoDa approaches** derived by adapting well-established concepts from ordinal data analysis.

Benefits illustrated for *descriptive analysis* of ordinal CoDa,
statistical inference from ordinal CoDa,
monitoring of ordinal CoDa process, and
ordinal compositional *time series* (CoTS).

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Ordinal Compositional Data

Descriptive Analysis

Descriptive Analysis of Ordinal CoDa

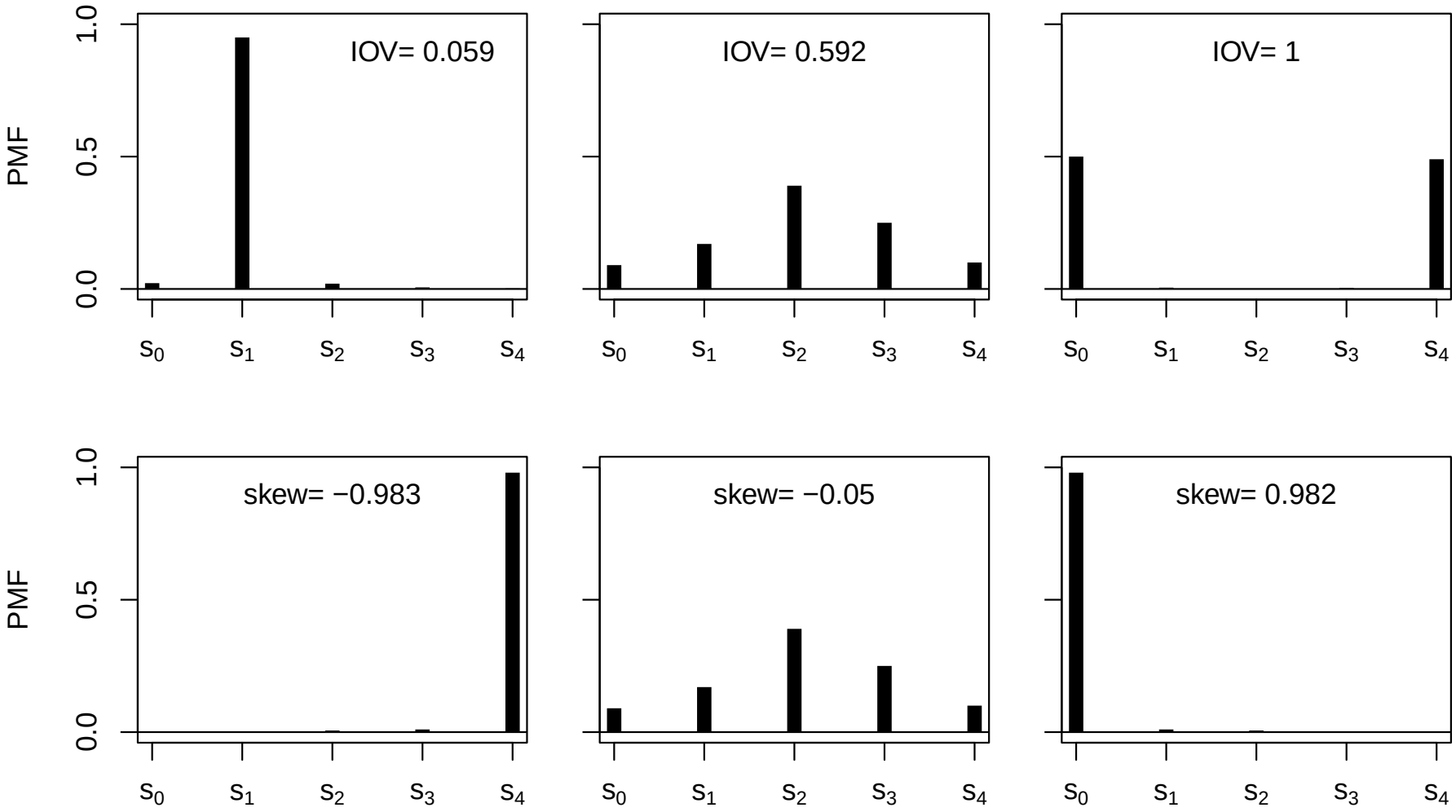
For **ordinal RV** Q, one does not focus on PMF $\boldsymbol{p} \in \mathbb{S}$, but on d-dimensional CDF vector $\boldsymbol{f} = (f_0, \ldots, f_{d-1})^\top \in [0;1]^d$, where $f_j = P(Q \leq s_j)$ (note that $f_d = 1$), because accumulation accounts for natural order in range $\mathcal{S}$.

Formally: $\boldsymbol{f} = \mathbf{T}\,\boldsymbol{p}$, where $\mathbf{T} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 1 & 0 \end{pmatrix}$.

Dispersion of Q, e. g., by $\mathrm{IOV}(\boldsymbol{f}) = \frac{4}{d} \sum_{i=0}^{d-1} f_i(1 - f_i)$.
(one-point vs. extreme two-point distribution)

Skewness of Q, e. g., by $\mathrm{skew}(\boldsymbol{f}) = \frac{2}{d} \sum_{i=0}^{d-1} f_i - 1$.
(extreme left (right) one-point vs. symmetric distribution)

Examples of IOV and skew:



Descriptive Analysis of Ordinal CoDa

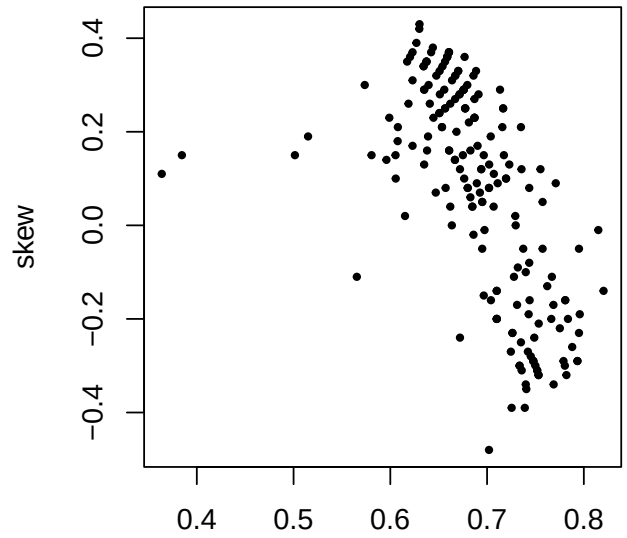
These ordinal measures can be applied to **ordinal CoDa** $\boldsymbol{x} \in \mathbb{S}$:

First, accumulate CoDa vectors to $\boldsymbol{c} := \mathbf{T}\,\boldsymbol{x}$, then evaluate ordinal dispersion and skewness via $\mathrm{IOV}(\boldsymbol{c})$ and $\mathrm{skew}(\boldsymbol{c})$.

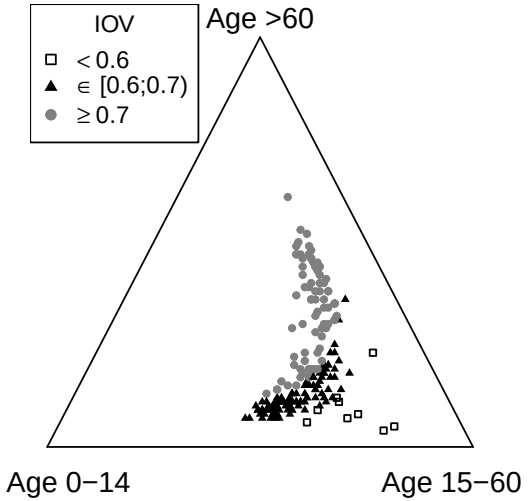
Example: `ageCatWorld` data from R-package `robCompositions`.
Age proportions $\boldsymbol{x}_i = (x_{i,0}, x_{i,1}, x_{i,2})^\top$ in $n = 195$ countries, people with age < 15 ($x_{i,0}$), 15–60 ($x_{i,1}$), and > 60 ($x_{i,2}$).

Location measures not much discriminative power:
median category s_1 in 191 out of 195 cases,
mode category s_1 in 163 out of 195 cases.

IOV-skew diagram and ternary diagram:



(a)



(b)

(c)

IOV-skew diagram easy to interpret. In the present example, similar insights into data as ternary diagram.

But IOV-skew diagram also for higher-dimensional CoDa!

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Ordinal Compositional Data

Statistical Inference

Statistical Inference for Ordinal CoDa

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i. i. d. ordinal CoDa with mean $\boldsymbol{p}$,

then standard CLT implies

$$\sqrt{n}\left(\overline{\boldsymbol{X}} - \boldsymbol{p}\right) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = (\sigma_{ij})_{i,j=0,\ldots,d}, \quad \sigma_{ij} = Cov[X_i, X_j].$$

Accumulate $\boldsymbol{C}_k = \mathbf{T}\,\boldsymbol{X}_k$ with $E[\boldsymbol{C}_k] = \boldsymbol{f} = \mathbf{T}\,\boldsymbol{p}$, then

$$\sqrt{n}\left(\overline{\boldsymbol{C}} - \boldsymbol{f}\right) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}') \quad \text{with } \boldsymbol{\Sigma}' = \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}^\top, \quad \sigma'_{ij} = \sum_{r=0}^{i} \sum_{s=0}^{j} \sigma_{rs}.$$

Asymptotics of IOV($\overline{\boldsymbol{C}}$) and skew($\overline{\boldsymbol{C}}$)

via Taylor expansions ("Delta method") ...

Theorem 1: $\mathrm{IOV}(\overline{\boldsymbol{C}})$ and $\mathrm{skew}(\overline{\boldsymbol{C}})$ asymptotically normally distributed with

$$E\big[\mathrm{IOV}(\overline{\boldsymbol{C}})\big] \;\approx\; \mathrm{IOV}(\boldsymbol{f}) \;-\; \frac{1}{n}\frac{4}{d}\sum_{i=0}^{d-1}\sigma'_{ii},$$

$$V\big[\mathrm{IOV}(\overline{\boldsymbol{C}})\big] \;\approx\; \frac{1}{n}\frac{16}{d^2}\sum_{i,j=0}^{d-1}(1-2f_i)(1-2f_j)\,\sigma'_{ij},$$

and

$$E\big[\mathrm{skew}(\overline{\boldsymbol{C}})\big] = \mathrm{skew}(\boldsymbol{f}), \qquad V\big[\mathrm{skew}(\overline{\boldsymbol{C}})\big] \;\approx\; \frac{1}{n}\frac{4}{d^2}\sum_{i,j=0}^{d-1}\sigma'_{ij}.$$

Example: closed-form formulae for Dirichlet distribution.

Simulation study: good finite-sample performance.

Example: `educFM` data from R-package `robCompositions`.

Proportions of low (s_0), medium (s_1), and high (s_2) education levels of father ($\boldsymbol{x}_{\mathsf{f};i}$) and mother ($\boldsymbol{x}_{\mathsf{m};i}$)
in $n = 31$ European countries.

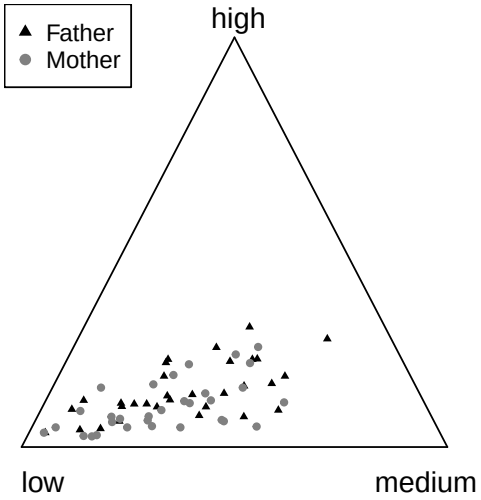
Category s_0 ("low") is median (mode) of $\boldsymbol{x}_{\mathsf{f};i}$ in 19 (22) cases, and of $\boldsymbol{x}_{\mathsf{m};i}$ in 22 (25) cases; otherwise, it is s_1.

Ternary plot: concentration in low-to-medium corner.

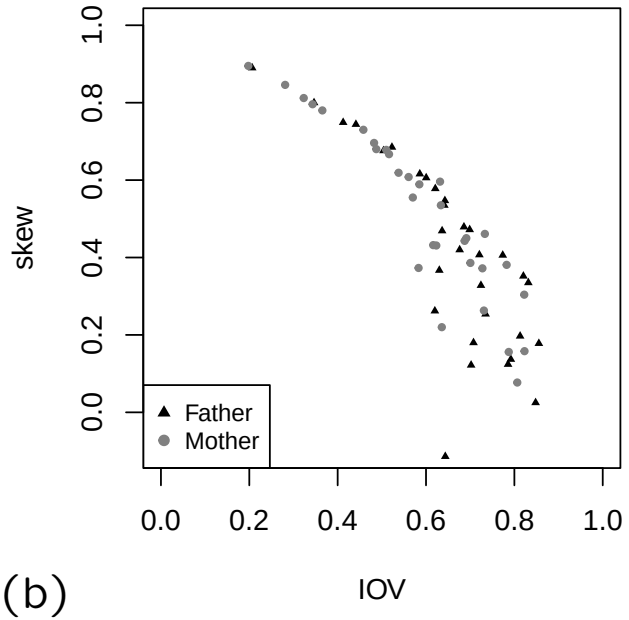
IOV-skew diagram: low dispersion if strong positive skewness.

Hardly any differences between father and mother visible.

Statistical Inference for Ordinal CoDa



(a)



(b)

Father	Estim.	Bias	SE
IOV($\overline{C}$)	0.731	-0.003	0.022
skew($\overline{C}$)	0.414	0.000	0.043
Mother			
IOV($\overline{C}$)	0.652	-0.002	0.033
skew($\overline{C}$)	0.516	0.000	0.041

(c)

Fit logistic normal distributions (Aitchison, 1986) to CoDa, approximate bias and standard error (SE):

Significantly (5%-level) less dispersion and stronger positive skewness for mothers $\Rightarrow$ unequal opportunities for education.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Monitoring Ordinal Compositional Data

Control Charts

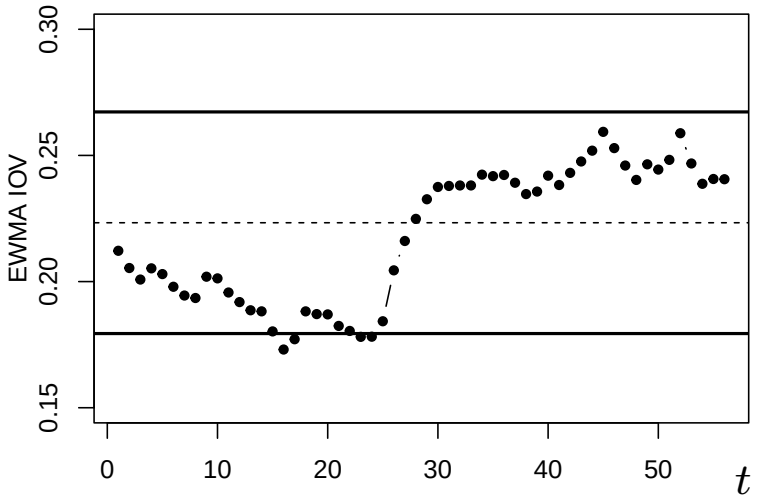
Control Charts for Ordinal CoDa **(Sketch)**

Default control chart for i. i. d. CoDa process: χ^2**-chart**

$$\mathrm{X}_t^2 \;=\; (\boldsymbol{Y}_t - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{Y}_t - \boldsymbol{\mu}_0) \quad \text{for } t = 1, 2, \ldots$$

Proposal: Complement by **IOV- and skew-chart**,
preferably with additonal EWMA smoothing, i. e.,
$\boldsymbol{C}_{t,\lambda} = \lambda\, \boldsymbol{C}_t + (1-\lambda)\, \boldsymbol{C}_{t-1,\lambda}$ with $\boldsymbol{C}_{0,\lambda} = \boldsymbol{f}_0$ and $\lambda \in (0; 1)$.

Plot IOV($\boldsymbol{C}_{t,\lambda}$)
or skew($\boldsymbol{C}_{t,\lambda}$)
for $t = 1, 2, \ldots$
against control limits:



HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Ordinal Compositional Time Series

Stochastic Modelling

Ordinal Compositional Time Series

Existing models for CoTS do not account for ordering.

Example: conditional Dirichlet model (Zheng & Chen, 2017), $\boldsymbol{X}_t \,|\, \mathcal{F}_{t-1} \sim \text{Dir}(\boldsymbol{p}_t, \nu)$ with $E[\boldsymbol{X}_t \,|\, \mathcal{F}_{t-1}] = \boldsymbol{p}_t$ and $\nu > 0$. "ARMA recursion" for $\boldsymbol{p}_t = \phi(\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-\text{p}}, \ \boldsymbol{p}_{t-1}, \ldots, \boldsymbol{p}_{t-\text{q}})$, with AR-order p and feedback terms $\boldsymbol{p}_{t-1}, \ldots, \boldsymbol{p}_{t-\text{q}}$.

Idea: To account for natural order, combine with cumulative logit approach in Pruscha (1993), Fokianos & Kedem (2003).

Proposal: Let $F_{0,1}(x) = [1 + \exp(-x)]^{-1}$ be inverse logit. Define $\tilde{\boldsymbol{X}}_t = (X_{t,0}, \ldots, X_{t,d-1})^\top$ and $\tilde{\boldsymbol{p}}_t = (p_{t,0}, \ldots, p_{t,d-1})^\top$, let $\boldsymbol{f}_t = (f_{t,0}, \ldots, f_{t,d-1})^\top$ be conditional CDF vector.

Full ARMA-type model for ordinal CoTS by

$$f_{t,i} \;=\; F_{0,1}\Big(\eta_i + \Sigma_{k=1}^{\mathsf{p}}\, \boldsymbol{\alpha}_k^\top\, \tilde{\boldsymbol{X}}_{t-k} + \Sigma_{l=1}^{\mathsf{q}}\, \boldsymbol{\beta}_l^\top\, \tilde{\boldsymbol{p}}_{t-l}\Big) \quad \text{for } i = 0, \ldots, d-1,$$

with $d(\mathsf{p}+\mathsf{q}+1)$ parameters and $-\infty < \eta_0 < \ldots < \eta_{d-1} < +\infty$.

More parsimonious model with $d + \mathsf{p} + \mathsf{q}$ parameters:

$$f_{t,i} \;=\; F_{0,1}\Big(\eta_i + \Sigma_{k=1}^{\mathsf{p}}\, \alpha_k\, \mathbf{1}^\top \tilde{\boldsymbol{X}}_{t-k} + \Sigma_{l=1}^{\mathsf{q}}\, \beta_l\, \mathbf{1}^\top \tilde{\boldsymbol{p}}_{t-l}\Big).$$

Include possible covariate $\boldsymbol{z}_t$ by summand "$+\boldsymbol{\gamma}^\top \boldsymbol{z}_t$".

Dirichlet model with additional scale parameter ν:

$\boldsymbol{X}_t \,|\, \mathcal{F}_{t-1} \;\sim\; \mathrm{Dir}(\boldsymbol{p}_t, \nu)$, where $\boldsymbol{p}_t$ from $\boldsymbol{f}_t$ by discrete differences.

Example: Yearly proportions of categories "not overweight" (BMI < 25), "overweight" (BMI in $[25; 30)$), and "obese" (BMI ≥ 30) in Germany for period 1975–2016 (data from WHO).

Partition: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{36}$ (1975–2010) for model fitting, $\boldsymbol{x}_{37}, \ldots, \boldsymbol{x}_{42}$ (2011–2016) for out-of-sample forecasting.

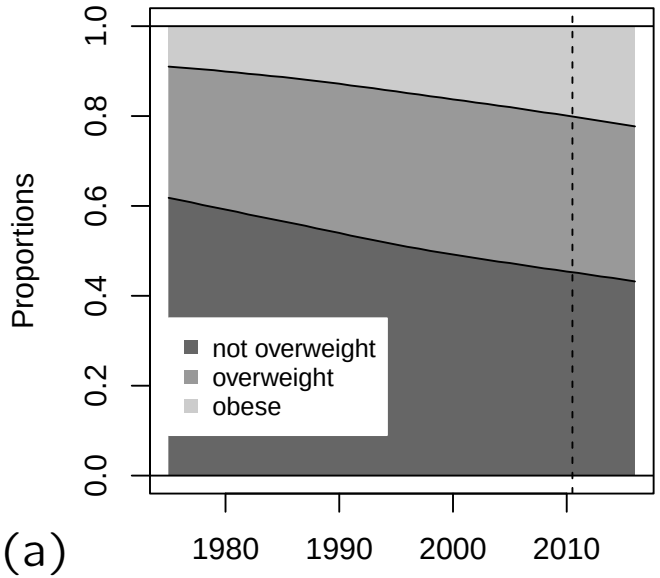
For similar data set, Mills (2010) proposed **linear alr-model**:

$$(1) \qquad E\big[\mathrm{alr}(\boldsymbol{x}_t)_i\big] \;=\; a_i + b_i\, t \quad \text{for } i = 1, \ldots, d.$$

$$(2) \qquad f_{t,i} \;=\; F_{0,1}(\eta_i + \gamma_i\, t) \quad \text{for } i = 0, \ldots, d-1.$$

$$(3) \qquad f_{t,i} \;=\; F_{0,1}(\eta_i + \gamma_i\, t + \boldsymbol{\alpha}_1^\top\, \tilde{\boldsymbol{X}}_{t-1}) \quad \text{for } i = 0, \ldots, d-1.$$

Plot of proportions and table with model fits and CSS values:



(a)

Model	i	η_{i-1} or a_i	γ_{i-1} or b_i	$\alpha_{1,i-1}$	CSS_{in}	CSS_{out}
(1)	1	1.957	-0.032		1.135	2.119
	2	1.230	-0.018			
(2)	1	0.480	-0.019		0.698	1.269
	2	2.350	-0.027			
(3)	1	0.469	-0.016	0.322	0.410	0.379
	2	2.354	-0.025	-0.651		

(b)

Novel (logistic-)linear model, especially with AR(1)-component, clearly superior in conditional sum of squares (CSS, times 10^3):

$$\text{CSS}(\boldsymbol{\theta}) := \sum_t \|\boldsymbol{X}_t - \boldsymbol{p}_t\|^2.$$

Conclusions

- Beneficial to consider natural order of ordinal CoDa:
  - (visual) descriptive analysis of ordinal CoDa,
  - statistical inference from i. i. d. CoDa samples,
  - EWMA control charts for i. i. d. CoDa processes,
  - conditional regression models for ordinal CoTS.
- **Future research:** Among others, . . .
  - performance of ordinal CoDa control charts for *serially dependent* CoDa (i. e., monitoring of CoTS);
  - *analysis* of ordinal CoTS by adapting methods from ordinal time series analysis (Weiß, 2020).

Thank You for Your Interest!

MATH
STAT

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

`weissc@hsu-hh.de`

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Projektnummer 516522977.

Literature

Weiß (2023) Ordinal compositional data and time series. *Statistical Modelling*, in press.

Agresti (2002) *Categorical Data Analysis*. Wiley.

Agresti (2010) *Analysis of Ordinal Categorical Data*. Wiley.

Aitchison (1986) *Statistical Analysis of Compositional Data*. Chapman.

Filzmoser et al. (2018) *Applied Compositional Data Analysis*. Springer.

Fokianos & Kedem (2003) Regression theory for . . . *Stat Sci* **18**, 357–376.

Mills (2010) Forecasting compositional . . . *Qual & Quant* **44**, 673–690.

Pawlowsky-Glahn & Buccianti (2011) *Compositional Data Analysis*. Wiley.

Pruscha (1993) Categorical time series . . . *Statistics* **24**, 43–57.

Vives-Mestres et al. (2014) Individual T^2 control . . . *JQT* **46**, 127–139.

Weiß (2020) Distance-based analysis of . . . *JASA* **115**, 1189–1200.

Zheng & Chen (2017) Dirichlet ARMA models . . . *JMA* **158**, 31–46.