Functional data analysis in Bayes Spaces

with an application to spatio-temporal population data

Matthias Templ

Institute of Data Analysis and Process Design
Zurich University of Applied Sciences, Switzerland

Keynote, UROS, 3rd Dec. 2020

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse
und Prozessdesign

Overview

Compositional data analysis (CoDa) is not (well-)known in official statistics and survey statistics, but we need some concepts of CoDa to motivate our topic today.

Overview

Compositional data analysis (CoDa) is not (well-)known in official statistics and survey statistics, but we need some concepts of CoDa to motivate our topic today.

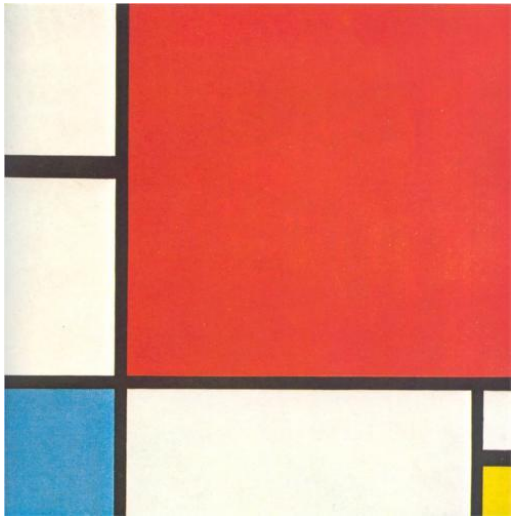
Part I (Compositional Data)

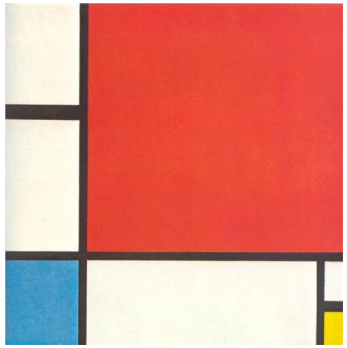
1. What are compositional data?
2. Examples of compositional data
3. Log-ratio analysis

Part II (Functional Data Analysis)

4. Application to probability density functions and functional data analysis

Piet Mondrian: composition with red, blue and yellow

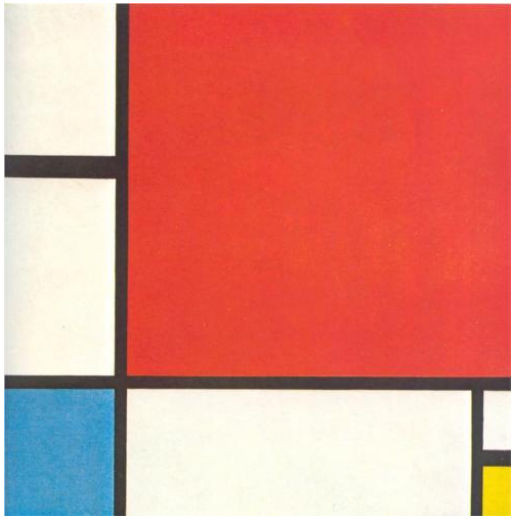


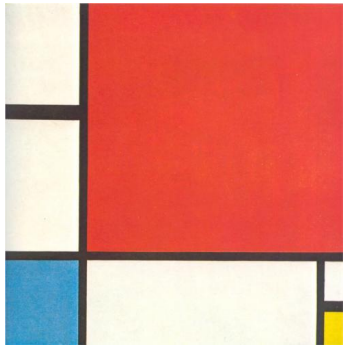


Size of each area (simplified, without location of areas and black lines): $\mathbf{x} = (x_1, ..., x_7)$

```
x <- c(1/15, 1/13, (1/15+1/13)*4, 1/19, 1/19*3.59898,
       1/50, 1/50) *100 # sum(x) == 100
```

Piet Mondrian: composition with red, blue and yellow





Size of each area (simplified, without location of areas and black lines): $\mathbf{x} = (x_1, ..., x_7)$

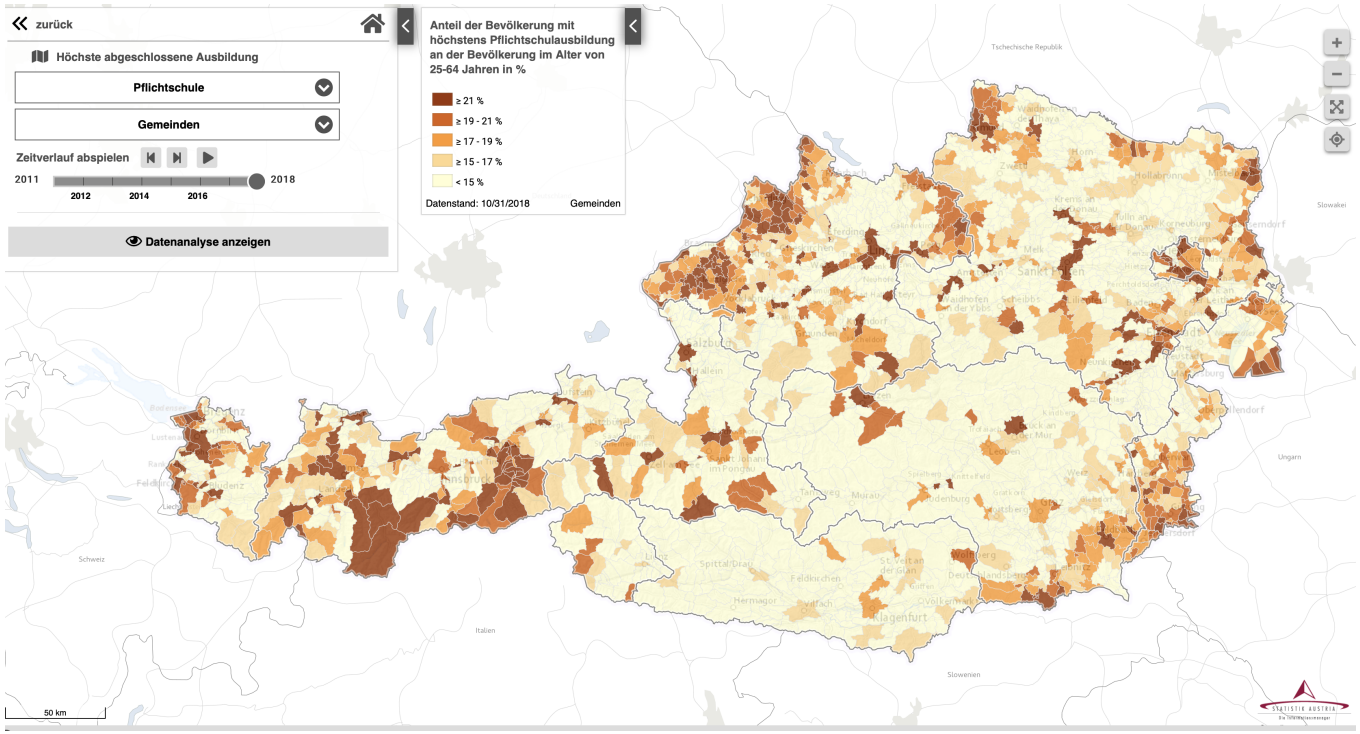
```
x <- c(1/15, 1/13, (1/15+1/13)*4, 1/19, 1/19*3.59898,
       1/50, 1/50) *100 # sum(x) == 100
```

- Absolute values are not important
- Ratios are important to express the *harmony of life* / the composition
- To describe the picture, either took all ratios, or weight each area with a mean of other/all areas

Compositional data (CoDa), OLD and outdated viewpoint

- observations are *closed* by a fixed constant (e.g. 1 or 100)

Example: https://www.statistik.at/atlas/



Modern view on CoDa (Filzmoser, Hron, and Templ 2018)

Each composition $\mathbf{x}$ represents a random vector consisting of strictly positive components in the D-part simplex space

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D : x_i > 0, i = 1, \ldots, D, \sum_{i=1}^{D} x_i = \kappa \right\}$$

Modern view on CoDa (Filzmoser, Hron, and Templ 2018)

Each composition $\mathbf{x}$ represents a random vector consisting of strictly positive components in the D-part simplex space

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D : x_i > 0, i = 1, \ldots, D, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- quantitative representation of **parts of a certain whole**, whereby the information on the whole is irrelevant $\rightarrow \kappa$ is a (fixed) constant, but be different for each composition in a data set

Modern view on CoDa (Filzmoser, Hron, and Templ 2018)

Each composition **x** represents a random vector consisting of strictly positive components in the D-part simplex space

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D : x_i > 0, i = 1, \ldots, D, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- quantitative representation of **parts of a certain whole**, whereby the information on the whole is irrelevant $\rightarrow \kappa$ is a (fixed) constant, but be different for each composition in a data set
- multivariate data whenever the analysis of **relative information** is important

Modern view on CoDa (Filzmoser, Hron, and Templ 2018)

Each composition **x** represents a random vector consisting of strictly positive components in the D-part simplex space

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D : x_i > 0, i = 1, \ldots, D, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- quantitative representation of **parts of a certain whole**, whereby the information on the whole is irrelevant $\rightarrow \kappa$ is a (fixed) constant, but be different for each composition in a data set
- multivariate data whenever the analysis of **relative information** is important
- units e.g. ppm, mg/kg, mg/l, EUR, Dollar, hours of a day, ...

Modern view on CoDa (Filzmoser, Hron, and Templ 2018)

Each composition **x** represents a random vector consisting of strictly positive components in the D-part simplex space

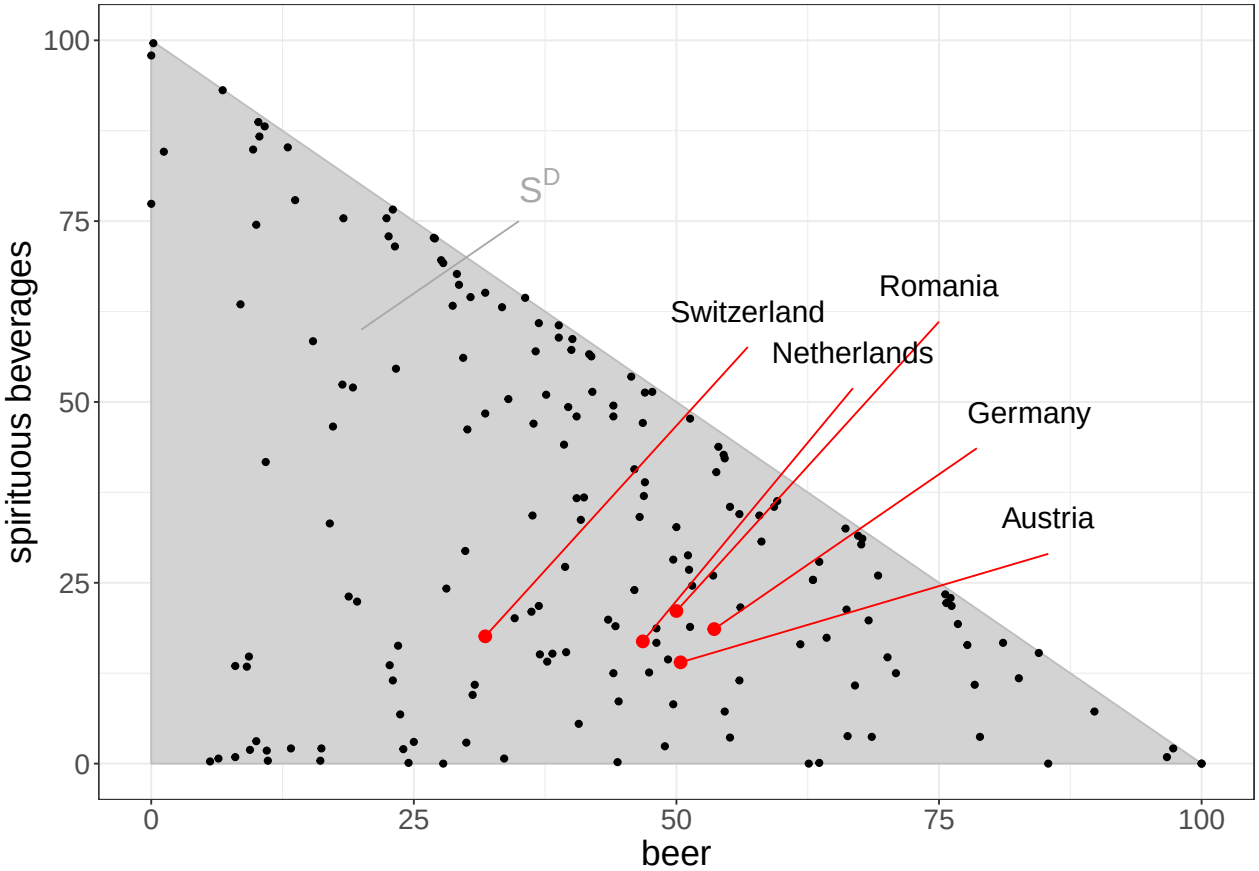
$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D : x_i > 0, i = 1, \ldots, D, \sum_{i=1}^{D} x_i = \kappa \right\}$$

- quantitative representation of **parts of a certain whole**, whereby the information on the whole is irrelevant $\rightarrow \kappa$ is a (fixed) constant, but be different for each composition in a data set
- multivariate data whenever the analysis of **relative information** is important
- units e.g. ppm, mg/kg, mg/l, EUR, Dollar, hours of a day, ...

Meaning of a simplex, example

```
alcohol %>% filter(country == "Romania")
```

```
##    country  year beer wine spirits other
## 1 Romania  2010   50 28.9    21.1     0
```

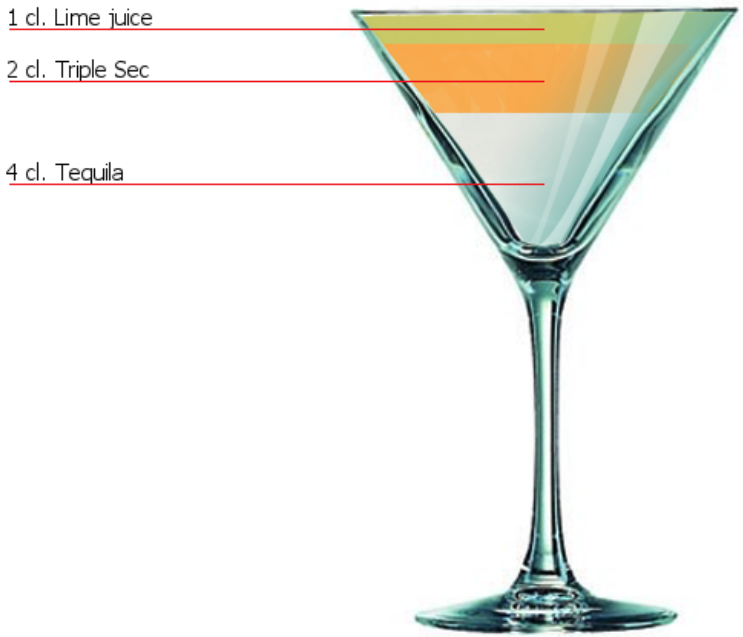


- non-Euclidean geometry. Correlation is forced to be negative.

Modern view on CoDa (Filzmoser, Hron, and Templ 2018)

Compositional data:

- follows an specific geometry (non-Euclidean, but Aitchison geometry)
- log-ratio presentation of data (defined later)
- some requirements (defined later)



The Margaritha-*data* can be a composition (when the flavour of a Margaritha is important) or non-compositional (absolute amount of Tequila is important to get drunk)

Examples of Compositional Data

- contingency, probability and compositional tables
- mortality data
- basket data
- hours of a day
- percentages or ratios of a whole
- distribution of species in certain areas
- tax, wage or expenditure data
- poll data
- sequencing data in genetics
- chemical components
- or fixed row-sums in official statistics (percentages, ratios, . . .)
- . . .

and even **probability density functions** ($\int_{-\infty}^{\infty} f(x)\,dx = \kappa \quad (= 1)$) are compositions

Examples of Compositional Data

Many data sets on

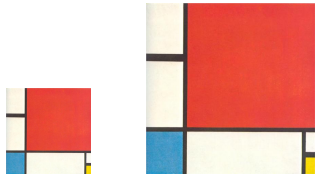
https://ec.europa.eu/eurostat/home

- Agriculture Lucas land coverage
- Covid19 tourism, number of trips
- Intra-EU28 trade, by Member State, total product
- Population counts as a percentage of EU27
- ...

Requirements of a concise methodology

Scale invarianz

Same results obtained when a composition is multiplied by a constant

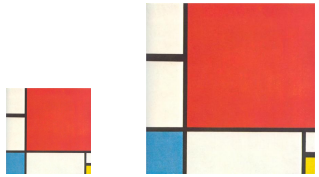


$$\mathbf{x} \propto \mathbf{y} \iff \mathcal{C}(\mathbf{y}) = \mathcal{C}(\mathbf{x})$$

Requirements of a concise methodology

Scale invarianz

Same results obtained when a composition is multiplied by a constant



$$\mathbf{x} \propto \mathbf{y} \iff \mathcal{C}(\mathbf{y}) = \mathcal{C}(\mathbf{x})$$

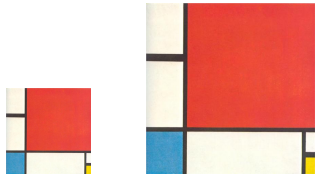
Permutation invariance

Reordering of parts (variables) leads to the same results

Requirements of a concise methodology

Scale invarianz

Same results obtained when a composition is multiplied by a constant



$$\mathbf{x} \propto \mathbf{y} \iff \mathcal{C}(\mathbf{y}) = \mathcal{C}(\mathbf{x})$$

Permutation invariance

Reordering of parts (variables) leads to the same results

Subcompositional coherence

Dominance: the distance between two compositions should be larger than the distance between a subcomposition of these two compositions.

$$\Delta_D(\mathbf{x}, \mathbf{y}) \geq \Delta_d(\mathbf{x}_d, \mathbf{y}_d) \text{ mit } d < D$$

Ratio preserving: non-selected variables have no influence on results.

Scale invariance and subcompositional coherence example

A	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
Absolute	1	1710	950	570	570	3800
information	2	540	300	180	180	1200
in EUR	3	900	500	300	700	2400

Scale invariance and subcompositional coherence example

A	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
Absolute	1	1710	950	570	570	3800
information	2	540	300	180	180	1200
in EUR	3	900	500	300	700	2400

B	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
in CHF	1	1846	1026	615	615	4102
in Dollar	2	583	324	194	194	1295
in CHF	3	972	540	324	756	2592

Scale invariance and subcompositional coherence example

A	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
Absolute	1	1710	950	570	570	3800
information	2	540	300	180	180	1200
in EUR	3	900	500	300	700	2400

B	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
in CHF	1	1846	1026	615	615	4102
in Dollar	2	583	324	194	194	1295
in CHF	3	972	540	324	756	2592

C	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
Information	1	45	25	15	15	100
expressed	2	45	25	15	15	100
in %	3	37.5	20.1	12.5	29.2	100

Scale invariance and subcompositional coherence example

A	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
Absolute	1	1710	950	570	570	3800
information	2	540	300	180	180	1200
in EUR	3	900	500	300	700	2400

B	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
in CHF	1	1846	1026	615	615	4102
in Dollar	2	583	324	194	194	1295
in CHF	3	972	540	324	756	2592

C	Obs.	*housing*	*foodstuff*	*transport*	*communications*	Sum
Information	1	45	25	15	15	100
expressed	2	45	25	15	15	100
in %	3	37.5	20.1	12.5	29.2	100

D	Obs.	*housing*	*foodstuff*	*transport*	Sum
Information	1	52.9	29.4	17.6	100
expressed	2	52.9	29.4	17.6	100
in %	3	52.9	29.4	17.6	100

- a compositional analysis gives the same results for A, B, C, and they would not be in contradiction to D

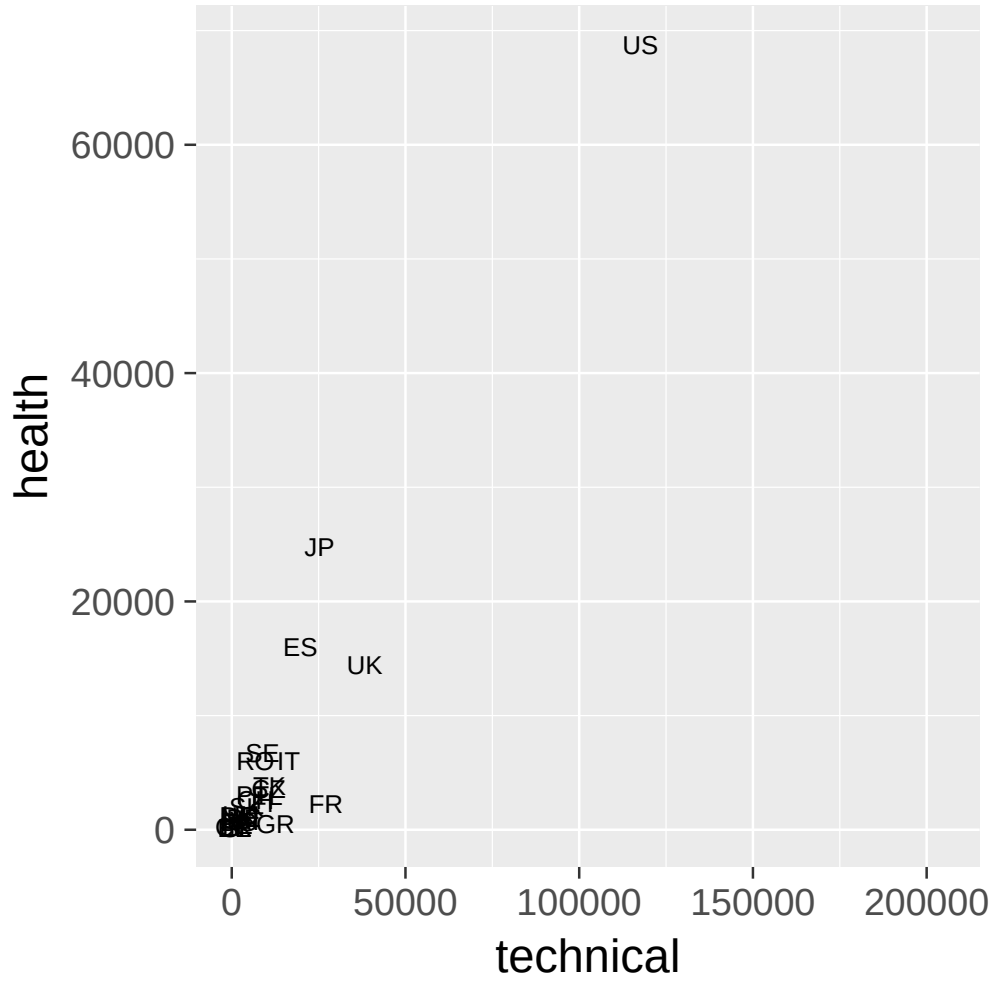
Classical statistics leads to arbitrary results

E.g correlations are always forced to be negative because of the simplex sample space.

Example PhD students in some countries

	total	male	female	technical	soc-eco-law	human	health	agriculture
BE	7500	59.0	41.0	3462	1469	997	1041	532
BG	5200	49.7	50.3	2064	1102	1170	666	198
CZ	22600	62.1	37.9	10668	3748	3518	3633	1035
DK	4800	54.2	45.8	1886	614	696	1210	394
EE	2000	46.5	53.5	847	424	420	196	112
IE	5100	52.1	47.9	2633	787	1124	450	107
GR	22500	55.6	44.4	12590	3941	5090	495	383
ES	77100	49.0	51.0	19751	20704	18885	16026	1733
FR	69800	53.9	46.1	27152	21429	18846	2303	70
IT	38300	48.3	51.7	16403	7621	5803	6035	2437
LV	1800	39.6	60.4	542	603	434	182	40
LT	2900	43.4	56.6	1183	916	400	293	107
HU	8000	53.0	47.0	2576	1648	1992	1304	480
AT	16800	54.3	45.7	4978	6374	4103	790	555
PL	32700	50.7	49.3	10202	7881	9974	3008	1635
PT	20500	44.0	56.0	6027	6191	4879	3034	369
RO	21700	51.7	48.3	6864	3801	3323	6017	1694
SI	1100	53.5	46.5	526	174	189	168	43
SK	10700	57.1	42.9	4220	2121	1971	2024	364
FI	22100	48.4	51.6	8875	4990	5365	2406	464
SE	21400	51.3	48.7	8872	2651	2694	6756	428
UK	94200	55.4	44.6	38266	19747	20408	14456	1323
CR	1300	53.3	46.7	601	94	286	235	84
TK	32600	60.6	39.4	10888	7922	7335	3814	2641

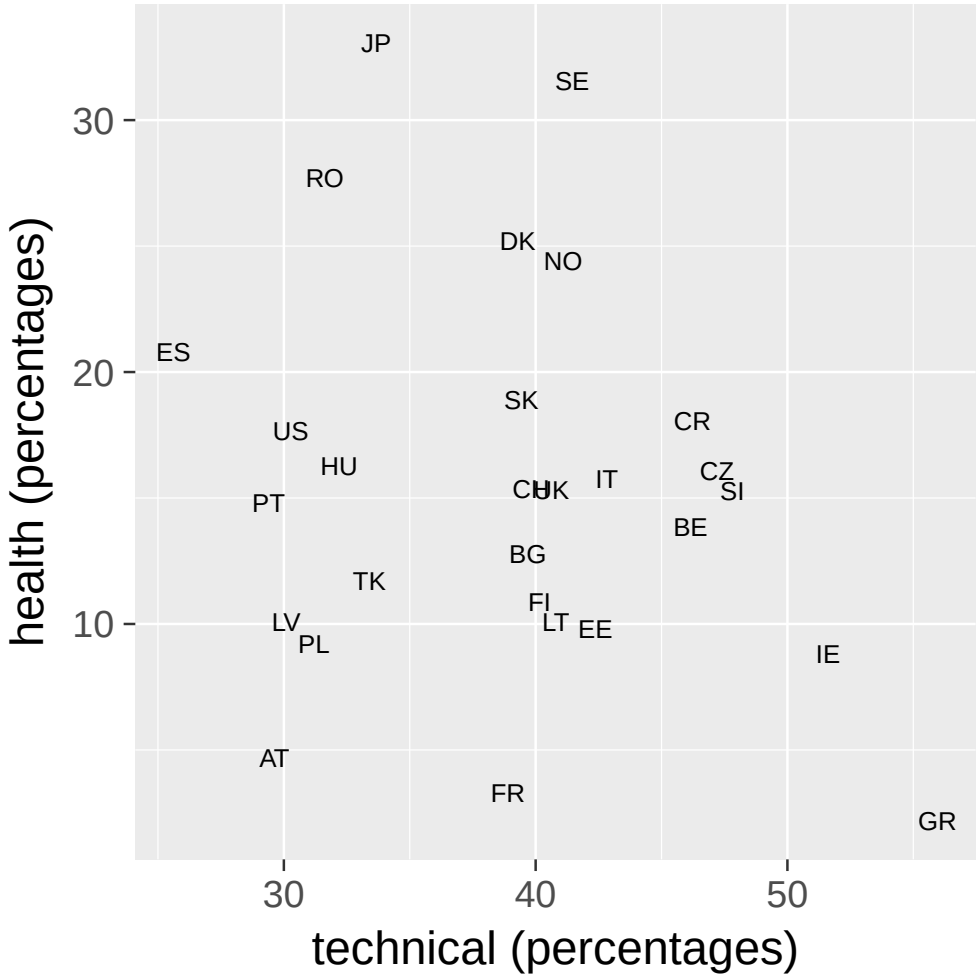
PhD students: absolute values



- not very informative, because correlation is just driven from large countries. The relative information would be more informative (ratios on technical, health, . . .).

PhD students (in %)

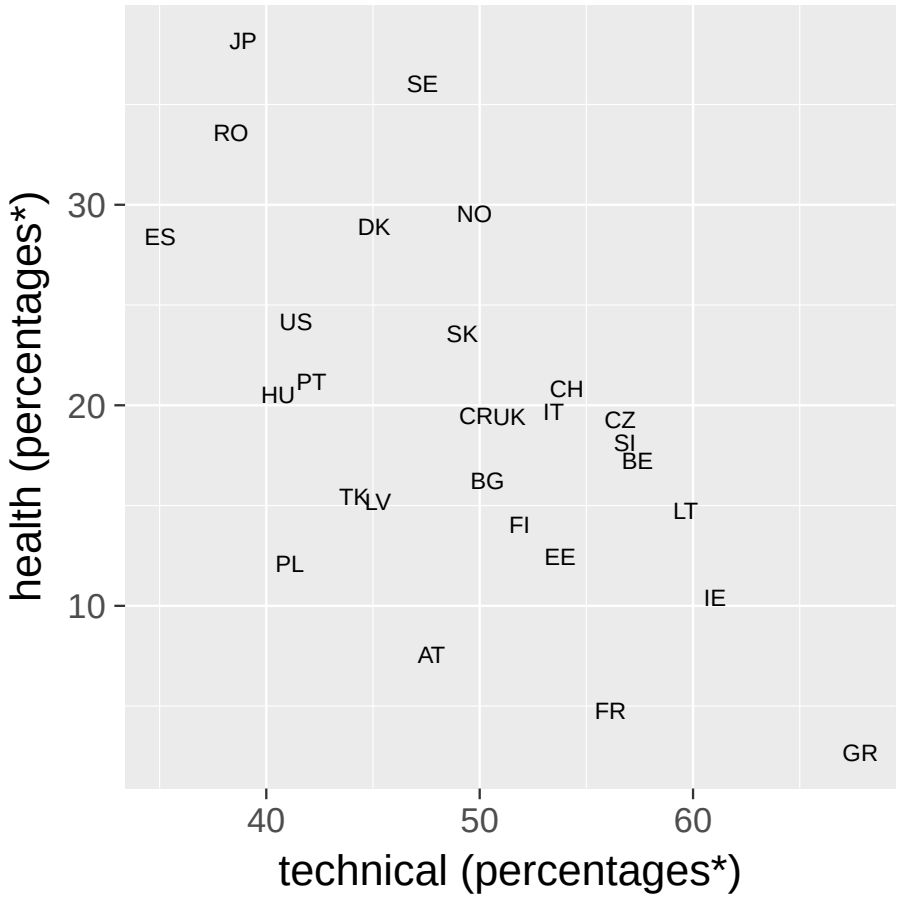
Therefore often data are constraint/normalized to 1 or 100.



Correlation is approx. zero.

PhD students: percentages built **WITHOUT** socio-econ. and law students

If we do not measure doctoral students in socioeconomics and law, the correlation changes (now negative). Reason: closure effect. Trad. correlation measures leads to arbitrary results.



From S^D to R^D

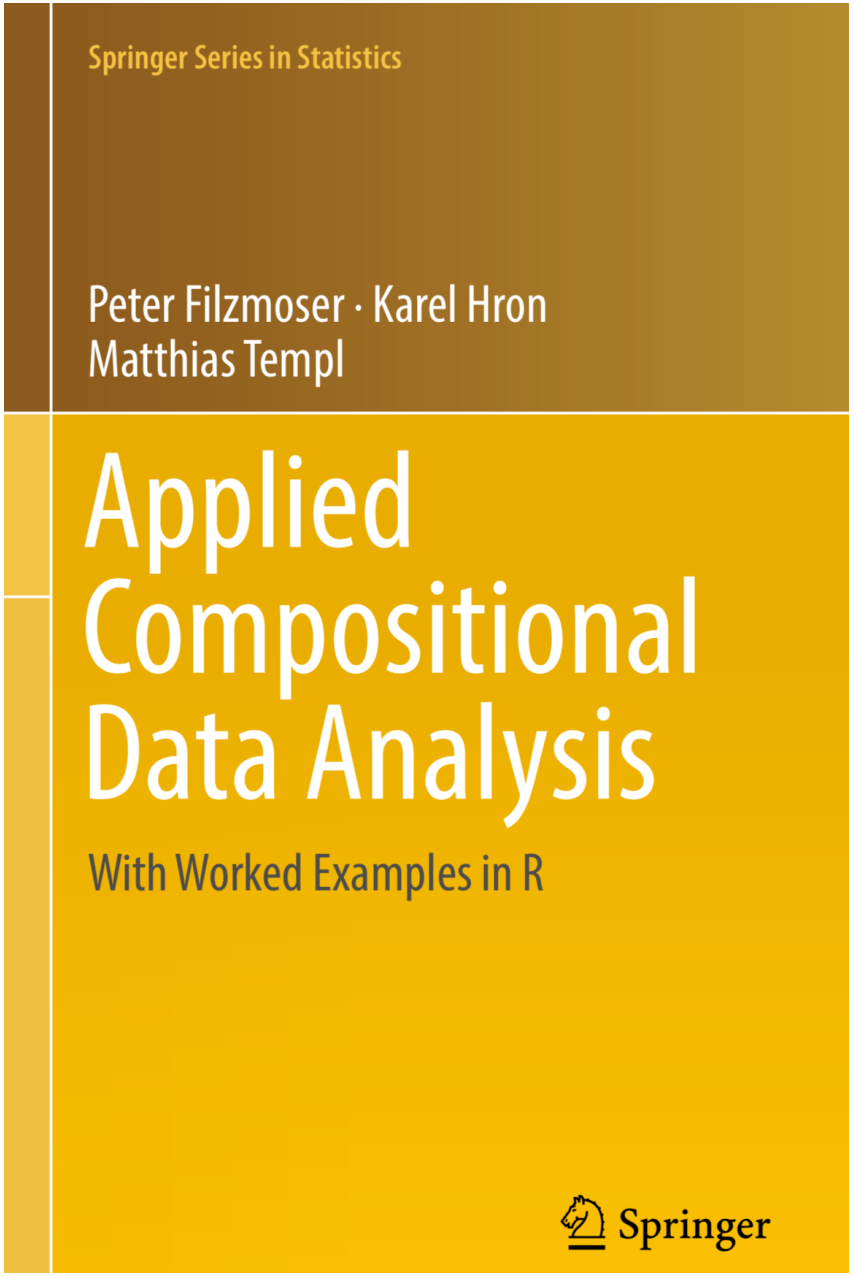
Centred log-ratio coordinates: (the simplest choice)

Divide each values of a composition $\mathbf{x} = (x_1, x_2, ..., x_D) \in \mathcal{S}^D$ by the geometric mean and take the log:

$$(\mathbb{R}^D \ni)\, \mathbf{y} = (y_1, \ldots, y_D)^t = \left(\log \frac{x_1}{\sqrt[D]{\prod_{i=1}^D x_i}}, \ldots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^D x_i}} \right)^t$$

Advantage: „Symmetric“.
Disadvantage: Singularity. Often better interpretation of results with *isometric log-ratio representations* and well-chosen *balances*.

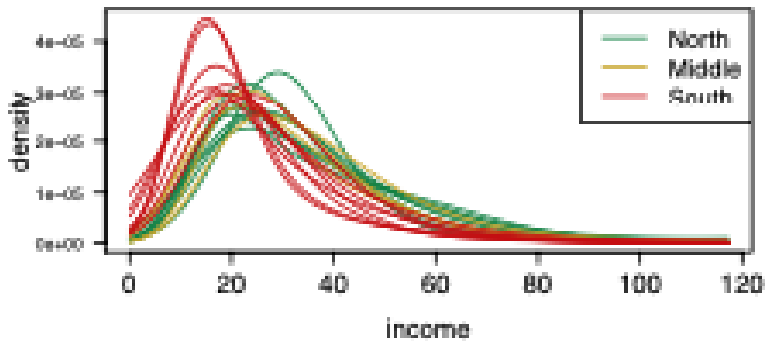
R Package robCompositions + book (2018):



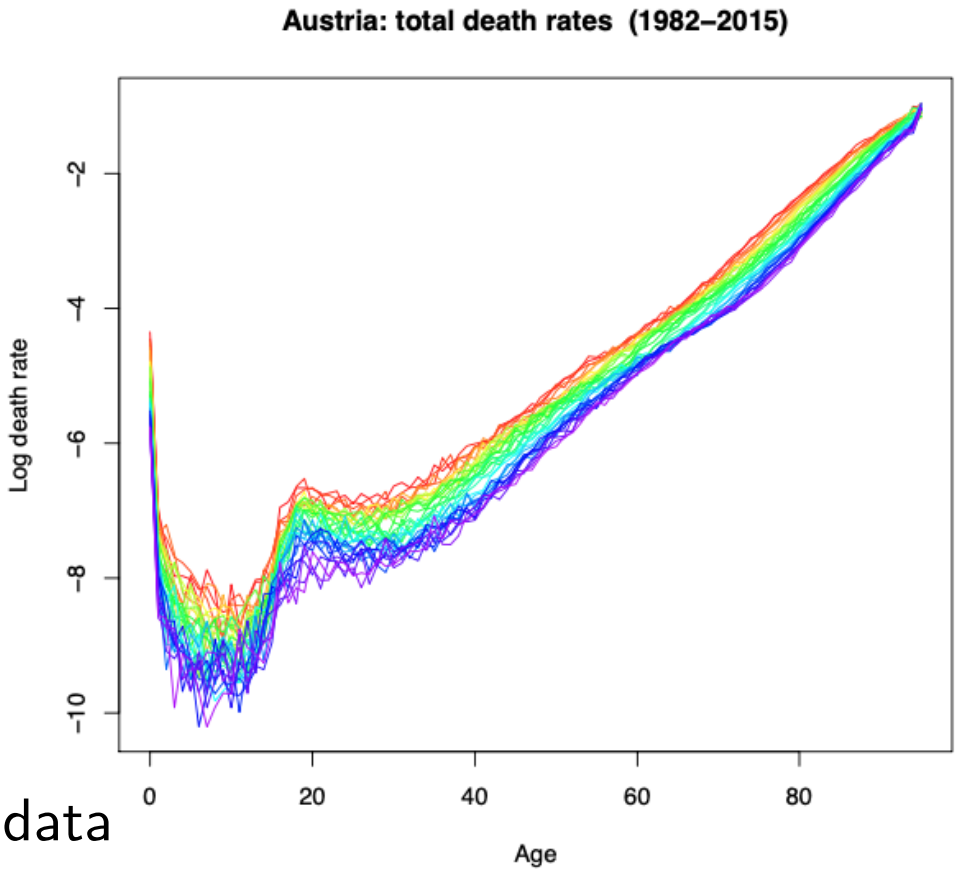
Part II: Functional Data

Example of compositional functional data

- Income of persons in regions



Example of compositional functional data



- Mortality data
- Mortality by age and cause
- Population counts for each age (class) in each municipality . . .
- spatio-temporial bike rentals per day and hour
- children's growth over time
- . . .

Functional data analysis (FDA)

- Samples whereby *each data point lays on a (random) curve*.
- X_i is the i-th observation at time t, **continuous** in $\mathcal{I} = [a, b]$
  - This functional notation **is conceptional**, because in practice $X_i(t)$ are only observed discretely on points $t_1, t_2, ..., t_n$
  - First step: from discrete data $\rightarrow$ *curves*.

Functional data analysis (FDA)

- Samples whereby *each data point lays on a (random) curve*.
- X_i is the i-th observation at time t, **continuous** in $\mathcal{I} = [a, b]$
  - This functional notation **is conceptional**, because in practice $X_i(t)$ are only observed discretely on points $t_1, t_2, ..., t_n$
  - First step: from discrete data $\rightarrow$ *curves*.

Most popular application in FDA:

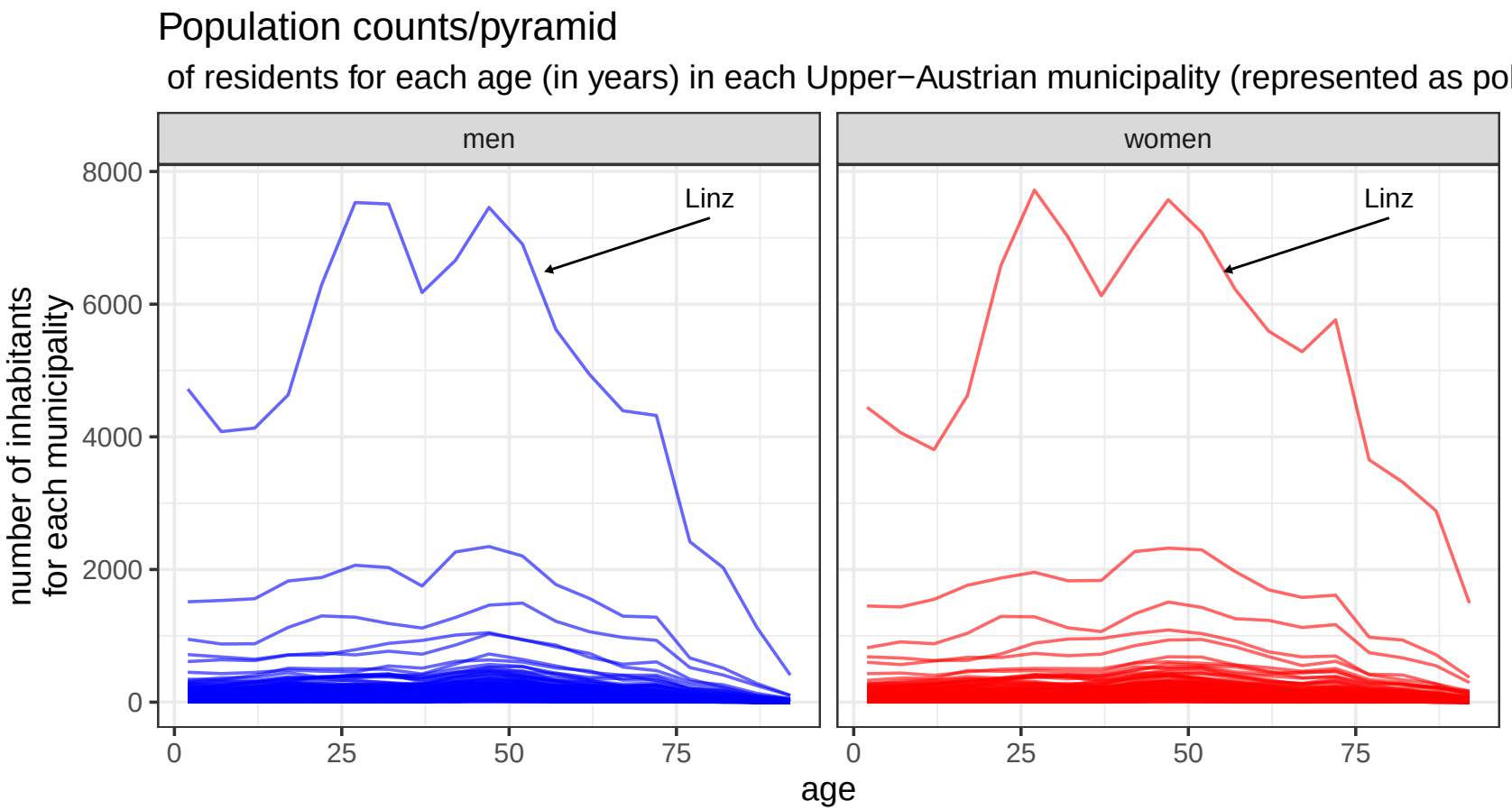
- **functional principal component analysis** (FPCA) for dimension reduction and analysis/interpretation in lower dimensions.
- input are usually probability density functions.
- We will see: **probability density functions are compositional**
- probability density functions are used almost everywhere

Motivating data set from Upper Austria

- Population data in 57 Municipalities $\times$ 19 age groups $\times$ gender

Motivated Guy from Linz-Land

Population pyramids for each municipality (raw data)



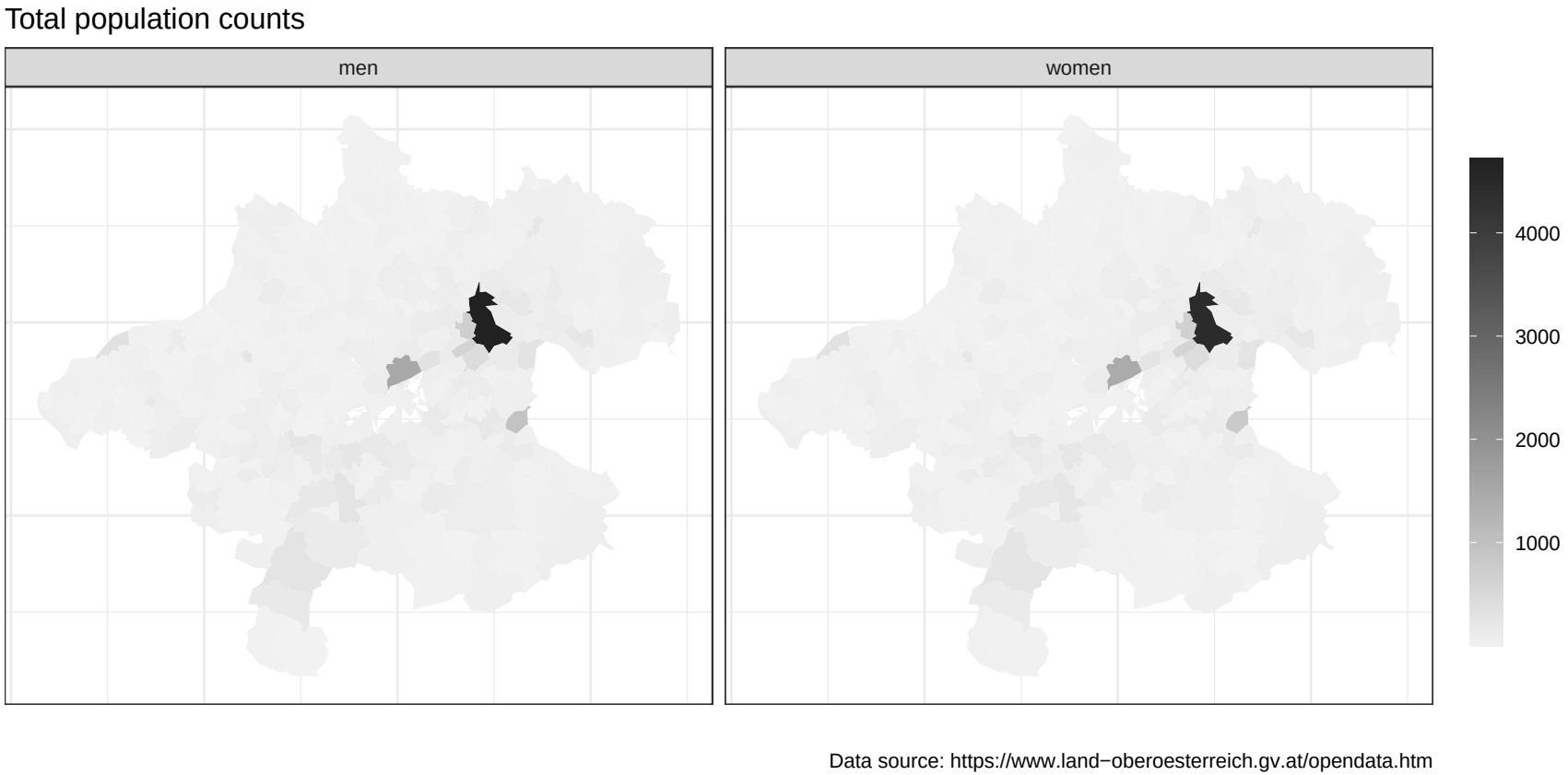
Data source: https://www.land−oberoesterreich.gv.at/opendata.htm

Indicators, e.g. total population counts

Total population counts

Data source: https://www.land−oberoesterreich.gv.at/opendata.htm

Indicators, e.g. total population counts



- such maps could be plotted for each age (group) as well
  - however, we would produce many maps (not easy to compare)

Indicators, e.g. total population counts

- such maps could be plotted for each age (group) as well
  - however, we would produce many maps (not easy to compare)
- or one map displaying the value of an indicator like the ratio of young to old people
  - however, one indicator does not tell the whole story

Ratios

Aim:

comparison of the **whole distrubtion** on age between all regions and gender

Ratios

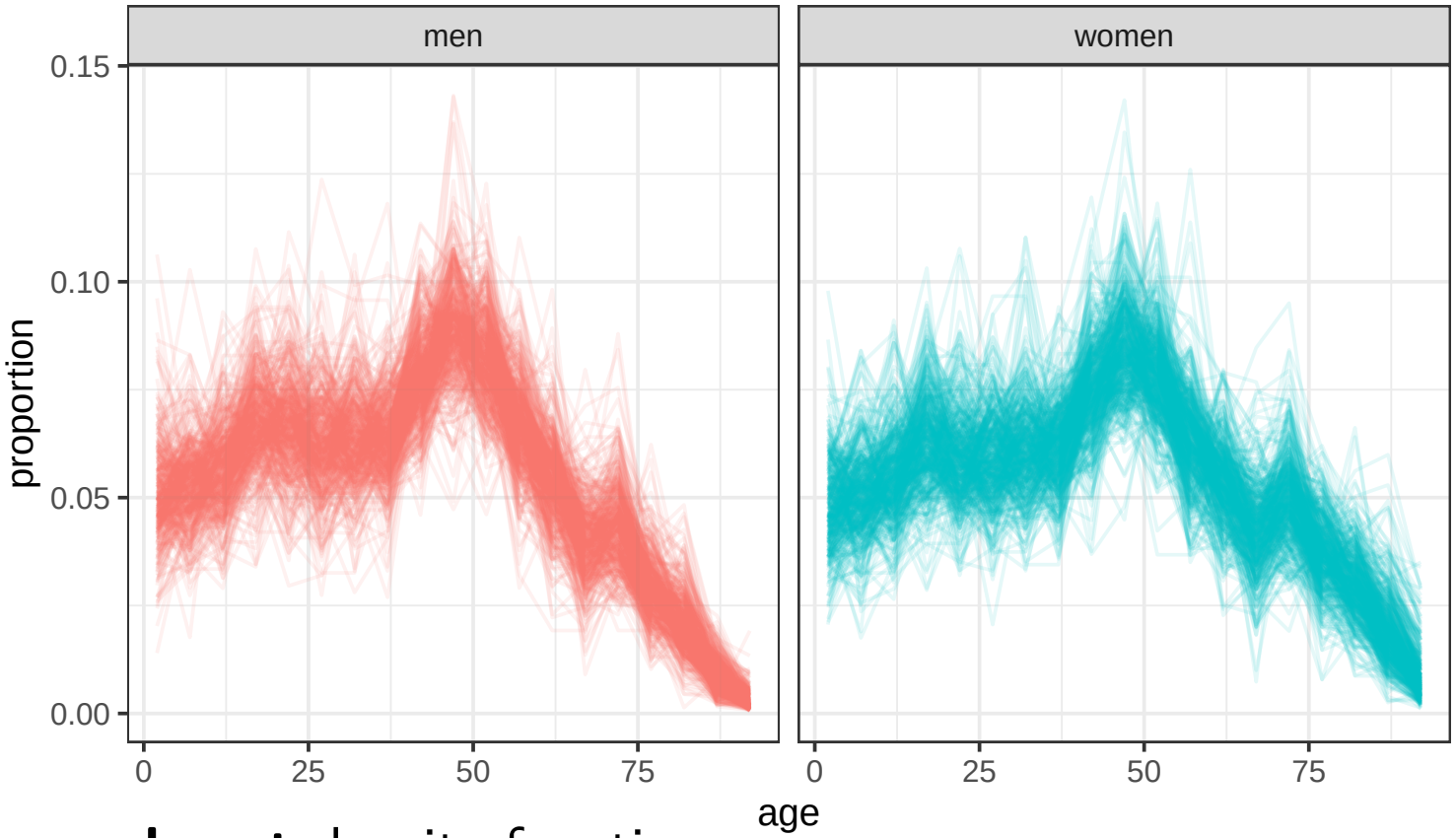
Aim:

comparison of the **whole distrubtion** on age between all regions and gender

First idea:

- Use ratios (**relative information**) between the counts in any municipality
- Absolute values were dominating (Linz, . . .), therefore we calculate counts per age (class) in a municipality divided by the total population count of the muncipality
- we are thus interested in ratios
  - thus questions like: will Linz have relatively a higher amount of young people than Gmunden, etc. . . .
- The important information is in the ratios

Ratios based on the total



- . . . **almost** density functions
- highdimensional data but still discrete information
- we still don't know how to compare these ratios

Goal: Analysing the densities of the age distribution after dimension reduction through principal component analysis (PCA)

Problems related to density functions

- densities integrates to 1 (even this is irrelevant for a compositional analysis)
- densities consists of **only relative information**

Problems related to density functions

- densities integrates to 1 (even this is irrelevant for a compositional analysis)
- densities consists of **only relative information**

Standard methods for FDA have serious flaws

- If we summate two densities (convolution of two densities), the result is not a density (integral is not 1, $\int_{-\infty}^{\infty} f \cdot g \, dx \neq 1$)
- If we multipy a density, the result is not a density (integral is not 1, $\int_{-\infty}^{\infty} f + g \, dx \neq 1$)

Problems related to density functions

- densities integrates to 1 (even this is irrelevant for a compositional analysis)
- densities consists of **only relative information**

Standard methods for FDA have serious flaws

- If we summate two densities (convolution of two densities), the result is not a density (integral is not 1, $\int_{-\infty}^{\infty} f \cdot g \, dx \neq 1$)
- If we multipy a density, the result is not a density (integral is not 1, $\int_{-\infty}^{\infty} f + g \, dx \neq 1$)
- FPCA: results can even include negative estimates of density values

Problems related to density functions

- densities integrates to 1 (even this is irrelevant for a compositional analysis)
- densities consists of **only relative information**

Standard methods for FDA have serious flaws

- If we summate two densities (convolution of two densities), the result is not a density (integral is not 1, $\int_{-\infty}^{\infty} f \cdot g \, dx \neq 1$)
- If we multipy a density, the result is not a density (integral is not 1, $\int_{-\infty}^{\infty} f + g \, dx \neq 1$)
- FPCA: results can even include negative estimates of density values
- density of 0.005 and 0.01 (doubling), 0.05 and 0.055 (1.1 times larger) has the same Euclidean distance, **but once it is doubling its value and the other time it is a small increase**.

Solutions

- a probability density function f can be considered as a compositional vector with infinitely many parts
- compositional data analysis and a log-ratio approach
- Bayes Hilbert Spaces $\mathcal{B}^2$ instead of Lebesque standard space $\mathcal{L}^2$ for integrable functions
- Norm, inner product, addition (perturbation), multiplication (powering) can be defined for $\mathcal{B}^2$ (Egozcue and Pawlowsky-Glahn 2006)(Boogaart, Egozcue, and Pawlowsky-Glahn 2014)

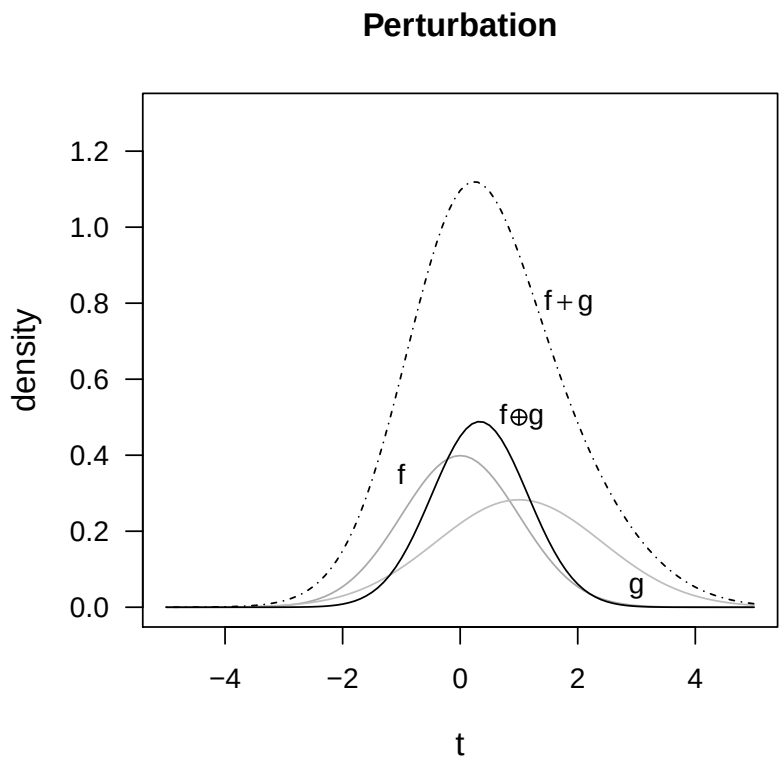
Solutions

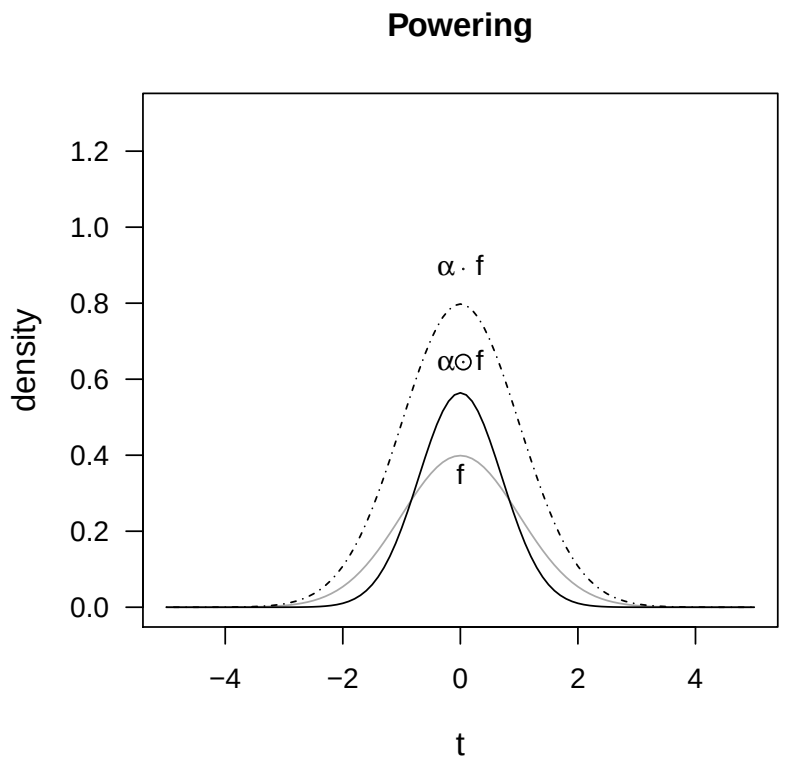
- a probability density function f can be considered as a compositional vector with infinitely many parts
- compositional data analysis and a log-ratio approach
- Bayes Hilbert Spaces $\mathcal{B}^2$ instead of Lebesque standard space $\mathcal{L}^2$ for integrable functions
- Norm, inner product, addition (perturbation), multiplication (powering) can be defined for $\mathcal{B}^2$ (Egozcue and Pawlowsky-Glahn 2006)(Boogaart, Egozcue, and Pawlowsky-Glahn 2014)
- Perturbation ($\oplus$) of two densities is again a density
- Powering ($\odot$) of two densities results in a density
- Isomorphism: Isometric *log-ratio* transformation as isometric isomorphism between Bayes-Space $\mathcal{B}^2$ and Euclidean Space $\mathcal{L}^2$

Addition and multiplication in $\mathcal{B}^2$

$f, g \in \mathcal{B}^2(I)$ integrable and α is a real constant $\alpha \in \mathbb{R}$:

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}.$$





Smoothing in FDA

Coming back to functional data analysis.

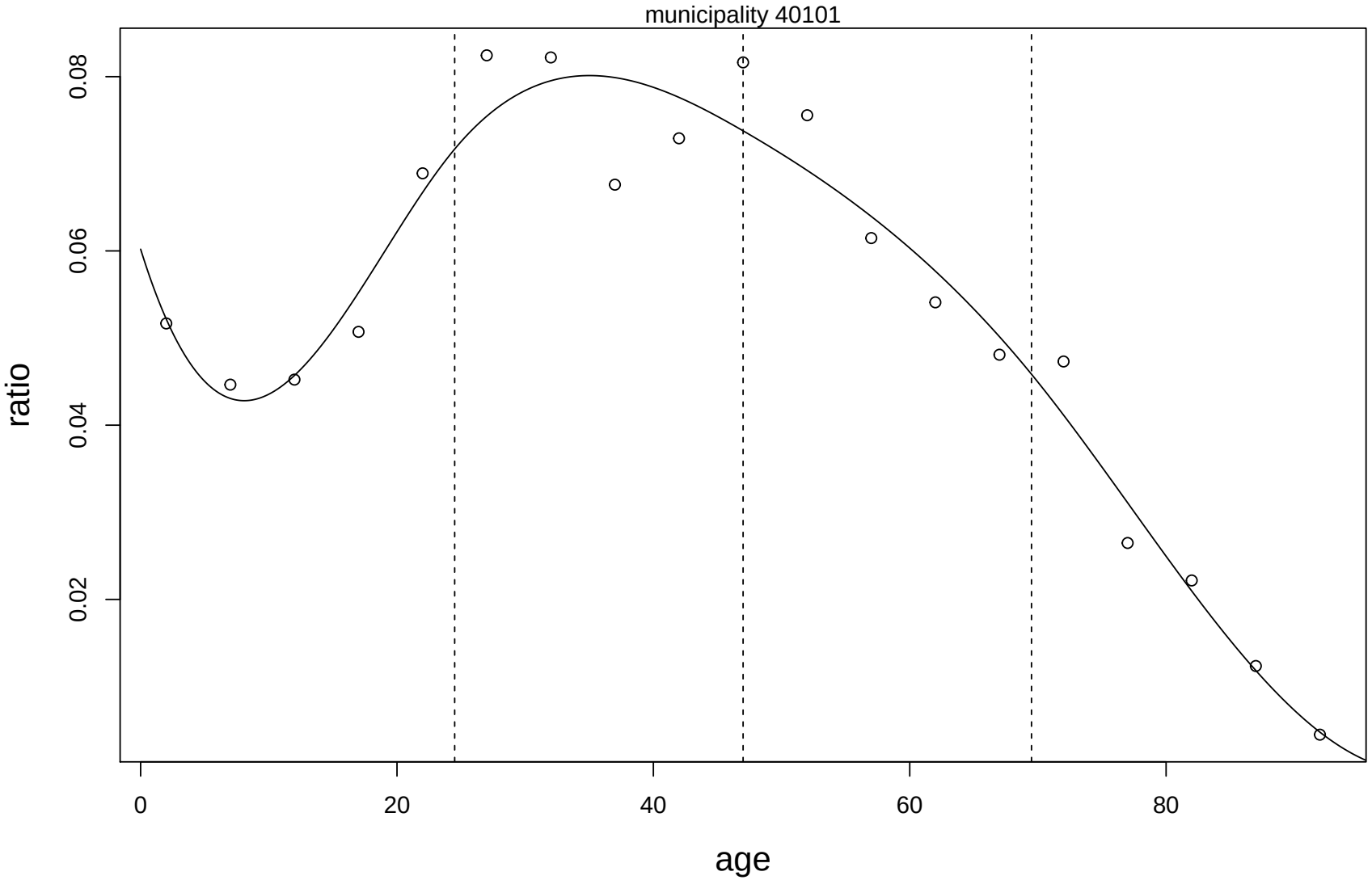
- Smoothing is necessary, because we start with discrete measurements

How is smoothing done?

- Many possibilites
- In practice often carried out with
  - natural cubic splines between all knots with some contraints (smoothness of the 2. derivation at knots, ...).
  - equivalent but with better numerical properties: B-Splines

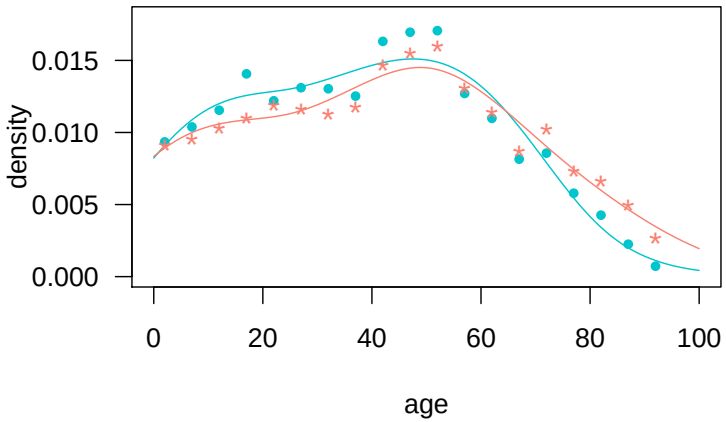
Smoothing in FDA

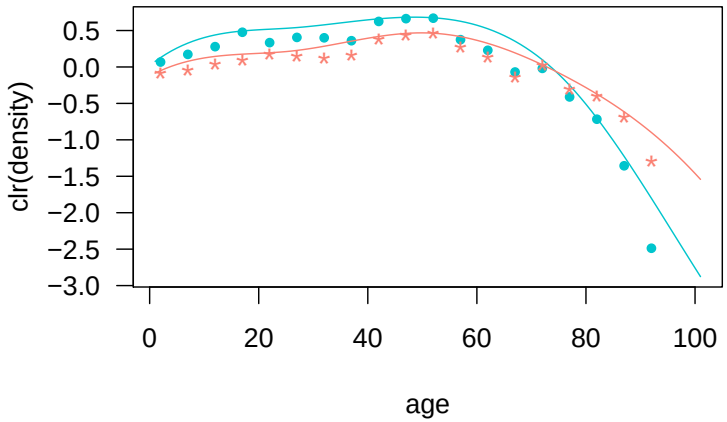
Lets take three knots at $Q_{0.25}(E)$, $Q_{0.5}(E)$ and $Q_0.75(E)$



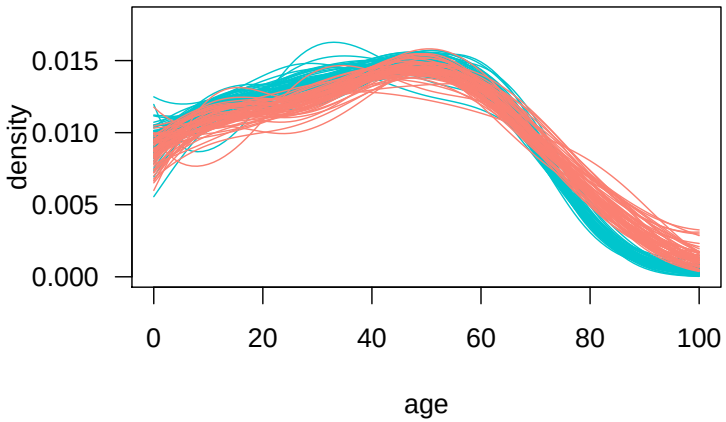
Darstellung in CLR Koordinaten und Splinesfit

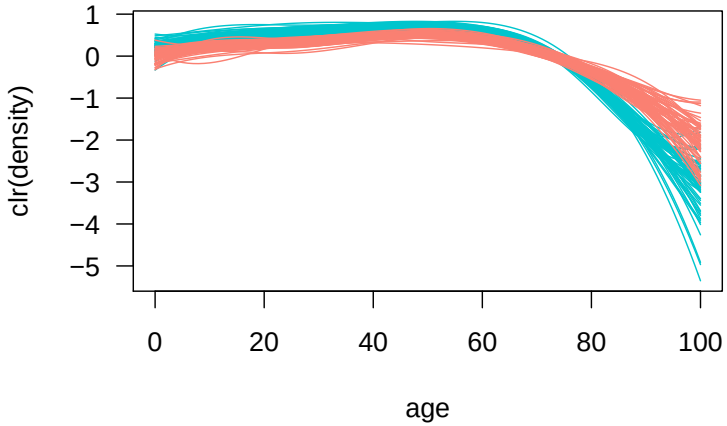
Splines for the age distribution density in community 40101





For all municipalities (right: in CLR coordinates)





Functional Principal Component Analysis (FPCA)

We already know how to fit functions to data, now we can proceed with the dimension reduction of our high-dimensional data

- centered **functional** random variable $X_1, ..., X_N$
- $\langle x, y \rangle_2 = \int_I x(t)y(t)dt$ ... defines the **inner product** between two elements x, y in $\mathcal{L}^2(I)$
- **Aim**: to describe the main variability of the data set with the help of K linear combinations of the original variables $X_i \approx \sum_{k=1}^{K} \langle X_i, \xi_k \rangle_2 \xi_k$
- FPCA: the first score vector ξ_1 in $\mathcal{L}^2(I)$ is obtained by maximisation of $\xi \in \mathcal{L}^2(I)$

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \xi \rangle_2^2 \quad \text{unter} \quad \|\xi\|_2 = 1.$$

Functional principal component analysis (FPCA)

- the remaining FPC's, $\{\xi_j\}_{j \geq 2}$ includes the remaining variablity of the data under the constraint of orthogonality $\langle \xi_k, \xi \rangle_2 = 0, k < j$
- FPC ξ_j are obtained with an eigenvalue decomposition of the covariance matrix

$\rightarrow$ **Output**: eigenfunctions of the covariance (**describing the variables**) and scores (**representing the observations**)

Interpretation is done using k functional principal componentes, e.g. the first 2, through visualisation (**biplot**)

Simplicial functional principal component analysis (SFPCA)

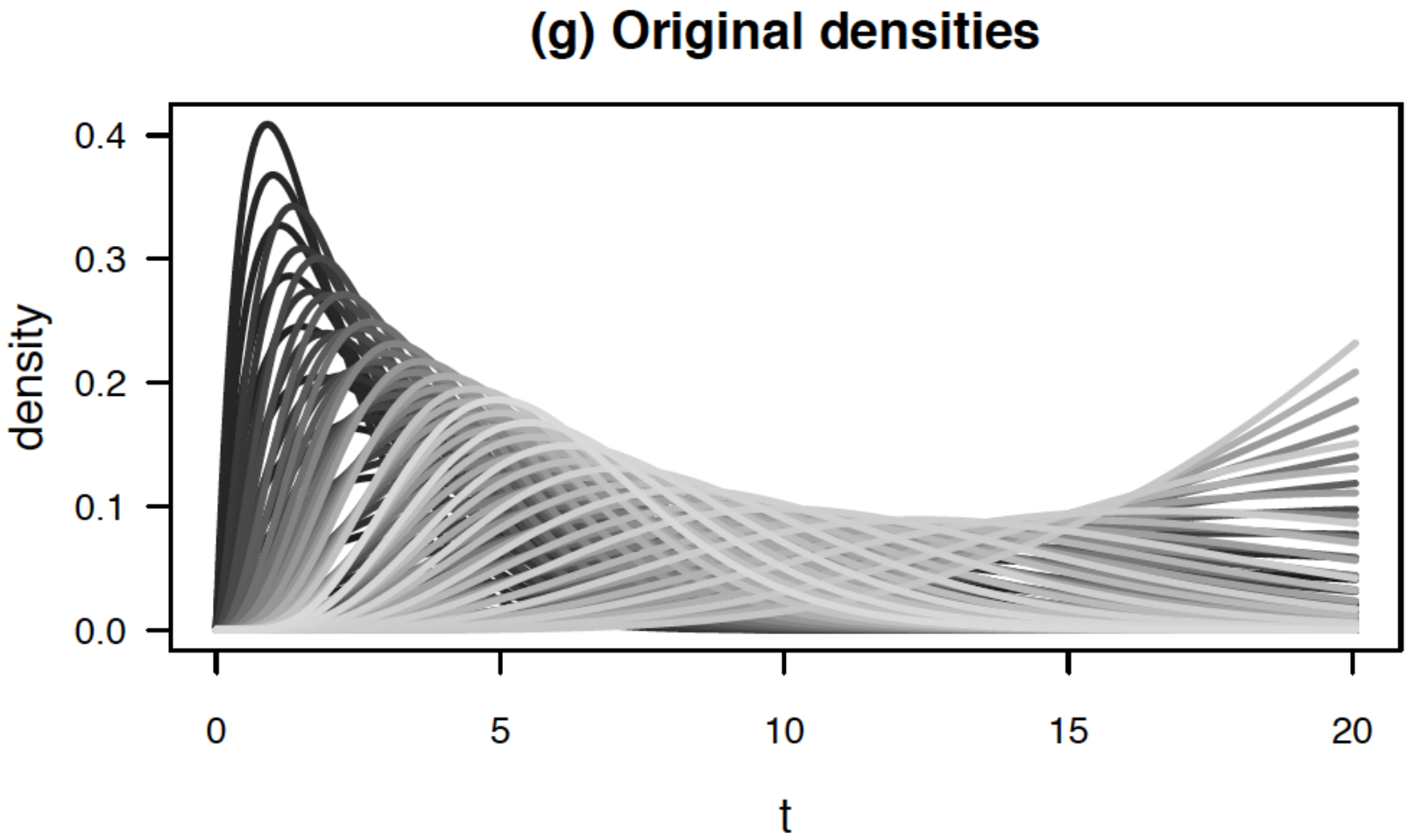
- $\rightarrow$ **SFPCA**: new formulation of FPCA in Bayes-Spaces for $X_1, ..., X_N$ (centred) sample in $\mathcal{B}^2(I)$
- Maximisation through $\zeta \in \mathcal{B}^2(I)$

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \zeta \rangle_B^2 \text{ under constraint } \|\zeta\|_B = 1;\ \langle \zeta_j, \zeta_k \rangle_B = 0,\ k < j$$

- $\rightarrow$ we can formulate and solve the problem, because $\mathcal{B}^2(I)$ is a Hilbert-Space (but equations are complicated and looooong, we refer to Hron et al. (2016))
- $\rightarrow$ **Problem**: Efficient implementation
- $\rightarrow$ **Solution**: with the help of *centred log-ratio* coordinates (division of the values of a composition with its geometric mean and taking the log)
- Smoothing and PCA in CLR-coordinates under certain constraints (Hron et al. 2016)

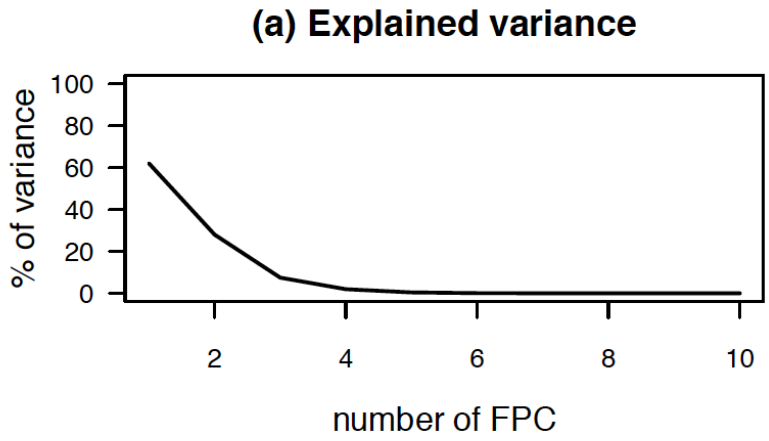
Simulation: comparing FPCA and SFPCA

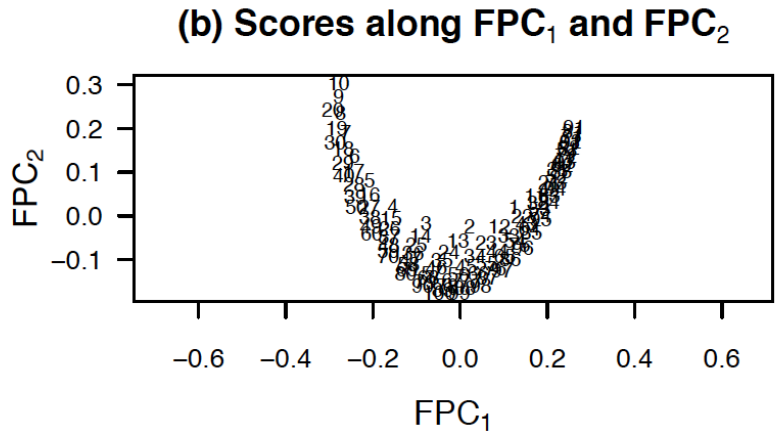
100 densities simulated from a Gamma distribution:

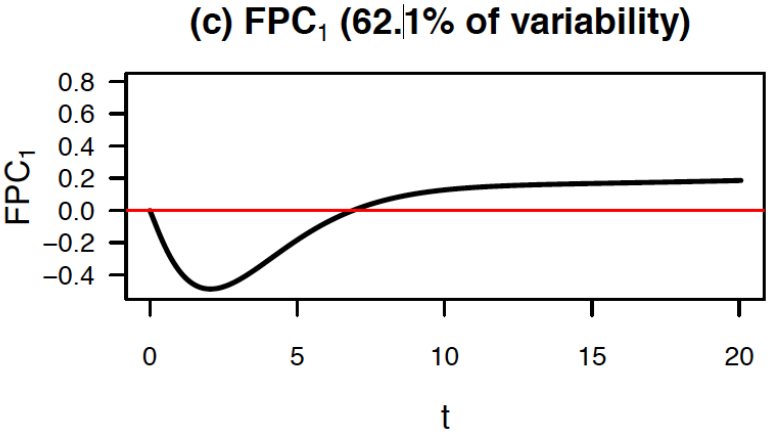


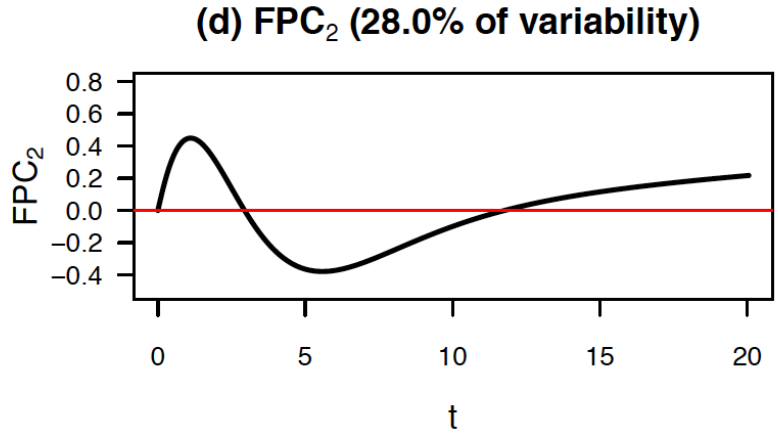
Simulation: comparing FPCA and SFPCA

with FPCA we see some non-linearities in the scores



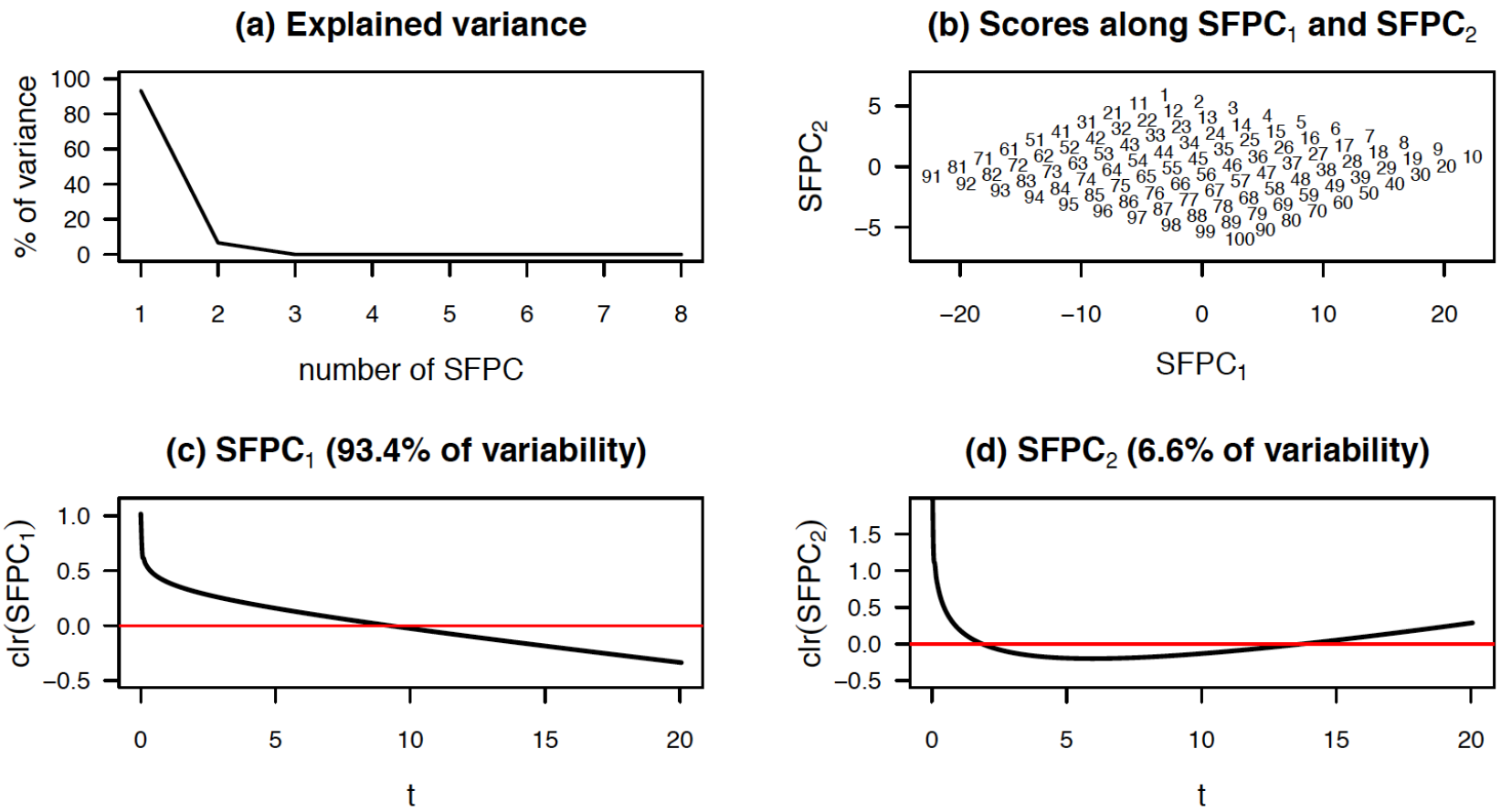






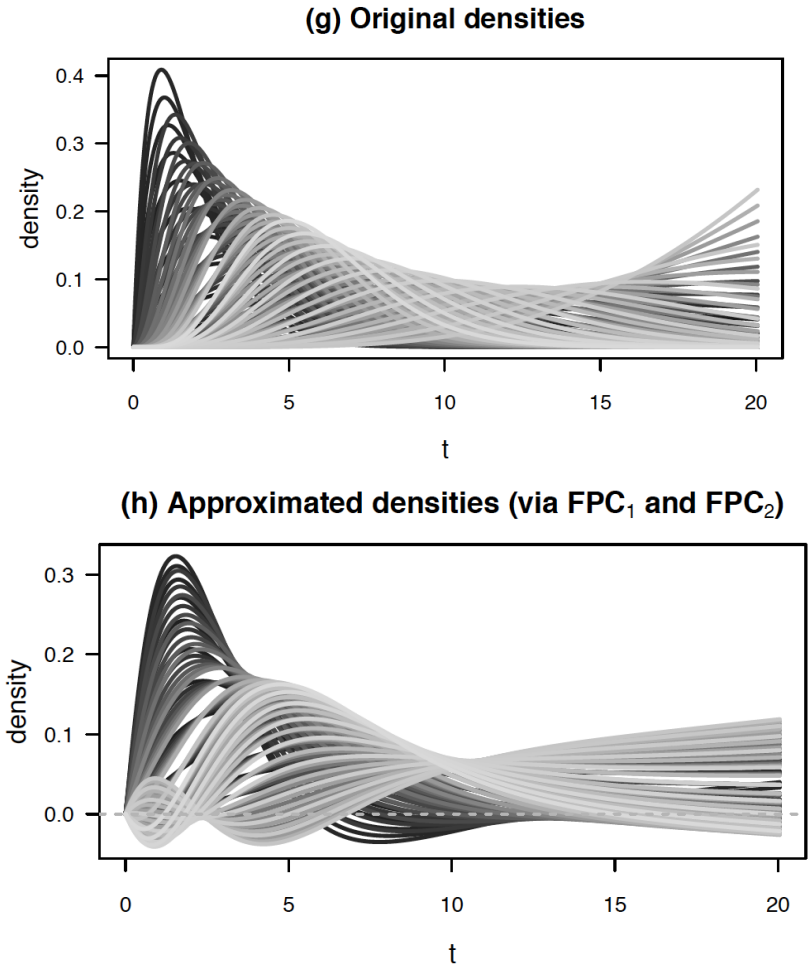
Simulation: comparing FPCA and SFPCA

With SFPCA we have higher exlained variance



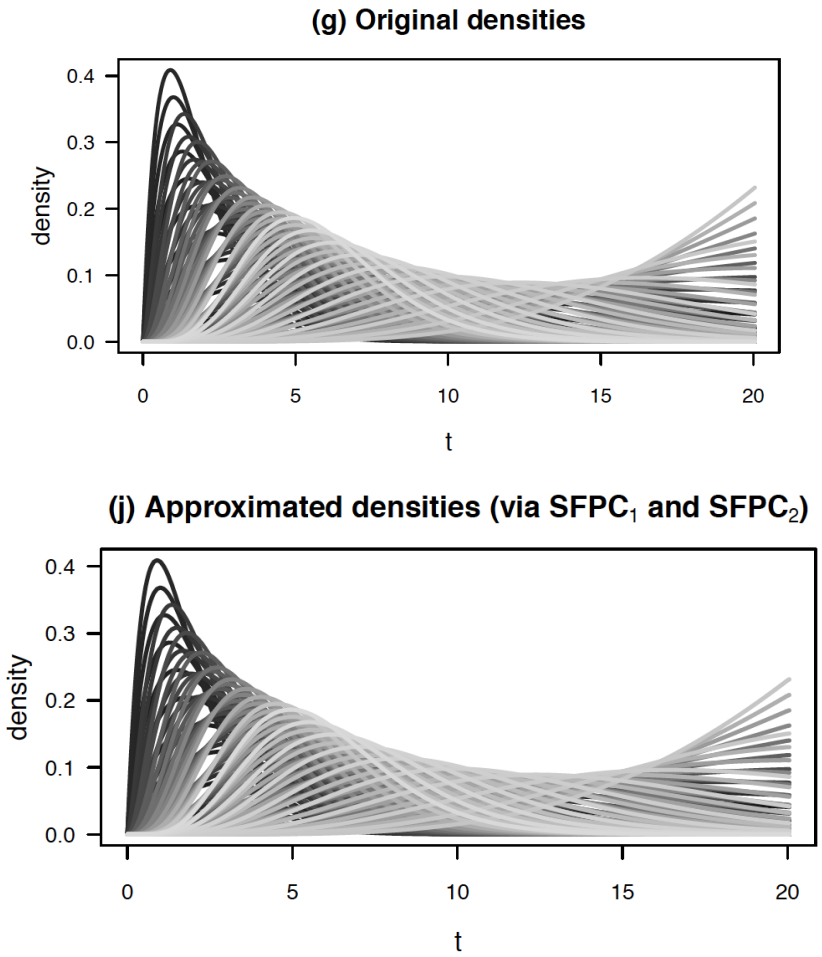
Simulation: comparing FPCA and SFPCA

Lets use the FPCA fit to fit the densities. We receive even negative values of the estimated densities



Simulation: comparing FPCA and SFPCA

With SFPCA, the estimates are again densities and we can reconstruct the original densities from our fit.



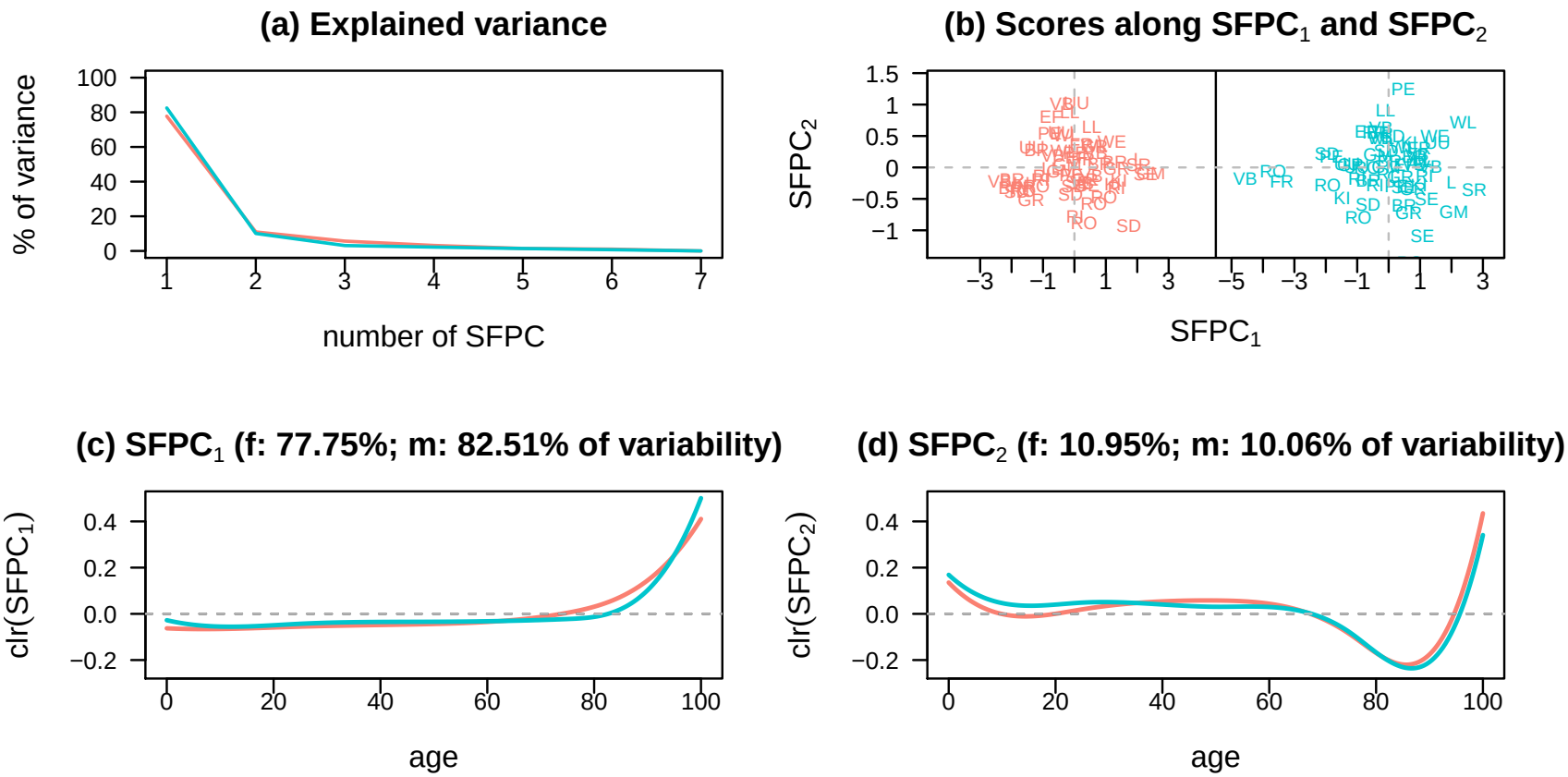
Application to population data from Upper Austria

- to characterise the main variability in population densities
- to make a dimension reduction in Bayes-Spaces $\mathcal{B}^2(\mathcal{I})$ in opposition to $L^2(\mathcal{I})$

We have already shown:

- we can expect much better results with SFPCA as in comparison to FPCA
- we have shown that this is also true for real-world data (Hron et al. (2016))
- in the following we only discuss the results from SFPCA

Scores (Upper Austrian Population Data)

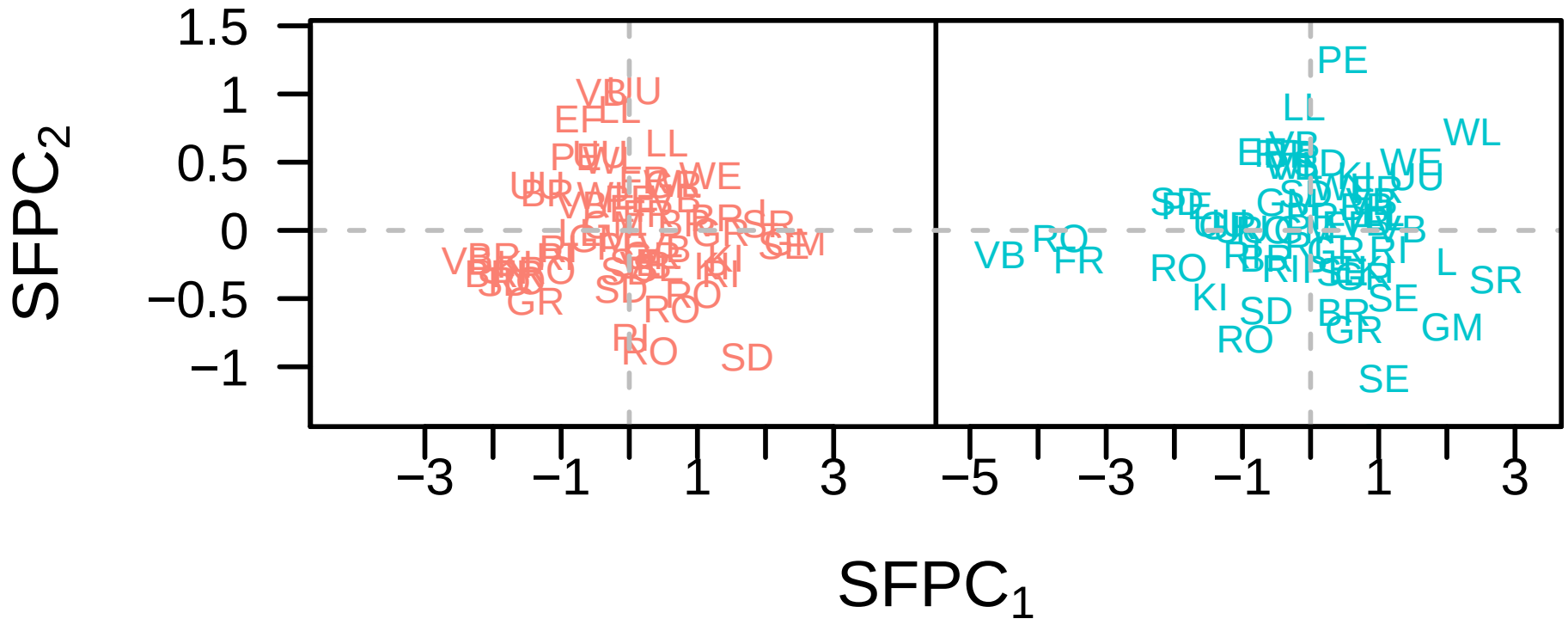


SFPC1: old/young: high scores, whenever the ratio of old people is high in comparison to the whole age distribution.

SFPC2: contrast of the 69-92 years old people to the rest of the age distribution

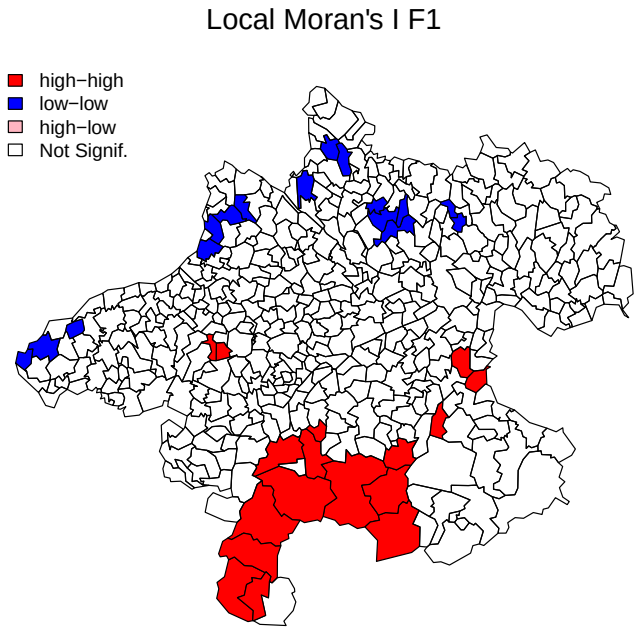
Scores (Upper Austrian Population Data)

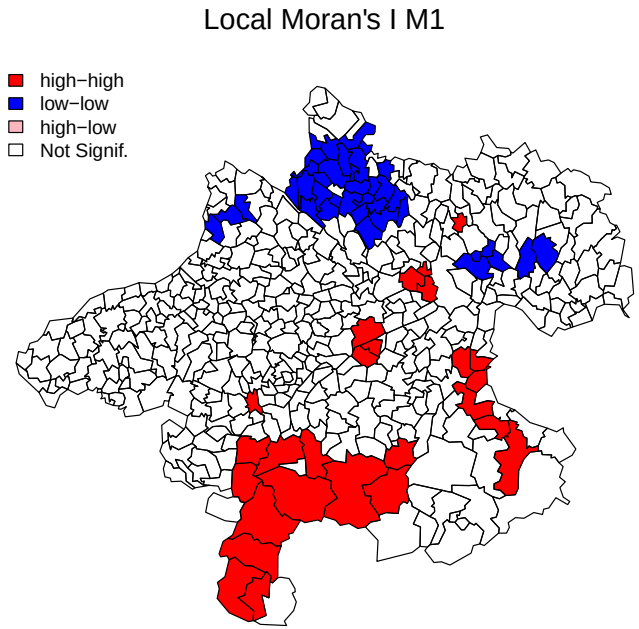
- Interpretation is done together with the help of the previous and following maps
- The observations (on municipalities) are represented by the abbreviations



Due to time constraints, we will not go into detail here, but show the scores in maps

Results - Lisa maps

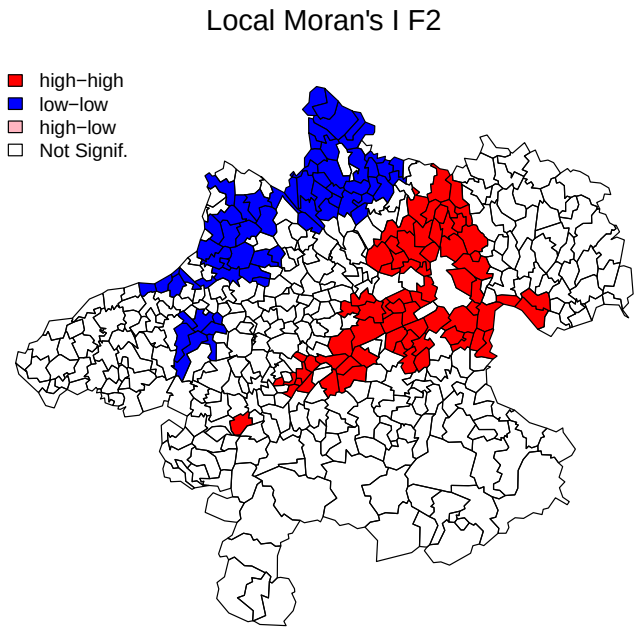


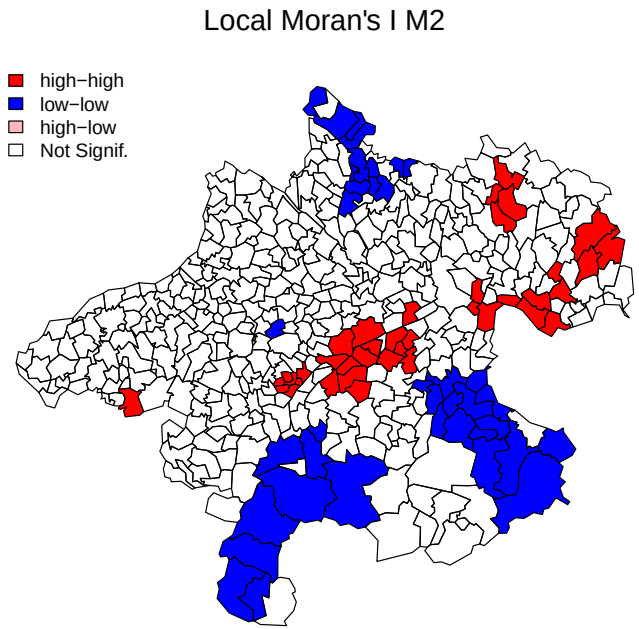


Gmunden and Ennstal: people are old, relatively to the whole age distribution in these regions. Young people were moving to the captial (Linz, and in the surrounding of Linz).

above Linz: young families moved there

Results - Lisa maps





Rohrbach, Schärding and surrounding: a way too much old people relatively to the whole age distribution. Women emigrated from these region.

Along the highway and industrial zones: many people fit for work are migrated

Conclusion

Main aim was to show:

- Many (of your) data sets are compositional.
- New possibilities and findings in respect to new research (*A primer on the need and usefulness of (new) research*).

Conclusion

Main aim was to show:

- Many (of your) data sets are compositional.
- New possibilities and findings in respect to new research (*A primer on the need and usefulness of (new) research*).

We also showed:

- With the help of CoDa methods we improved FDA
- This was a very specific application, but should show you the success of the CoDa methods for compositional data
- The succesful application of a log-ratio analysis of CoDa in correlation, cluster, discriminant, principal component, regression analysis, compositional tables, (and many more methods) is shown in (Filzmoser, Hron, and Templ 2018).
- Applications in Small Area Estimation of proportions, and new theory in fitting confidence intervals of ratios for complex surveys (Hron, Templ, and Filzmoser 2013)

Conclusion

Main aim was to show:

- Many (of your) data sets are compositional.
- New possibilities and findings in respect to new research (*A primer on the need and usefulness of (new) research*).

We also showed:

- With the help of CoDa methods we improved FDA
- This was a very specific application, but should show you the success of the CoDa methods for compositional data
- The succesful application of a log-ratio analysis of CoDa in correlation, cluster, discriminant, principal component, regression analysis, compositional tables, (and many more methods) is shown in (Filzmoser, Hron, and Templ 2018).
- Applications in Small Area Estimation of proportions, and new theory in fitting confidence intervals of ratios for complex surveys (Hron, Templ, and Filzmoser 2013)

Outlook compositional FDA

- SFPCA was also (successfully) applied to mortality statistics, with nice results
- The method was extended to functional regression with functional response (Talská et al. 2018)
- Further mathematical foundations of probability density functions using B-splines for Bayes Spaces in Machalová et al. (2020)

Software for compositional FDA and CoDa

Main reference for Compositional data analysis:

- R packages `robCompositions` (Templ, Hron, and Filzmoser 2011, 2020) and `compositions` (Boogaart, Tolosana, and Bren 2020)

Example of functions in `robCompositions`:

- `compositionalSpline` ... spline fit
- `pcaCoDa` ... principal component analysis
- `impCoda` ... imputation of missing values
- `intTab`, `indTab`, `rSDev.test` ... contingency tables and related tests
- `isomLR`, `cenLR`, ... transformations
- `daCoDa` ... discriminant analysis
- ...

and 45 compositional data sets.

References I

Boogaart, K.G. van den, J.J. Egozcue, and V. Pawlowsky-Glahn. 2014. "Bayes Hilbert Spaces." *Australian & New Zealand Journal of Statistics* 56 (2): 171–94. https://doi.org/10.1111/anzs.12074.

Boogaart, K.G. van den, R. Tolosana, and M. Bren. 2020. *Compositions: Compositional Data Analysis*. http://CRAN.R-project.org/package=compositions.

Egozcue, J.J., and V. Pawlowsky-Glahn. 2006. "Compositional Data Analysis in the Geosciences: From Theory to Practice." In, edited by A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, 145–60. Geological Society, London.

Filzmoser, P., K. Hron, and M. Templ. 2018. *Applied Compositional Data Analysis: With Worked Examples in R*. Springer Series in Statistics. Cham, Switzerland: Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-96422-5.

Hron, K., A. Menafoglio, M. Templ, K. Hrůzová, and P. Filzmoser. 2016. "Simplicial Principal Component Analysis for Density Functions in Bayes Spaces." *Computational Statistics & Data Analysis* 94 (C): 330–50.

References II

Hron, K., M. Templ, and P. Filzmoser. 2013. "Estimation of a Proportion in Survey Sampling Using the Logratio Approach." *Metrika: International Journal for Theoretical and Applied Statistics* 76 (6): 799–818. https://doi.org/10.1007/s00184-012-0416-6.

Machalová, Jitka, Renáta Talská, Karel Hron, and Aleš Gába. 2020. "Compositional Splines for Representation of Density Functions." *Computational Statistics*. https://doi.org/10.1007/s00180-020-01042-7.

Talská, R., A. Menafoglio, J. Machalová, K. Hron, and E. Fišerová. 2018. "Compositional Regression with Functional Response." *Computational Statistics & Data Analysis* 123: 66–85. https://doi.org/https://doi.org/10.1016/j.csda.2018.01.018.

Templ, M., K. Hron, and P. Filzmoser. 2011. "robCompositions: An R-Package for Robust Statistical Analysis of Compositional Data." In *Compositional Data Analysis*, 341–55. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch25.

———. 2020. *robCompositions: Robust Estimation for Compositional Data*. http://CRAN.R-project.org/package=robCompositions.