Einführung in funktionale Daten

Institut für Statistik
Ludwig-Maximilians-Universität München

Jona Cederbaum

12.10.2015

Funktionale Daten

Einfache Statistiken für funktionale Daten

Ausblick: Glättung und Registrierung

Funktionale Daten

- i.d.R. liegen diskrete Messungen vor
- Annahme: aufeinanderfolgende Werte "gehören zusammen" und es kann theoretisch beliebig fein gemessen werden

  $\rightarrow$ den Daten liegt eine glatte Funktion zugrunde
- Funktionen haben einen stetigen Definitionsbereich, z.B. Zeit, Wellenlänge oder Raum
- meist Kurven, aber auch Oberflächen oder 3D-Daten möglich
- Beobachtungseinheit: eine Funktion/Funktionsauswertung

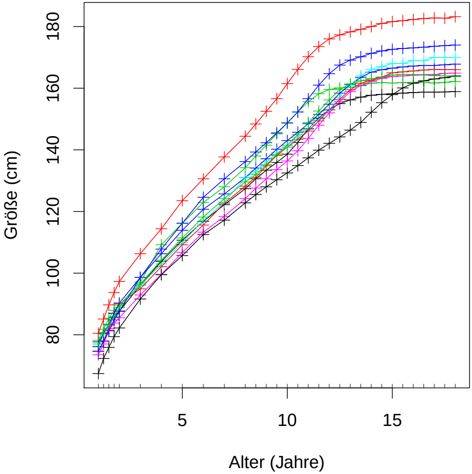
(Ramsay und Silverman 2005)

Handschriftliches fda (Position des Stiftes)

Cursive handwriting samples (Ramsay und Silverman 2005)

Wachstumskurven

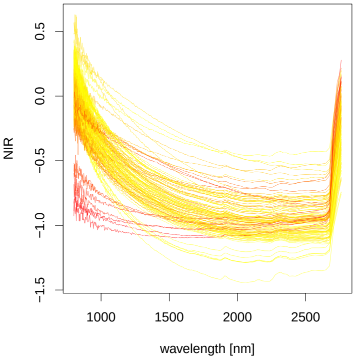
Wachstumskurven von 10 Mädchen

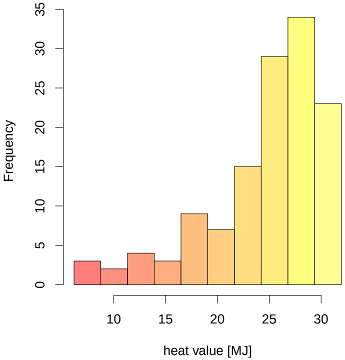


Berkeley Growth Study data (Ramsay und Silverman 2005)

Spektraldaten von fossilen Brennstoffen

Spektraldaten von fossilen Brennstoffen zur Vorhersage des Brennwertes

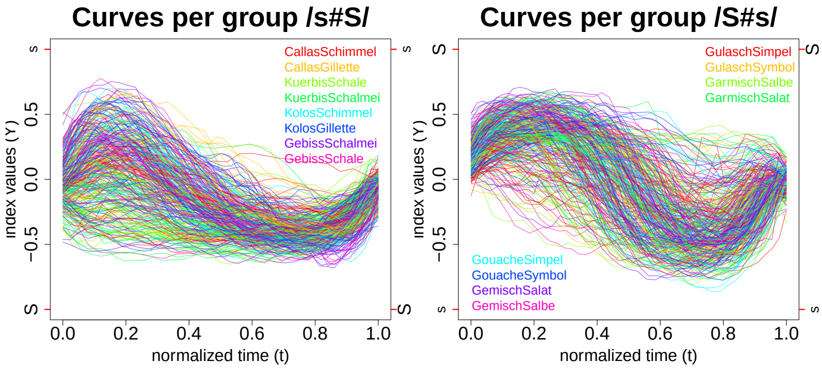




Spectral data of fossil fuels (Fuchs et al., 2015)

Konsonanten Daten

Funktionale Zielgröße

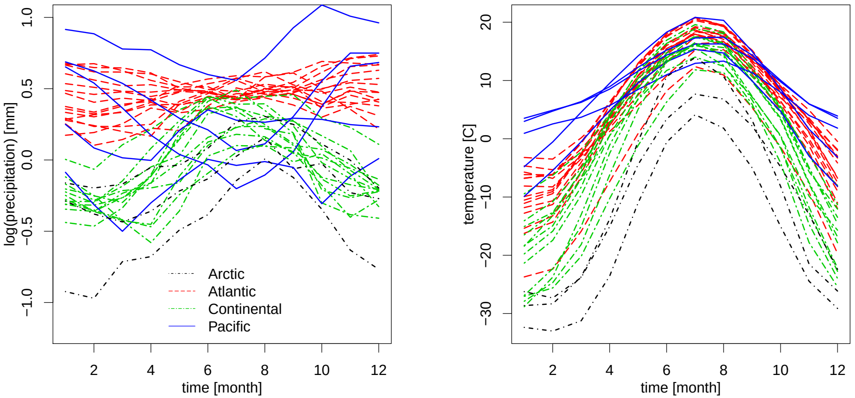


Consonant assimilation (Cerderbaum et al., 2015)

Kanadische Wetterdaten

Funktionale Einfluss- und Zielgröße

z.B. Vorhersage des log. Niederschlags aus der Temperatur



Canadian average annual weather cycle (Ramsay und Silverman 2005)

Mittelwert, Varianz und (Auto-)kovarianz

- Betrachte funktionale Variable $X(t)$, mit $t \in \mathcal{T}$ und $\mathcal{T}$ Intervall in $\mathbb{R}$
- Stichprobe $x_i(t)$, $i = 1, \ldots, n$
- funktionaler Mittelwert:

$$\hat{\mu}_X(t) = \bar{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t)$$

- funktionale Varianz:

$$\hat{s}_X^2(t) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t) - \bar{x}(t)]^2$$

- funktionale (Auto-)kovarianz:

$$\hat{K}_X(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t_1) - \bar{x}(t_1)][x_i(t_2) - \bar{x}(t_2)]$$

Mittelwert, Varianz und (Auto-)kovarianz

- Betrachte funktionale Variable $X(t)$, mit $t \in \mathcal{T}$ und $\mathcal{T}$ Intervall in $\mathbb{R}$
- Stichprobe $x_i(t)$, $i = 1, \ldots, n$
- funktionaler Mittelwert:

$$\hat{\mu}_X(t) = \bar{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t)$$

- funktionale Varianz:

$$\hat{s}_X^2(t) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t) - \bar{x}(t)]^2$$

- funktionale (Auto-)kovarianz:

$$\hat{K}_X(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t_1) - \bar{x}(t_1)][x_i(t_2) - \bar{x}(t_2)]$$

Mittelwert, Varianz und (Auto-)kovarianz

- Betrachte funktionale Variable $X(t)$, mit $t \in \mathcal{T}$ und $\mathcal{T}$ Intervall in $\mathbb{R}$
- Stichprobe $x_i(t)$, $i = 1, \ldots, n$
- funktionaler Mittelwert:

$$\hat{\mu}_X(t) = \bar{x}(t) = n^{-1} \sum_{i=1}^{n} x_i(t)$$

- funktionale Varianz:

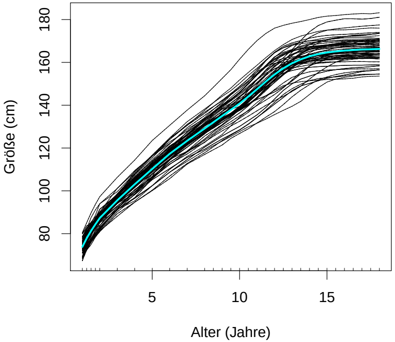
$$\hat{s}_X^2(t) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t) - \bar{x}(t)]^2$$

- funktionale (Auto-)kovarianz:

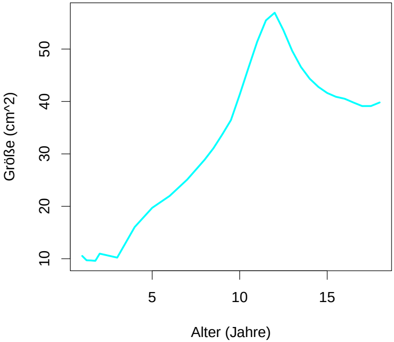
$$\hat{K}_X(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t_1) - \bar{x}(t_1)][x_i(t_2) - \bar{x}(t_2)]$$

Beispiel Mittelwert und Varianz

Wachstumskurven von 54 Mädchen

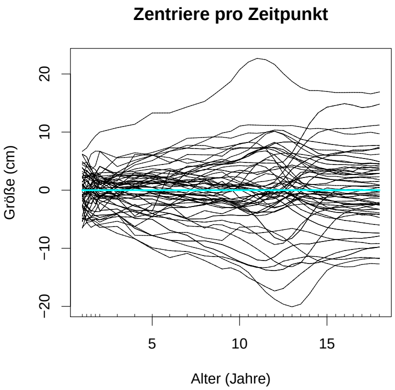


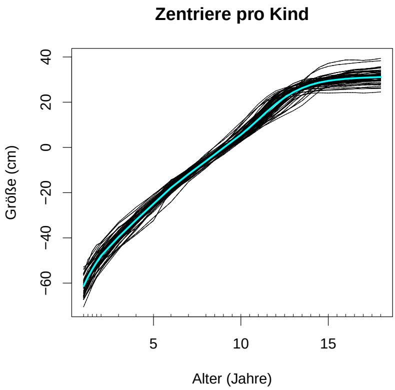
geschätzte Mittelwerts*funktion*



geschätzte Varianz*funktion*

Beispiel Mittelwert: Zwei Arten der Zentrierung



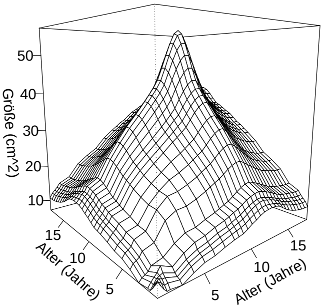


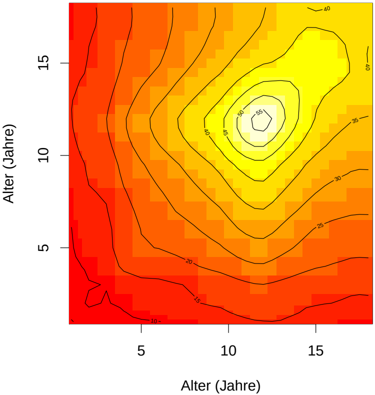
$$x_i^*(t) = x_i(t) - \bar{x}(t)$$

$$\tilde{x}_i(t) = x_i(t) - \int x_i(t)dt$$

Beispiel Kovarianzoberfläche

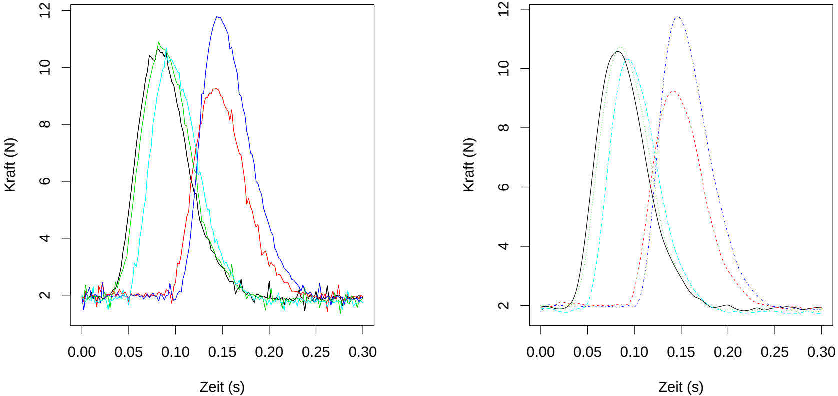
$$\hat{K}_X(t_1, t_2) = (n-1)^{-1} \sum_{i=1}^{n} [x_i(t_1) - \bar{x}(t_1)][x_i(t_2) - \bar{x}(t_2)]$$





Glättung

- funktionale Variable $X(t)$ mit Fehler beobachtet:
  $y_i(t) = X_i(t) + \epsilon_i(t)$
- Ziel: Finde die zugrundeliegenden glatten Funktionen $X_i(t)$



Pinch force (Ramsay und Silverman 2005)

Glättung

- Vielzahl an Methoden zur Glättung, zum Beispiel
  - Splines (P-Splines, Glättungssplines)
  - lokale Glätter (Nächste-Nachbarn, Running Mean/Median, Loess)
  - funktionale Hauptkomponentenanalyse (fPCA)

- Bias-Varianz-Trade-Off / Konflikt zwischen Datentreue und Glattheit

Glättung

- Vielzahl an Methoden zur Glättung, zum Beispiel
  - Splines (P-Splines, Glättungssplines)
  - lokale Glätter (Nächste-Nachbarn, Running Mean/Median, Loess)
  - funktionale Hauptkomponentenanalyse (fPCA)

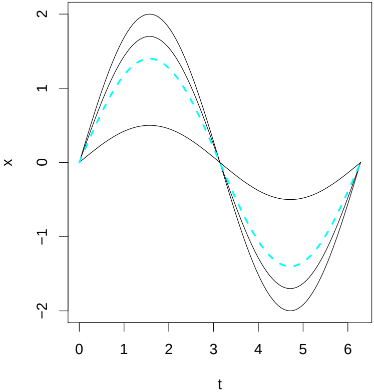
- Bias-Varianz-Trade-Off / Konflikt zwischen Datentreue und Glattheit

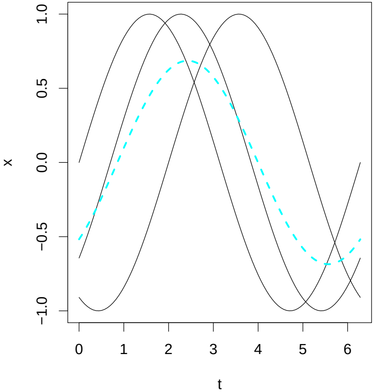
Registrierung: Motivation

Zwei Arten von Variation

- Amplitudenvariation,
  'Variation in y-Richtung'

- Phasenvariation,
  'Variation in x-Richtung'





- meist Interesse an Amplitudenvariation

Registrierung

- Vermeidung der Vermischung der beiden Variationsquellen durch Transformation der Funktionsargumente

- Vielzahl an Methoden zur Registrierung, zum Beispiel
  - Shift Registrierung:
    - einfache lineare Transformation der t-Variable
    - gesucht $x_i^*(t) = x_i(t + \delta_i)$
    - finde δ_i, sodass die vertikale Abweichung aller Kurven von der geschätzten Mittelwertkurve minimal ist
  - Landmark Registrierung:
    - Landmarks sind spezielle Charakteristika von Kurven, z.B. Extrema, Nullstellen
    - Kurven werden an Landmarks ausgerichtet
    - verwende streng monotone Transformationsfunktion $h_i(t)$

Registrierung

- Vermeidung der Vermischung der beiden Variationsquellen durch Transformation der Funktionsargumente
- Vielzahl an Methoden zur Registrierung, zum Beispiel
  - Shift Registrierung:
    - einfache lineare Transformation der t-Variable
    - gesucht $x_i^*(t) = x_i(t + \delta_i)$
    - finde δ_i, sodass die vertikale Abweichung aller Kurven von der geschätzten Mittelwertkurve minimal ist
  - Landmark Registrierung:
    - Landmarks sind spezielle Charakteristika von Kurven, z.B. Extrema, Nullstellen
    - Kurven werden an Landmarks ausgerichtet
    - verwende streng monotone Transformationsfunktion $h_i(t)$

Registrierung

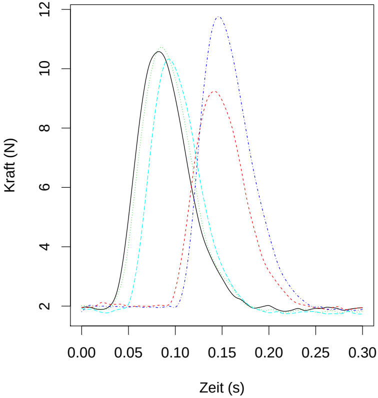
- Vermeidung der Vermischung der beiden Variationsquellen durch Transformation der Funktionsargumente

- Vielzahl an Methoden zur Registrierung, zum Beispiel
  - Shift Registrierung:
    - einfache lineare Transformation der t-Variable
    - gesucht $x_i^*(t) = x_i(t + \delta_i)$
    - finde δ_i, sodass die vertikale Abweichung aller Kurven von der geschätzten Mittelwertkurve minimal ist

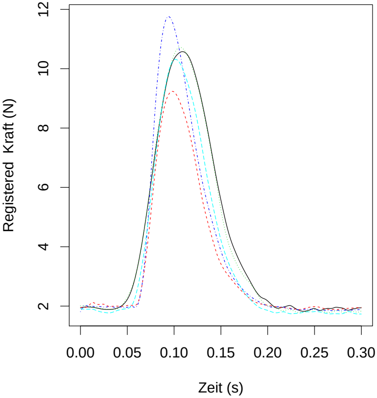
  - Landmark Registrierung:
    - Landmarks sind spezielle Charakteristika von Kurven, z.B. Extrema, Nullstellen
    - Kurven werden an Landmarks ausgerichtet
    - verwende streng monotone Transformationsfunktion $h_i(t)$

Beispiel Shift Registrierung

Pinch force (Ramsay und Silverman 2005)



Geglättete Daten



Registrierte Daten

Literatur

Cederbaum, J., Pouplier, M., Hoole, P. & Greven, S. (2015). 'Functional Linear Mixed Models for Irregularly or Sparsely Sampled Data.' *Submitted.*

Fuchs, K., Scheipl, F. & Greven, S. (2015). 'Penalized scalar-on-functions regression with interaction term.' *Computational Statistics & Data Analysis*, **81**, 38–51.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, Springer, New York.