Analyse von relativer Information

Matthias Templ

Fachvortrag ZHAW, 10.05.2016

Inhalt der Präsentation:

Kompositionsdatenanalyse. Warum dieses Thema gewählt?

Der falsche Begriff: Kompositionsdaten

Grundlagen von *Kompositionsdaten*

Anwendungen

Einbringen in ZHAW

Alles falsch?

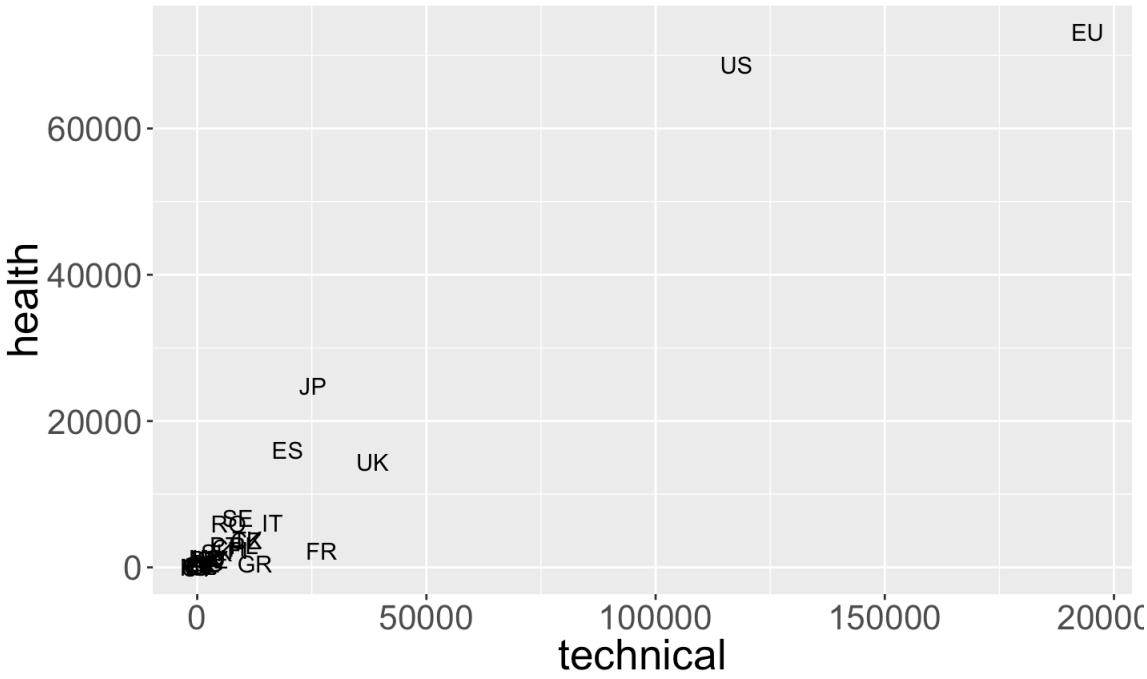
Motivation PhD Daten

```
library("robCompositions")
data(phd)
str(phd)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ country           : chr  "EU" "Belgien" "Bulgarien" "Tschech.R
##  $ countryEN         : chr  "EU" "Belgium" "Bulgaria" "CzechRepub
##  $ country2          : chr  "EU" "BE" "BG" "CZ" ...
##  $ total             : num  516.5 7.5 5.2 22.6 4.8 ...
##  $ male              : num  52.4 59 49.7 62.1 54.2 46.5 52.1 55.6
##  $ female            : num  47.6 41 50.3 37.9 45.8 53.5 47.9 44.4
##  $ technical         : num  36.9 46.2 39.7 46.4 39.3 42.3 49.2 55
##  $ socio.economic.law: num  22.9 19.6 21.2 16.3 12.8 21.2 14.7 17
##  $ human             : num  21.6 13.3 22.5 15.3 14.5 21 21 22.6 2
##  $ health            : num  13.9 13.9 12.8 15.8 25.2 9.8 8.4 2.2
##  $ agriculture       : num  2.8 7.1 3.8 4.5 8.2 5.6 2 1.7 2 0.1 .
```

Motivation PhD Daten

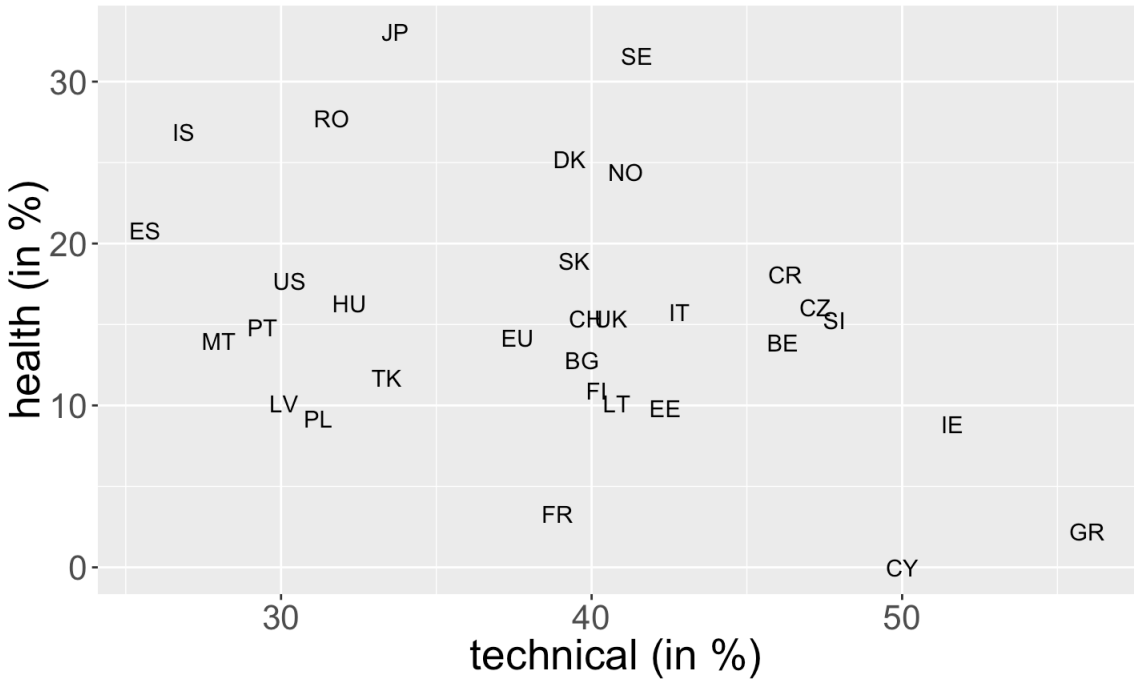
Absolutinformation



5/38

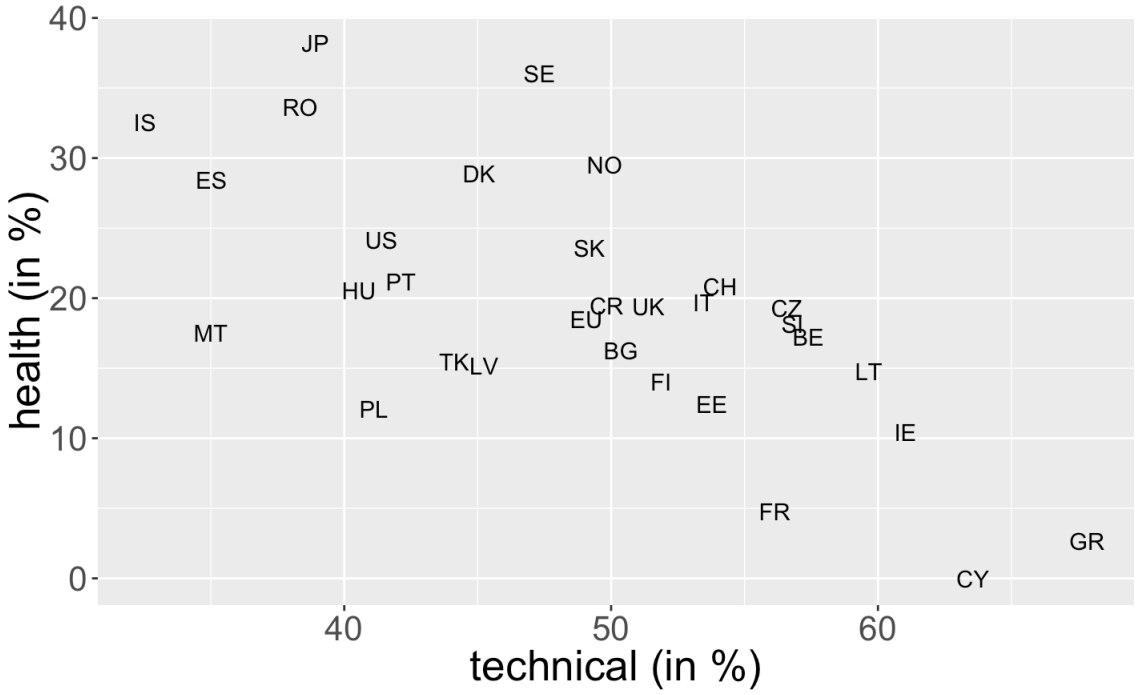
Motivation PhD Daten

In %



Motivation PhD Daten

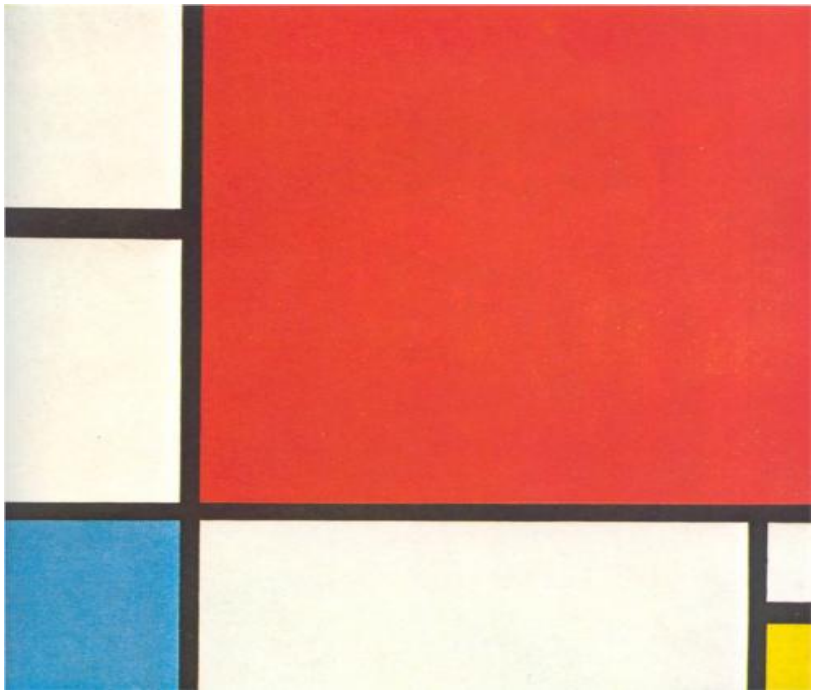
Ohne socio-economic und law sciences



Kompositionsdaten

Piet Mondrian:

Composition with Red, Blue and Y.



Multivariate Daten mit konstanter Zeilensumme

$$\mathbf{x} = (x_1, \ldots, x_D)^t, \quad x_i > 0, \quad \sum_{i=1}^{D} x_i = \kappa$$

Die Menge aller Kompositionen mit positiven Werten liegt im sogenannten **Simplex** $\neq$ Euklidischer Geometrie $\rightarrow$ jegliche klassische statistische Analyse ist fehl am Platz.

Beispiele für *Kompositionsdaten*: Haushaltsausgaben, Steuerkomponenten, Wahrscheinlichkeitstabellen, geochemische Daten, Spektren, Artenreichtum, 24h Tag, …

Das Hauptaugenmerk liegt auf den Verhältnissen zwischen den Anteilen.

Anmerkungen

Spurious Correlation und Simpsons Paradoxon nur weil in der falschen Geometrie gearbeitet wird.

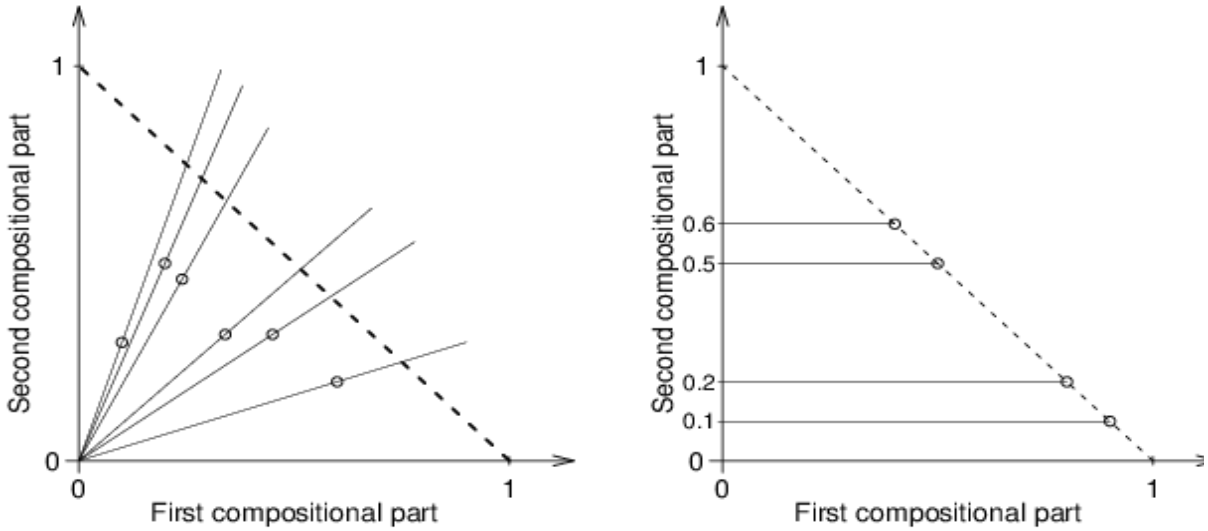
Skaleninvarianz, Subcompositional Coherence, Perumtationsinvarianz wichtig und alles verletzt in klassischer Statistik von relativer Information.

$\rightarrow$ Kompositionsdatenanalyse

$\rightarrow$ Arbeiten in Koordination von log-ratios

Analyse in Koordinaten

Aitchison Distanz



Links: 2-part Komposition. Ratios bleiben erhalten.

Rechts: Standard-Euklidische Distanz nicht geeignet.

Arbeiten in Koordinaten mit ilr

$$ilr(\mathbf{x}) = (z_1, \ldots, z_{D-1})^t, \quad z_j = \sqrt{\frac{D-j}{D-j+1}}\ \ln \frac{\cdot}{\sqrt[D-j]{\cdot}}$$

mit $j = 1, \ldots, D-1$.

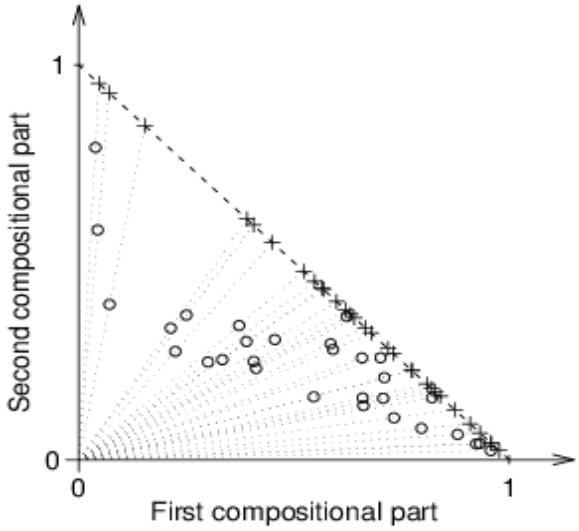
Diese Wahl garantiert, dass Werte in x_1, $z_2, \ldots, z_{D-1}$ nicht beeinflussen.

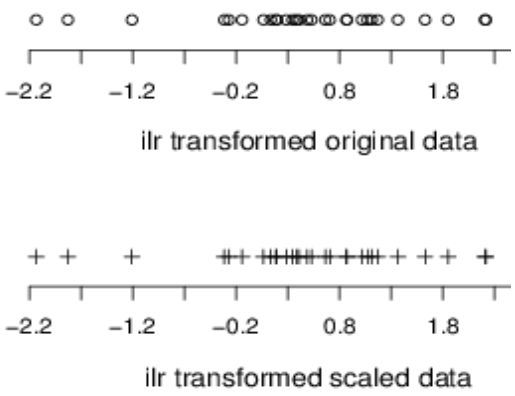
$$d_A(\mathbf{x}, \mathbf{y}) = d_E(ilr(\mathbf{x}), ilr(\mathbf{y}))$$

ZB 3-part Komposition:

$$z_1 = \sqrt{\frac{2}{3}}\ \ln \frac{x_1}{\sqrt[2]{x_2 x_3}} \quad , \quad z_2 = \sqrt{\frac{1}{2}}\ \ln \frac{x_3}{x_2}$$

Eigenschaften der ilr Transformation





Linke Grafik: 2-Part Kompositionsdaten konstante Zeilensumme (Symbole: ○), und Zeilensumme 1 (Symbole: +).

Rechte Grafik: Darstellung in Koordinaten

Anwendungen

Bier

Bier-Daten

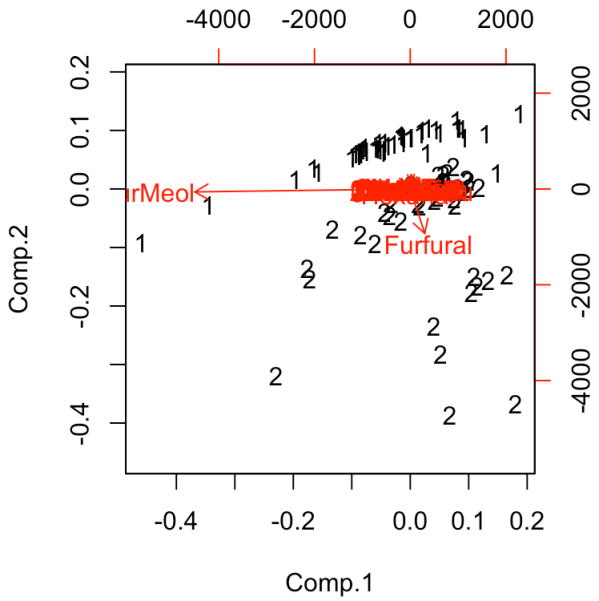
```
str(beer)
```

```
## 'data.frame':    86 obs. of  19 variables:
##  $ Betrieb : Factor w/ 45 levels "101RO-M","102RO-C",..: 25 43 44
##  $ BetrNr  : Factor w/ 10 levels "1","2","3","4",..: 1 1 1 1 1 1
##  $ newold  : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 3MeBu-al: num  23.4 16.3 101.4 18.3 152.9 ...
##  $ 3MeBu-on: num  10.3 49.5 29.5 7.4 45.4 4 9.1 24.7 6.7 8.2 ...
##  $ 2MeBu-al: num  44.4 99.8 223.9 492.2 142.3 ...
##  $ Hexanal : num  180 167 200 137 296 ...
##  $ 2FurMeol: num  3287 1555 2228 1895 3934 ...
##  $ Heptanal: num  4.2 3.9 4.2 4 8 3.8 4.3 4.9 4 4.4 ...
##  $ 2AcFur  : num  40.7 32.2 41.7 29.6 43.3 21.2 29.9 16.9 27.4 27
##  $ 5Me2Fur : num  45 41.1 51.9 34 54.6 29.1 34.6 26.7 32.2 32.3 .
##  $ EssFuEst: num  33.5 39.9 38.8 15.9 49.5 10.5 21 33.2 16.2 15.9
##  $ 2Ac5MeFu: num  26 14.8 27.8 15.6 8 13.9 24.1 4.5 7.1 6.3 ...
##  $ 2PhEt-al: num  21.1 15.6 14.9 11.3 23.2 7.8 12 8.1 9.1 7.1 ...
##  $ NicEtEst: num  10.1 17.2 14.7 28.3 30.9 5.7 11.4 11.2 9.1 9.3
##  $ 2PhEssEt: num  0.9 1.6 1.4 1.5 2 0.9 0.9 0.9 0.9 0.8 ...
##  $ gNonalac: num  82.6 93.8 85.3 77.6 221.7 ...
##  $ Furfural: num  21 15 14 13 30 17 23 18 16 20 ...
##  $ HMF     : num  0.92 0.8 0.72 0.95 0.79 0.86 0.76 0.86 0.81 1.1
```

Bier Daten PCA

Originalstudie (Varmuza et.al, 2002)

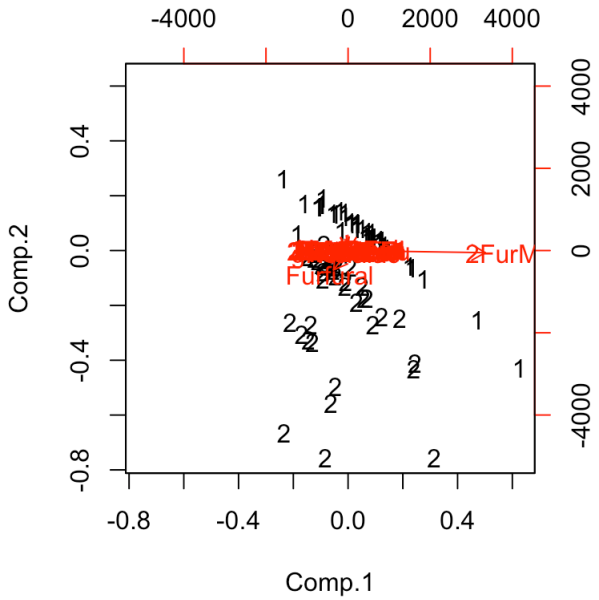
```
vals <- 4:ncol(beer)
biplot(princomp(beer[, vals]), xlabs=as.numeric(beer$newold))
```



Bier Daten PCA

Robuste PCA kann auch nichts reparieren

```
biplot(princomp(beer[, vals],
        covmat=covMcd(beer[, vals])),
       xlabs=as.numeric(beer$newold))
```

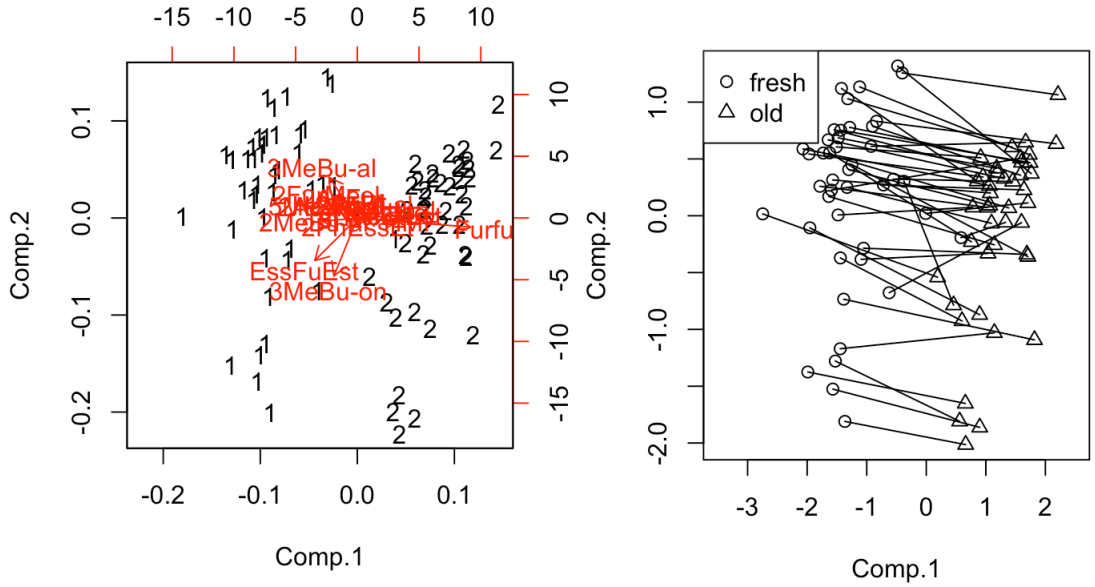



```
# --> strange, of course, it is the classical (wrong) approach.
```

Bier Daten PCA

Intern (`pcaCoDa`): ILR + robuste PCA + orthog. Projektion

`?robCompositions::pcaCoDa` (Templ, Hron, Filzmoser; 2016)



Eigene Entwicklungen in CoDA

Clusteranalyse mit Kompositionsdaten

Clusteranalyse mit Kompositionsdaten (Templ, Filzmoser, Reimann; 2008; Applied Geochemistry)

Anwendung geochemische Daten von Norwegen, Finnland und Russland; Gesteinsdaten von Aut. Blätter von Oslo.


Nachweis von Umweltverschmutzung, etc.. CoDa Methoden besser.

Imputation von fehlenden Werten

(Hron, Templ, Filzmoser; 2011, CSDA)

EM-Algorithmus zur Imputation

ILR Transformationen

Robustheit

KNN Methode für Kompositionsdaten

CoDa Methoden (weit) besser als jegliche klassische Methode

Imputation von zensierten Daten

(Templ, Hron, Filzmoser; 2016; JAS)

EM-Algorithmus zur Imputation (Tobit-Regression)

Kein Vergleich mehr zu nicht-CoDa Methoden ;-)

Imputation von hochdim. Daten

(Templ, Hron, Filzmoser, Alzbetka; 2016; Chemolab). Algorithmus der mit hochdim. Daten und Detection Limits umgehen kann. Sehr komplex

Ausreisser mit CoDa und Nullen

(Templ, Hron, Filzmoser; 2016; JAS). Schätzung der rob. Kovarianz von imputierten Daten. Rob. Mahalanobis-Distanzen in Sub-Kompositionen + Ausreisser in Nullerstruktur

Robuste Diskriminanzanalyse mit Kompositionsdaten

(Filzmoser, Hron, Templ, 2012)

CoDa Fisher-DA

Funktionale PCA

(Hron, Menafoglio, Templ, Filzmoser; 2015; CSDA)

SFPCA. CoDa, da $\int_{-\infty}^{\infty} f(x)dx = c$. Math. anspruchsvollste Arbeit

Kontingenztafeln und Wahrscheinlichkeitstabellen mit CoDa

(Egoscue, Pawlowsky, Templ, Hron; 2015; Comunications in Statistics) (Hron, Facekovia, Templ, Todorov; 2014; JAS) (Facevicova, Hron, Todorov, Templ; 2015; Scandinavian Journal of Statistics.)

Jede Tabelle ist ein Kompositionsproblem. Pearsons χ^2-Test theoretisch falsch.

Neuer Test

Serie von Tabellen. ILR. Anwendungen bei UNIDO und Statistik Austria.

Kontingenztafeln und Wahrscheinlichkeitstabellen mit CoDa



Survey-Statistik mit CoDa

(Hron, Templ, Filzmoser; 2013; Metrika)

Stratified random sampling mit geometrischen Mittel von Anteilen (arithmetische Mittel falsch).

Nicht-symmetrische Konfidenzintervalle

Klassische Inferenzstatistik falsch sobald Anteile geschätzt werden.

Robuste Mulitvariate Methoden mit CoDa & Buch

R Paket `robCompositions`

Springer Buch CoDa (2016) mit Peter Filzmoser und Karel Hron

Bisherige Bücher zu theoretisch

Neuer Ansatz (ohne konstante Zeilensumme)

Hoffentlich durch Buch Akzeptanz der Methoden in angewandten Wissenschaften.

Prognose Stelle

Grundsätzlich

Aufteilung der Arbeit ist eindeutig ein kompositionales Problem 

CV: Interesse an allen was mit Daten aber auch mit Lehre zu tun hat

Keine Angst vor Neuem

Uebergeordnetes Interesse: Kontakt zur Schweizer Statistischen Gesellschaft. D-AUT existiert aber D-AUT-CH noch nicht sichtbar. Austrian Journal of Statistics?

Wieviel Freiheit in Tun und Lassen aufgeben?

DAS

Data Science (Statistik + Spezialgebiete der Informatik) ist das Zukunftsthema (seit langem). Wichtig: Kombination Mathe + Angewandte Statistik + Informatik (ohne Logik, etc).

DAS ist sehr gut aufgestellt.

Mathematischer Background und Interesse in Angewandter Wissenschaft und Methodenentwicklung/Computational Statistics

TU Wien geprägt. Statistical Computing / R Kompetenz.

Data Science und R. Buch *Simulation for Data Science in R*; Packt Publishing, 2016.

Verstärkung von DAS

Lehre

Nach Bedarf, aber Vermittelung von R, Reproduzierbarkeit, Kompetenz in wissenschaftlichen Arbeiten.

Nahe am Studenten (Regelmässigkeit)

Keine Theorie ohne Anwendung, keine Anwendung ohne Theorie.

Ausgleich: Anekdoten und Fun

"Leider": relativ fordernd, aber mit viel Motivation.

Forschungsgelder einbringen

Gutes Netzwerk

Anträge *in der Schublade* in Syntetischer Datengenerierung / Microsimulation, Ausreissererkennung, Imputation, CoDa, Bio-Sciences, Data Privacy, Big Data.

Situationsbedingt

Eher nur bedingt ab und zu EU wegen geringer Akzeptanzrate. Oft besser mit Wirtschaft/Organisationen. Nationale Funds aber auch eher Richtung Austauschprogrammen?