Funktionale Hauptkomponentenanalyse

Philipp Lintl

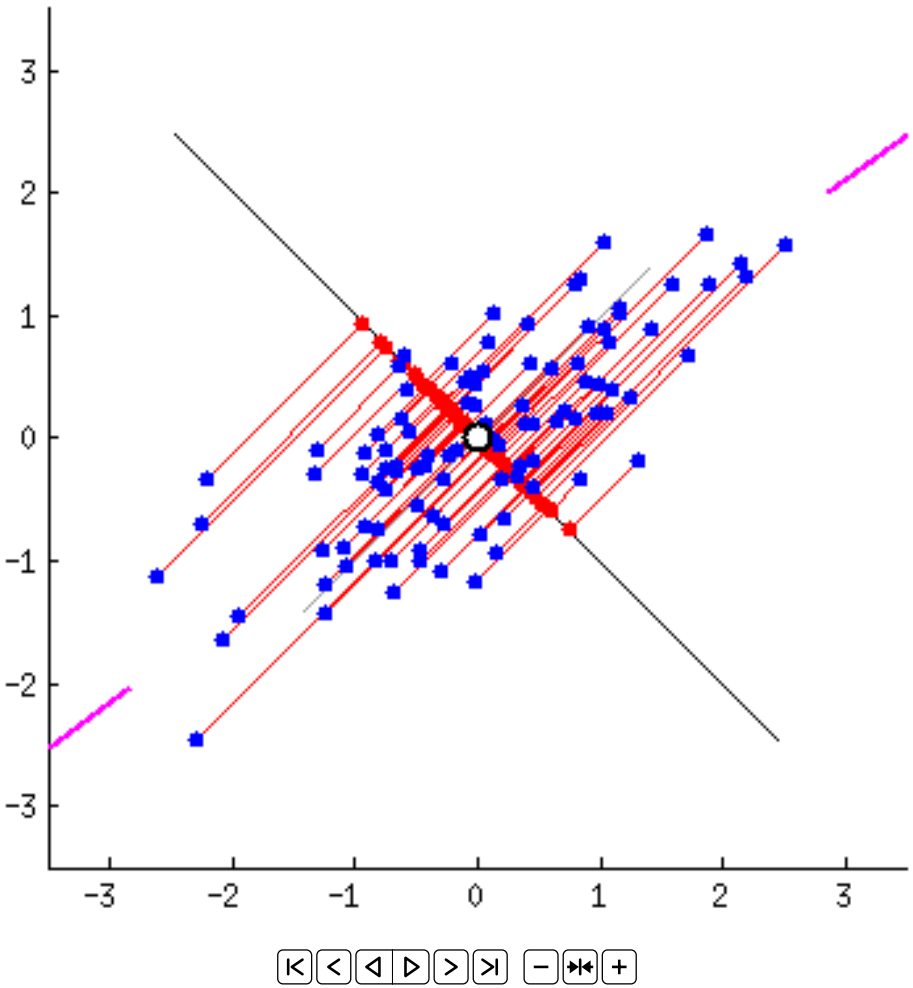
12. Januar 2018

Inhaltsverzeichnis

Hauptkomponentenanalyse

- umfangreiche Datensätze zu strukturieren, zu vereinfachen und zu veranschaulichen
- Grundidee: Datenreduktion bzw. Dimensionsreduktion
- betrachtete Daten auf möglichst wenige Hauptkomponenten reduzieren, ohne zu großen Informationsverlust
- $\rightarrow$ Hauptkomponenten sollen einen möglichst großen Teil der Varianz der Daten erklären
- $\Rightarrow$ funktional: wichtigste Arten der Variabilität / Strukturen in Daten

Beispiel



- Variablen x_1,x_2
- gesucht: Linearkombination $\boldsymbol{\alpha}^T \mathbf{x} = \alpha_1 x_1 + \alpha_2 x_2$ mit maximaler Varianz
- HK1: Gerade, bei der die Summe der Fehlerquadrate minimal ist (rote Linien)
- HK2: zu HK1 ortogonal, wieder maximierte Varianz

Für multivariate Daten

- Datenmatrix: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T, \quad \mathbf{X} \in \mathbb{R}^{Nxp}$
- N Beobachtungen, p Variablen
- zentrierte Daten: $\tilde{x}_{ij} = x_{ij} - \frac{1}{N} \sum_{i=1}^{N} x_{ij}$
- Varianz: $\widehat{Var}(\tilde{x}_j) = \frac{1}{N-1} \sum_{i=1}^{N} \tilde{x}_{ij}^2$
- Kovarianz: $\widehat{Cov}(\tilde{x}_j, \tilde{x}_k) = \frac{1}{N-1} \sum_{i=1}^{N} \tilde{x}_{ij} \tilde{x}_{ik}$
- Kovarianzmatrix: $\mathbf{V} = \frac{1}{N-1} \boldsymbol{X}^T \mathbf{X}, \quad \mathbf{V} \in \mathbb{R}^{pxp}$

- Idee: Varianz in $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ durch unkorrellierte Variablen $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^T$ beschreiben

- $\boldsymbol{\xi}$ als Linearkombination der originalen Variablen und den **Gewichten** $\boldsymbol{\phi}$

$$\xi_1 = \sum_{j=1}^{p} \phi_{j1} x_j$$

$$\xi_1 = \boldsymbol{\phi}_1^T \mathbf{x}$$

- Schrittweise finden der derjenigen **Hauptkomponenten** $\boldsymbol{\phi_i}$, die Varianz der $\boldsymbol{\xi_i}$ **maximieren**

- 1. Hauptkomponente: Der Gewichtsvektor $\boldsymbol{\phi_1} = (\phi_{11}, \ldots, \phi_{p1})^T$ für den

$$\xi_{i1} = \sum_j \phi_{j1} x_{ij} = \boldsymbol{\phi_1^T} \mathbf{x}_i, \quad i = 1, \ldots, N, j = 1, \ldots, p$$

- Varianz **maximal**: $Var(\boldsymbol{\xi}_1) = \frac{1}{N-1} \sum_i {\xi_{i1}}^2 \rightarrow max$

- $\Rightarrow$ erklärt stärkste Art der Variabilität in den Daten

- Eindeutigkeit durch Normierung: $\|\boldsymbol{\phi_1}\|^2 = \sum_j \phi_{j1}^2 = 1$

- m-te HK: Gewicht ϕ_m mit

$$Var(\xi_m) \to max, \quad \|\phi_m\|^2 = 1$$

- und m-1 zusätzlichen Bedingungen

$$\langle \phi_k, \phi_m \rangle = \phi_k^T \phi_m = 0, \quad k < m.$$

- zueinander orthogonale Hauptkomponenten $\to$ jede HK erklärt Neues
- zu jedem Schritt Varianzmaximierung: insgesamt sinkende erklärte Varianz
- $\Rightarrow$ Oft Großteil der Varianz schon durch wenige HK

Lösung des Optimierungsproblems

- wird Optimierungsproblem zu

$$max\ Var(\boldsymbol{\xi}) = max\ \frac{1}{N-1}\sum(\phi^T \boldsymbol{x})^2 = max\ \frac{1}{N-1}\phi^T \boldsymbol{X}^T \boldsymbol{X}\phi$$

$$= max\ \phi^T \mathbf{V}\phi \quad mit\ NB: \|\phi\|^2 = 1,$$

- $\Rightarrow$ Suche nach größten Eigenwerten aus

$$\boldsymbol{V}\phi = \lambda\phi$$

- mit (λ_m, ϕ_m) Eigenwert-Eigenvektor Paare, $Var(\boldsymbol{\xi_m}) = \lambda_m$ m-größter Eigenwert

Hauptkomponentenanalyse für funktionale Daten

Für funktionale Daten

- Zufallsstichprobe reellwertiger Funktionen $x_1(t), \ldots, x_N(t)$ auf Intervall $\mathcal{T} = [0, T]$
- Individuell: Realisierungen eines eindimensionalen stochastischen Prozesses $X = X(t)$
- Wieder zentrierte Daten: $x_i(t) = \tilde{x}_i(t) - \frac{1}{N} \sum_{j=1}^{N} \tilde{x}_j(t)$
- unendlich dimensionale funktionale Daten $\xrightarrow{FPCA}$ endlich dimensionale Darstellung

	PCA	FPCA
Variablen	$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\mathbf{x}_i = (x_{1i}, \ldots, x_{Ni})$, $i = 1, \ldots, p$	$(x_1(t), \ldots, x_N(t))$, $t \in [0, T]$
Daten	Vektoren $\in \mathbb{R}^p$	Kurven $\in L_2(\mathcal{T})$
Mittelwert	$\mu = \mathbb{E}(X)$	$\mu(t) = \mathbb{E}(X(t))$
Kovarianz	$Cov(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{V}_{jk}$	$Cov(x(s), x(t)) = \gamma(s, t)$

- Kovarianzfunktion:

$$\widehat{\gamma}(s,t) = \frac{1}{N-1}\sum_{i=1}^{N} x_i(s)x_i(t)$$

- Multivariate Linearkombination:

$$\langle \boldsymbol{\phi}, \mathbf{x} \rangle = \sum_{j=1}^{p} \boldsymbol{\phi}_j \mathbf{x}_j \quad \xrightarrow{\text{wird zu}} \quad \langle \phi, x \rangle = \int_{\mathcal{T}} \phi(t)x(t)dt$$

- $\Rightarrow$ Gewichts-/Hauptkomponentenfunktionen $\phi(t)$
- $\Rightarrow$ Hauptkomponentenscores

$$\xi_{i1} = \langle \phi_1, x_i \rangle = \int \phi_1(t)x_i(t)dt.$$

- *Karhunen-Loeve Erweiterung* zentrierter Daten:

$$x(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t)$$

- Dimensionsreduktion, wenn

$$x(t) \approx \sum_{j=1}^{k} \xi_j \phi_j(t)$$

- gute Approximation der unendlichen Summe

- gleiche schrittweise Prozedur:
- 1. $\phi_1(t)$ aus:

$$Var(\boldsymbol{\xi}_1) = \frac{1}{N-1} \sum_{i=1}^{N} (\xi_{i1})^2 \rightarrow \ max$$

- Bedingung:

$$\|\phi_1\|^2 = \langle \phi_1, \phi_1 \rangle = \int \phi_1(t)^2 dt = 1$$

- 2. weitere HK durch maximale Varianz der m-ten HK
- und zusätzlich

$$\int \phi_k(t)\phi_m(t)dt = 0, k < m$$

- Gesuchte Gewichtsfunktionen $\phi(t)$ lösen:
- **funktionales Eigenwertproblem**:

$$\int \hat{\gamma}(s,t)\phi(t)dt = \lambda\phi(s)$$

- Kovarianzoperator $\mathbf{\Gamma}$ einer Funktion ϕ:

$$\mathbf{\Gamma}\phi(s) = \int \hat{\gamma}(s,t)\phi(t)dt$$

- Eigenwertproblem also wieder der Form:

$$\mathbf{\Gamma}\phi = \lambda\phi$$

- ϕ nun Eigenfunktionen und $V(\xi_m) = \lambda_m$ m-größter Eigenwert von $\mathbf{\Gamma}$.

- Unterschied zum multivariaten Fall: Anzahl mgl. Eigenwert-Eigenfunktionspaare

- theoretisch: max $\#\textit{Eigenfunktionen} = \#\textit{Funktionswerte}\ x(t)$
  $\Rightarrow$ unbegrenzt

- In der Praxis: Basisdarstellung der Funktionen $x_i(t)$:

$$\hat{x}_i(t) = \sum_{m=1}^{k} c_{im} v_m(t)$$

  gemäß bekannter Basisfunktionen $v_m(t)$ (Spline, Fourier,...)

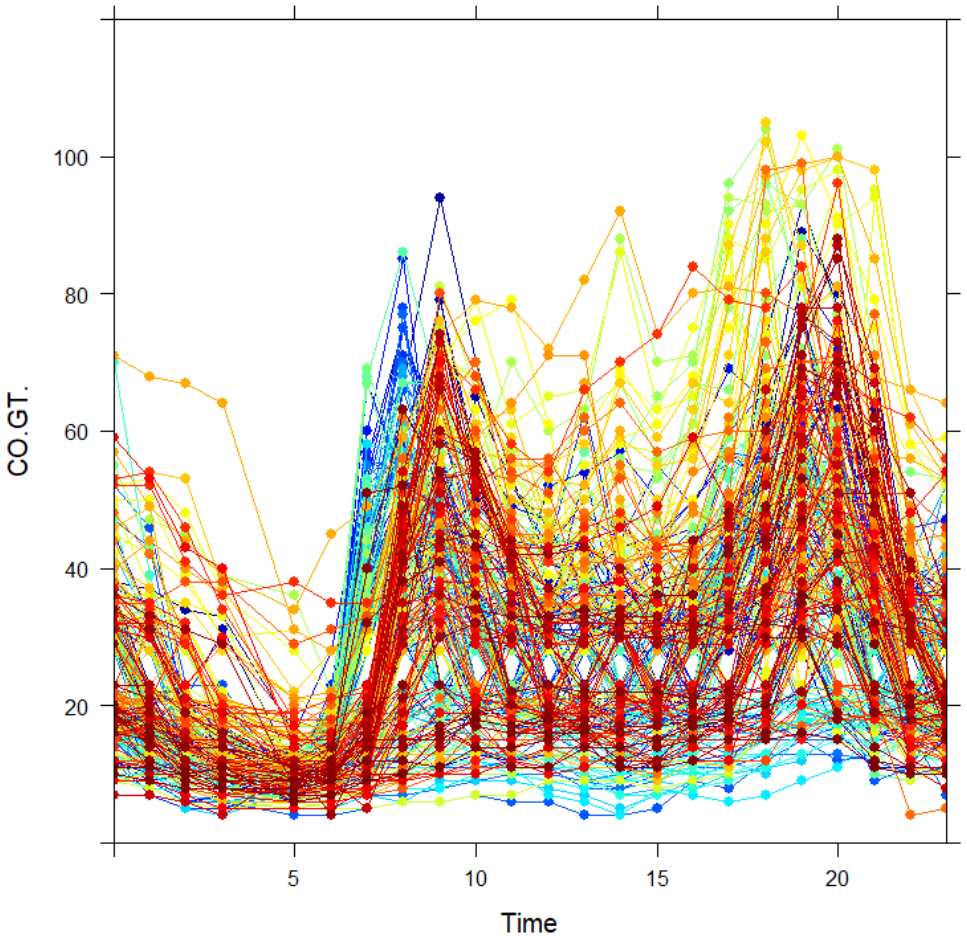
Lösung Eigenwertproblem: Diskretisierung

- diskretisieren gefitteter $x_i \in L^2(\mathcal{T})$ auf Gitter mit K Werten gleichen Abstands
- neue Datenmatrix $\mathbf{X} \in \mathbb{R}^{NxK}$
- $\Rightarrow$ multivariate Hauptkomponentenanalyse
- neues Eigenwertproblem $\Rightarrow$ neue Eigenvektoren $\xrightarrow{\textit{Rücktransformation}}$ Funktionen

Anwendung auf Luftverschmutzungsdaten

Datensatz

- Luftverschmutzung einer italienischen Stadt
- stündliche Mittelwerte CO-Konzentration $[mg/m^3]$
- Zeitraum von 03.2004 - 04.2005
- eine Beobachtung: 0-23 Uhr (4 Uhr oft NA -> entfernt)
- 282 Beobachtungen zur Analyse
- Quelle: University of California, Irvine

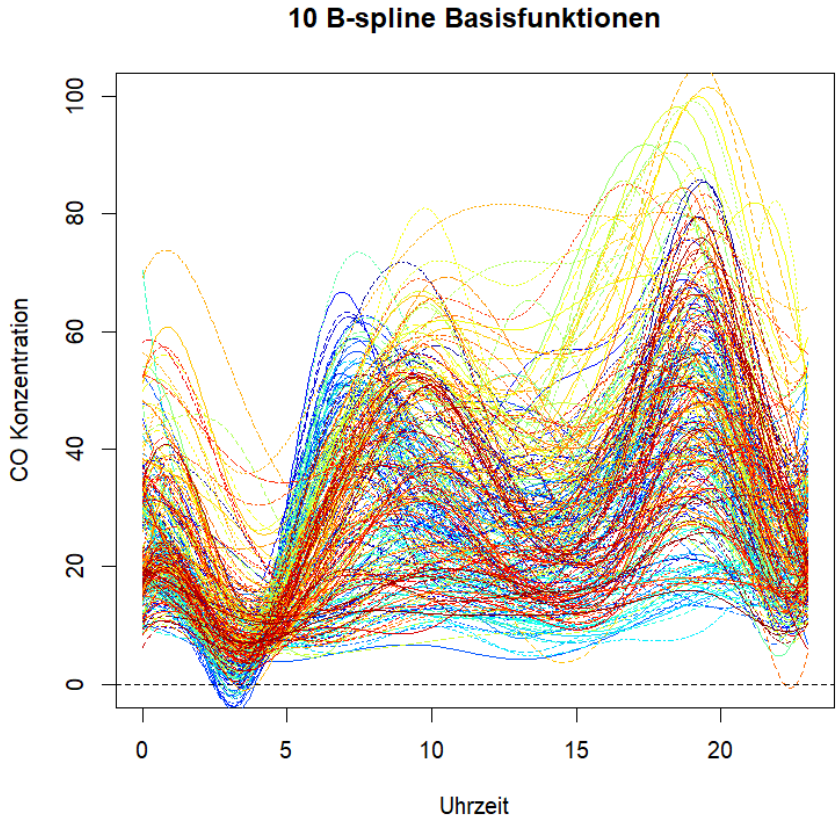


- bimodal (ca. 9 Uhr und 20 Uhr)
- An welchen Stellen liegt nun größte Variabilität in den Daten vor?

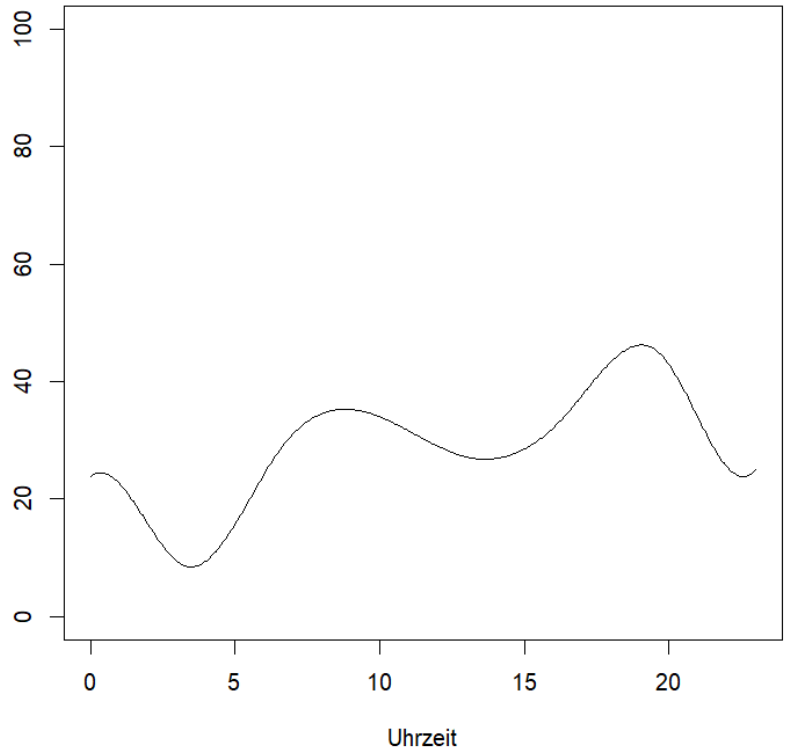
  $\Rightarrow$ Hauptkomponenten

Visualisierung

- Daten gemäß 10 Basisfunktionen



- Mittelwertsfunktion $\hat{\mu}(t)$



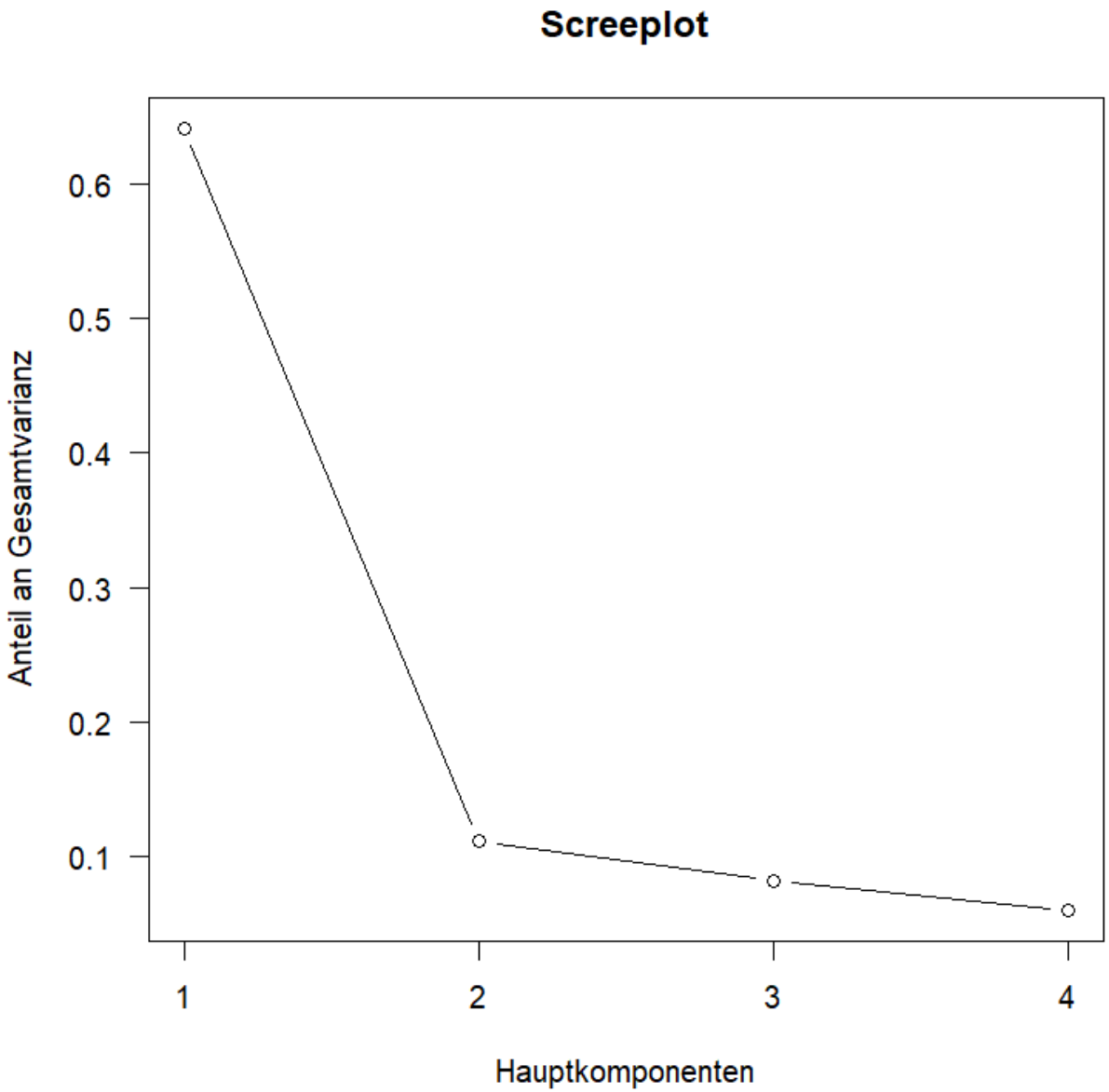
Wahl der Anzahl an Hauptkomponenten

- intuitiv: wähle so viele HK, bis t% der Gesamtvarianz in den Daten erklärt
- t meist zwischen 70% und 100%
- wegen $Var(\xi_m) = \lambda_m$, gilt $\sum_k^\infty \lambda_k = p$, mit Varianz der k-ten HK λ_k und Gesamtvarianz in den Daten p

- daher entfallen auf ersten m Hauptkomponenten

$$t_m = 100 \frac{\sum_{k=1}^m \lambda_k}{p}$$

- Sobald $t_k \geq t$ Anzahl k an Hauptkomponenten gefunden

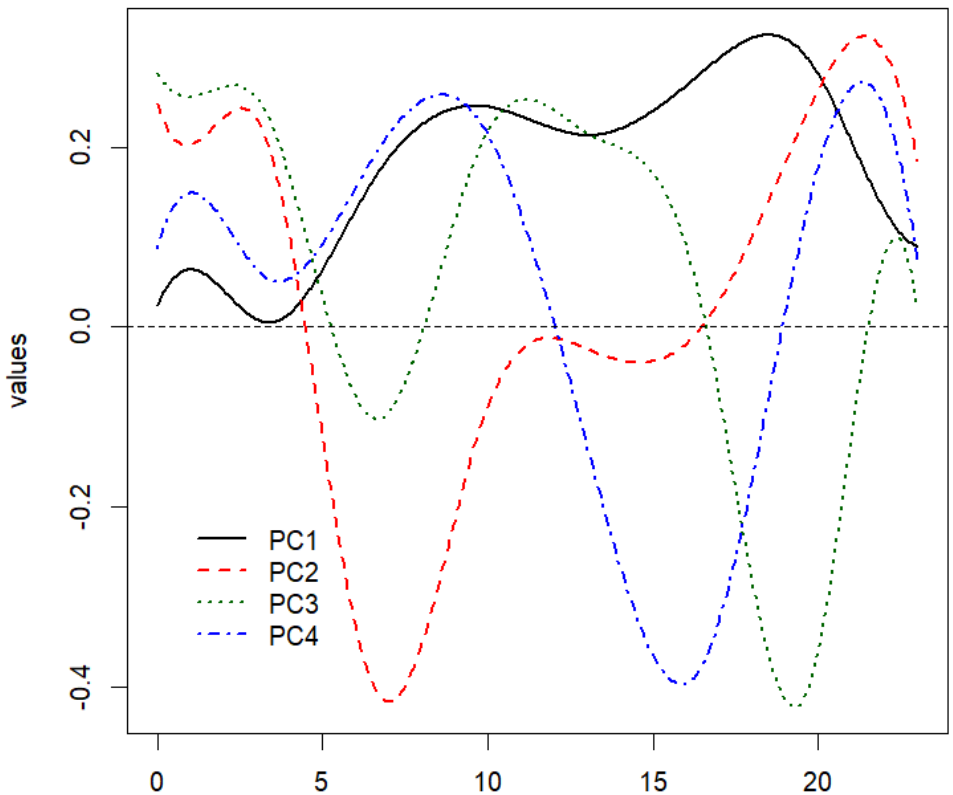


- HK 1: meiste Varianz

  Scree: Nur die 1. HK betrachten
- Zur Veranschaulichung: 4 Hauptkomponenten

Visualisierung: Gewichtsfunktionen

- Gewichtsfunktionen $\phi_1, \ldots, \phi_4$



- HK 1: Ähnlich zu $\hat{\mu}(t)$

  6 - 21 Uhr stark gewichtet
  $\rightarrow$ größte Variabilität

  Luftverschmutzungsverlauf
- HK 2:

  Nacht- & Abendstunden positiv

  $\rightarrow$ zweite Art der Variabilität

  Verschmutzung Nachts und Abends

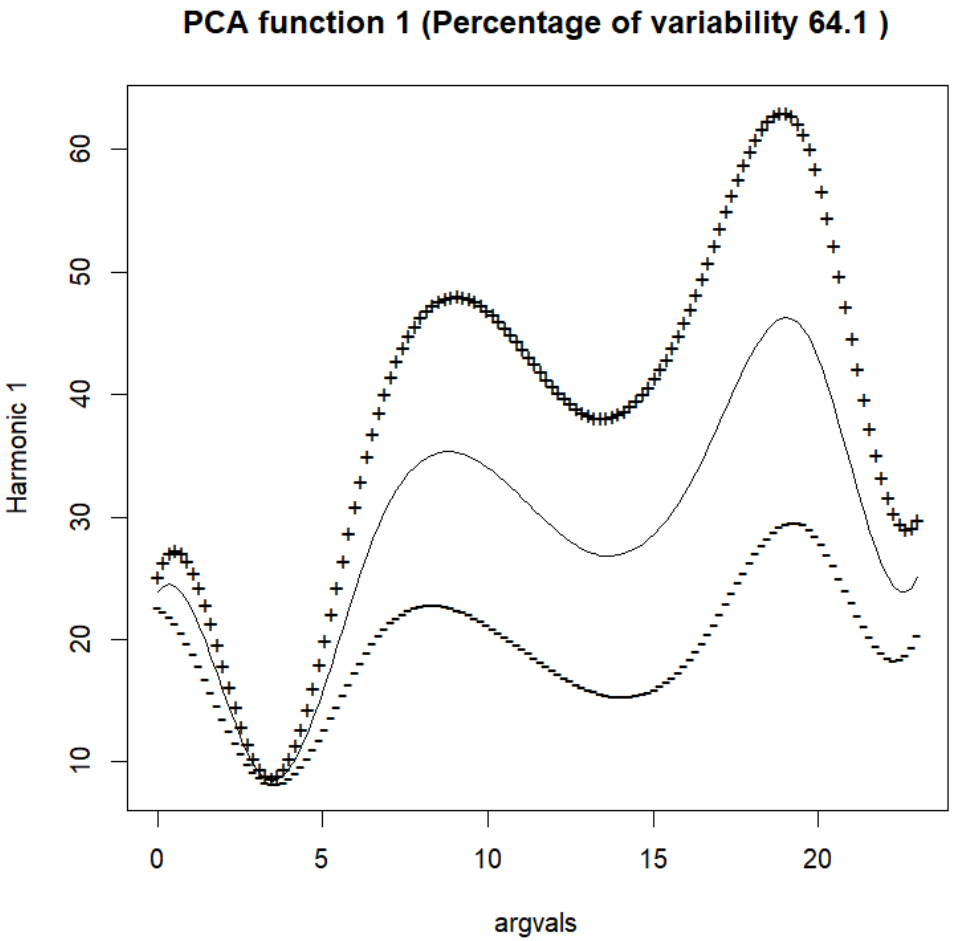
- *Karhunen-Loeve* nicht zentrierter Daten:

$$x(t) \approx \hat{\mu}(t) + \sum_{j=1}^{k} \xi_j \phi_j(t)$$

- Idee: Neue Funktion $\tilde{x}$, mit Score-Vektor: $\tilde{\boldsymbol{\xi}} = (\pm c, 0, ..., 0)$
- Die Karhunen-Loeve Darstellung davon:

$$\tilde{x}(t) = \hat{\mu}(t) + \sum_{j=1}^{k} \tilde{\xi}_j \phi_j(t) = \hat{\mu}(t) \pm c\phi_1(t)$$

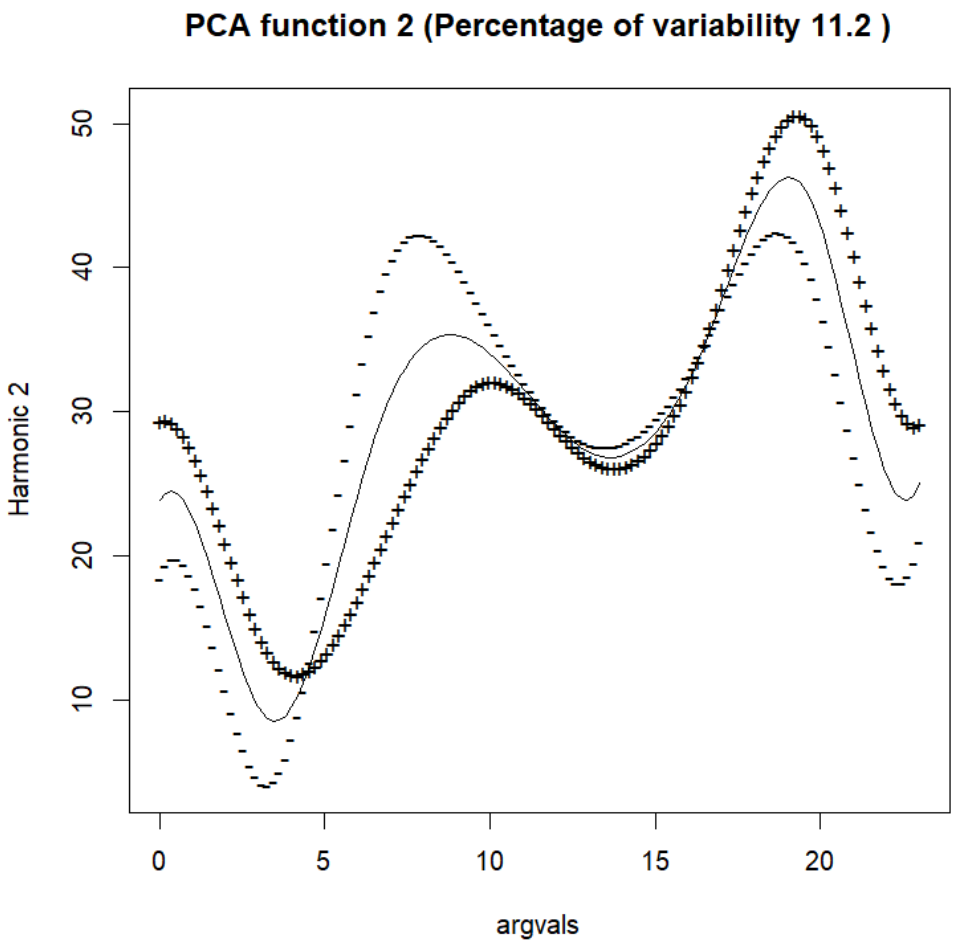
Visualisierung: Auswirkung auf Mittelwert



- HK 1: genereller Luftverschmutzungsverlauf
- Beobachtung mit hohem ξ_1:

  Hohe Verschmutzung (überdurchschnittlich) von 6-21 Uhr, besonders an Gipfeln
- niedriges ξ_1:

  geringe Luftverschmutzung (unterdurchschnittlich)
- $\Rightarrow$ Größte Art der Variabilität (64.1%)

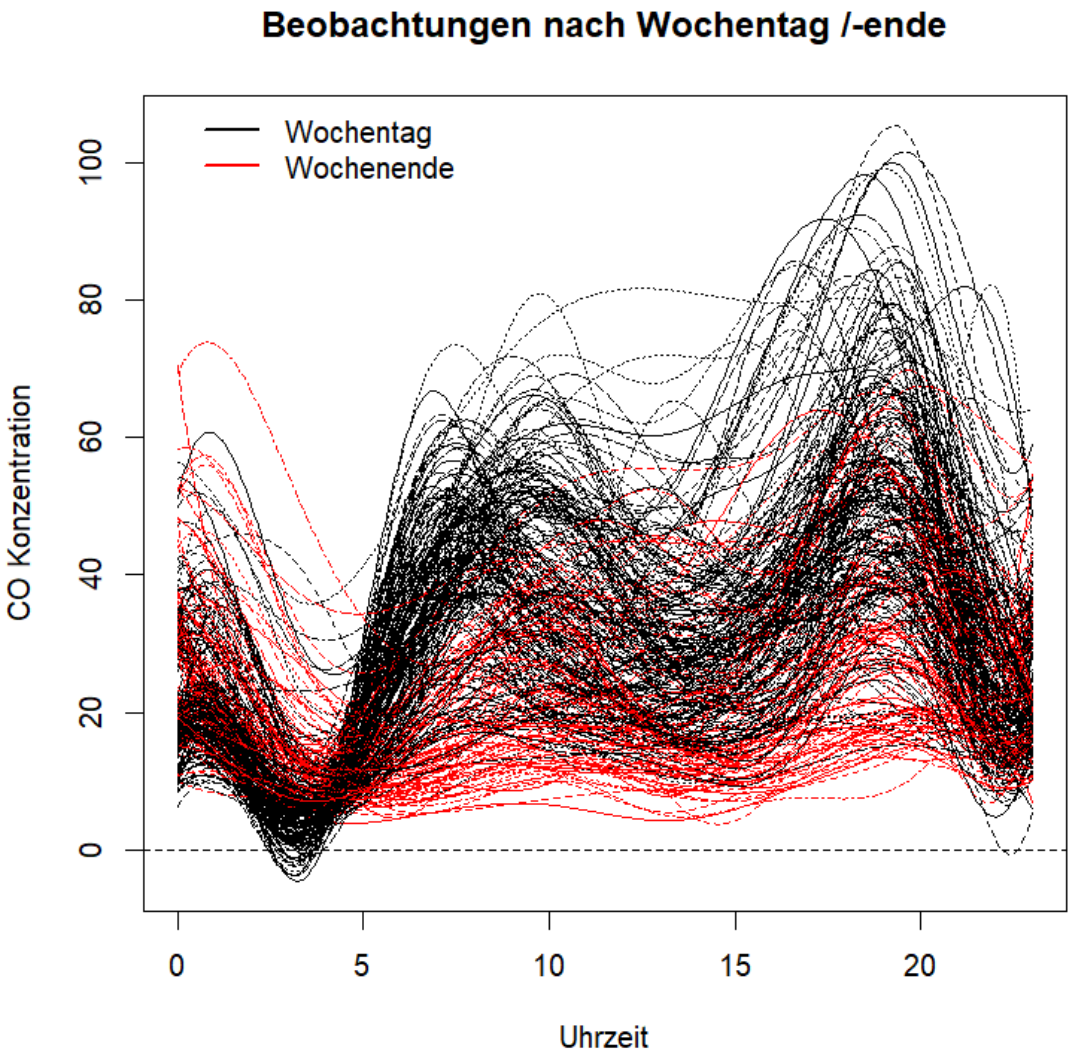


- HK 2 Variabilität nachts / spät
- Beobachtung mit hohem ξ_2:

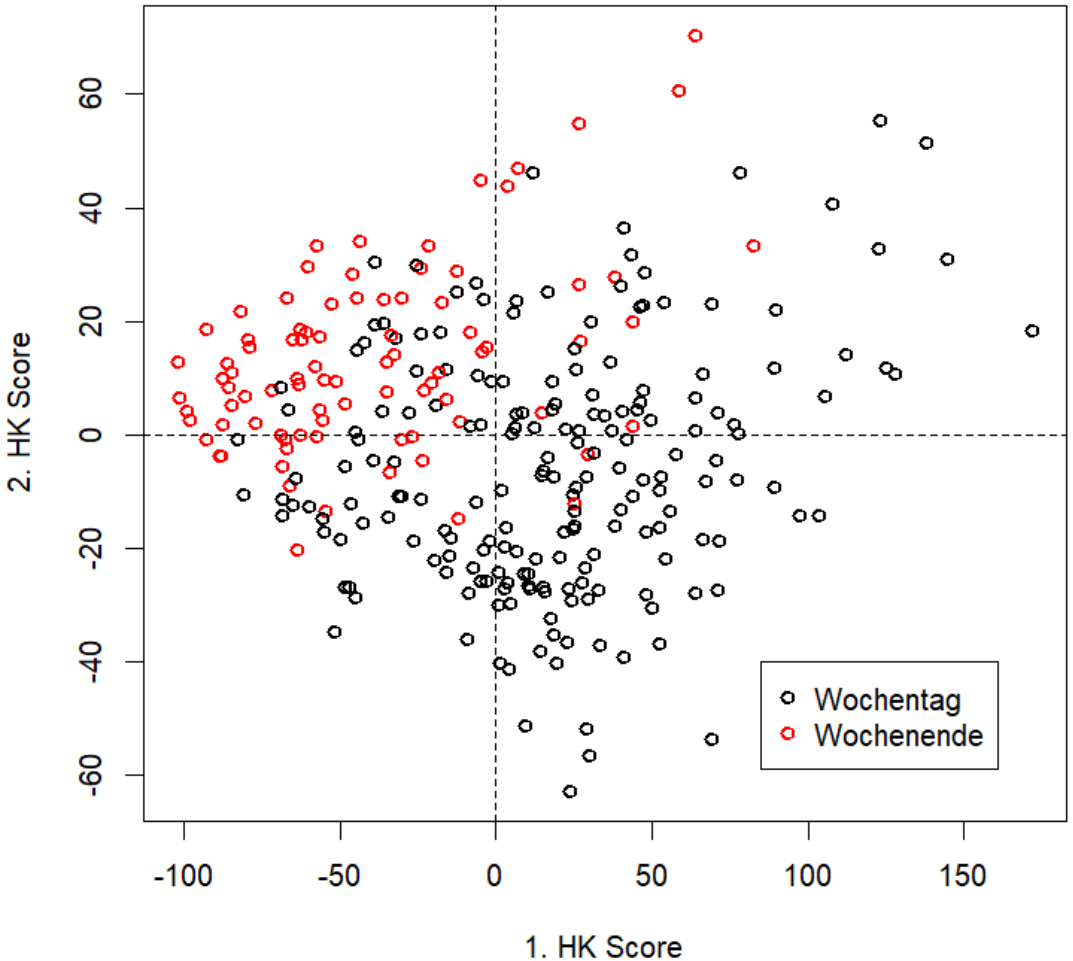
  Hohe Verschmutzung 0-5 und 16-23 Uhr, sonst niedrig
- niedriges ξ_2:

  hohes CO zwischen 5 und 15 Uhr, sonst niedrig
- $\Rightarrow$ Zweitgrößte Variation (11.2%)

Visualisierung der Hauptkomponenten: Scores



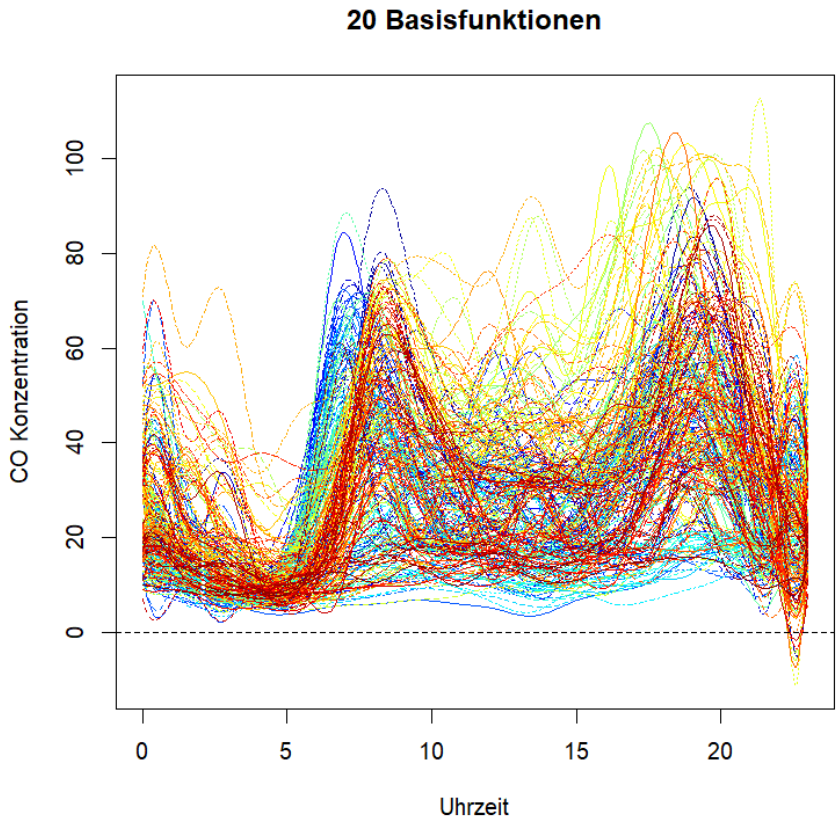
- Nachts-/Abendstunden teilweise höhere Belastung
- über den Tag verteilt allerdings unterdurchschnittlich
- erwartbar: kleines ξ_1 und hohes ξ_2

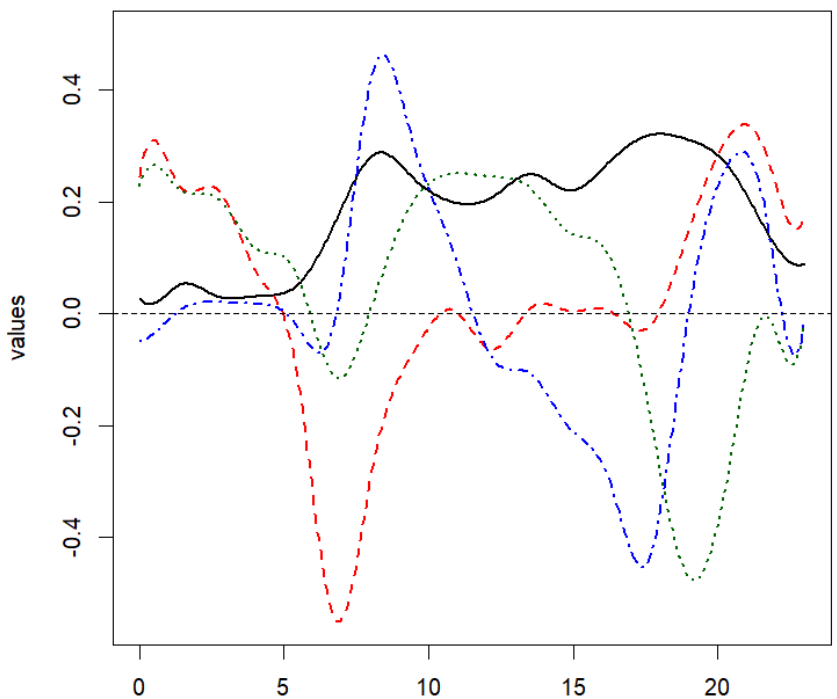


- Wochenendbeobachtungen vor allem links oben
- also niedrige generelle Verschmutzung
- höhere Belastung früh, spät und unterdurchschnittlich bei erstem Gipfel

Glättung der Hauptkomponenten

- Für rauhe Daten: 20 anstatt 10 Basisfunktionen
  $\rightarrow$ rauhe HK $\rightarrow$ schlecht interpretierbar





- Glattheitsanforderung für weitere Verwendung der Hauptkomponenten
- 2 Ansätze: Daten vor HKA glätten (Splines) vs. Hauptkomponenten glätten

- mittels **Penalisierungsterm** in der Hauptkomponentenanalyse
- Einführung Rauheitsmaß:

$$PEN_2(\phi) = \|D^2\phi\|^2 = \int \phi''(t)^2 dt$$

- 2. Ableitung kontrolliert Krümmung (Rauheit)

- zuvor: $\underset{\|\phi_m\|^2=1}{max}\ Var(\xi_m)$ gelöst durch λ, ϕ

- jetzt: zusätzliche Berücksichtigung der Rauheit $\Rightarrow$ Einführung Glattheitsparameter $\lambda \geq 0$

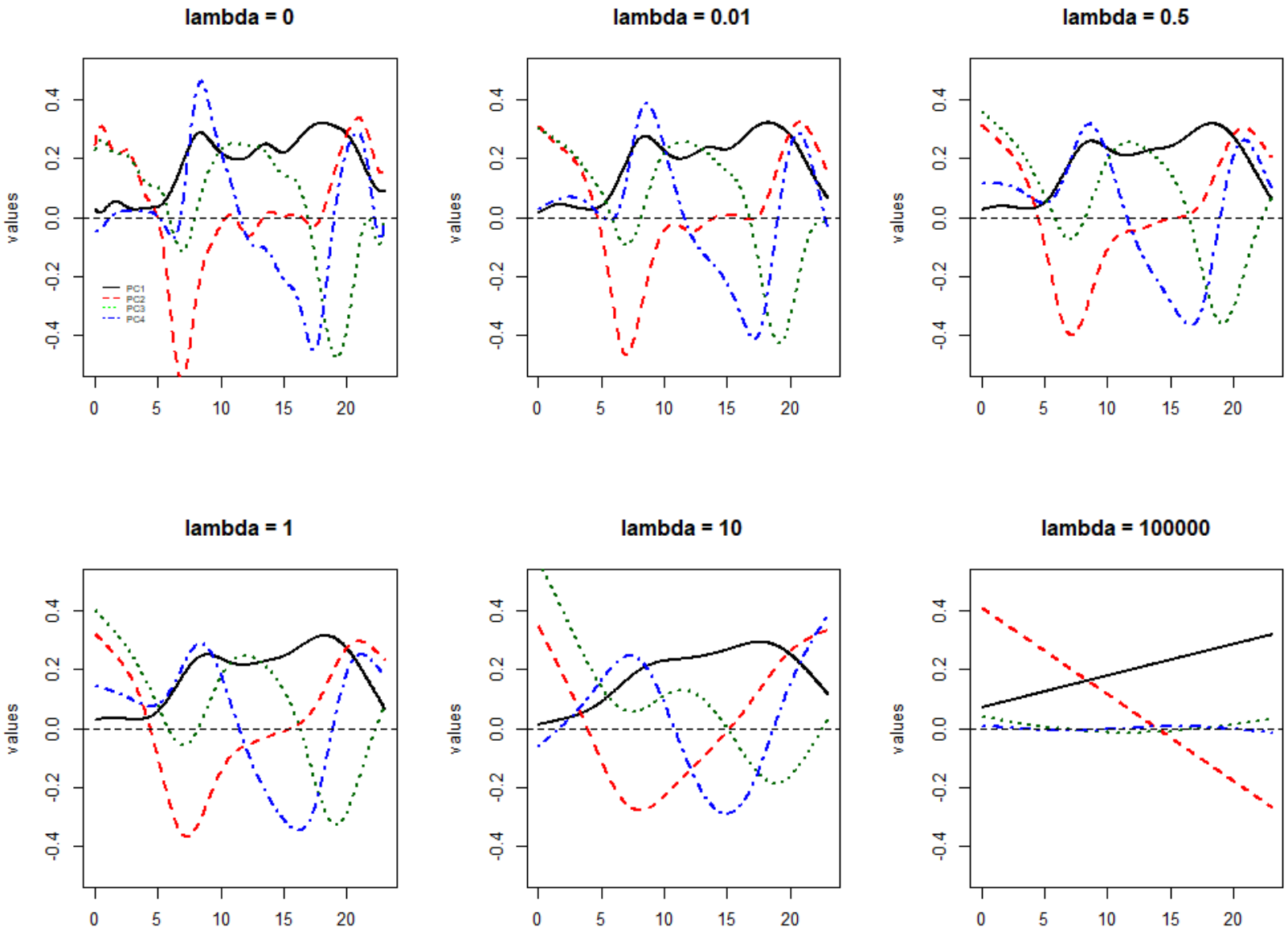
- **penalisierte Varianz**

$$PCAPSV(\xi) = \frac{Var(\int \phi x_i)}{\|\phi\|^2 + \lambda PEN_2(\phi)}$$

  - $\lambda \to 0 : PCAPSV(\xi) \to \frac{var(\int \phi x_i)}{1}$, gleiche Hauptkomponente wie zuvor
  - $\lambda \to \infty$, ergibt Konstante $\phi = a$ im periodischen Fall oder $\phi = a + bt$ im nichtperiodischen Fall

- Für optimales λ: Leave-one-out Kreuzvalidierung

Anwendung der Glättung



Zusammenfassung

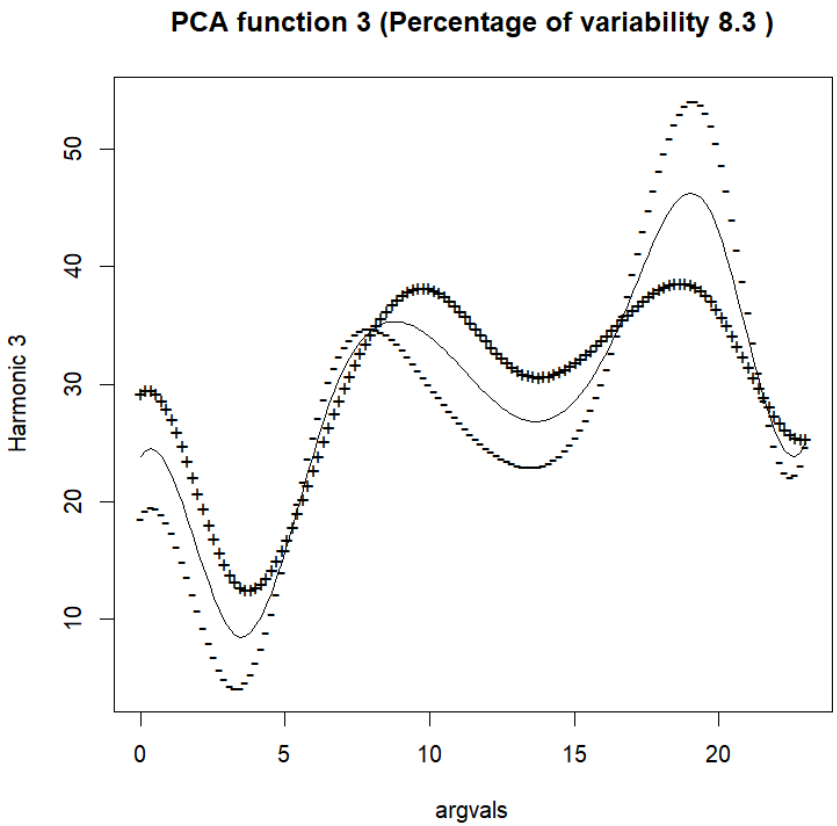
- PCA: beobachtete Variablen $\xrightarrow{\textit{Orthogonaltransformation}}$ linear unabhängige Hauptkomponenten (Dimensionsreduktion ohne zu großen Informationsverlust)
- FPCA: Explorative Methode zur Erkennung von Mustern und Variationsquellen in funktionalen Daten
- Anwendung: Meiste Variabilität in den Tagesverläufen (mit Peaks bei Hauptverkehrszeiten)
- Unterschied Werktage / Wochenendtage
- Glättung durch Basenwahl oder Rauheitsmaß (ähnliche Ergebnisse)

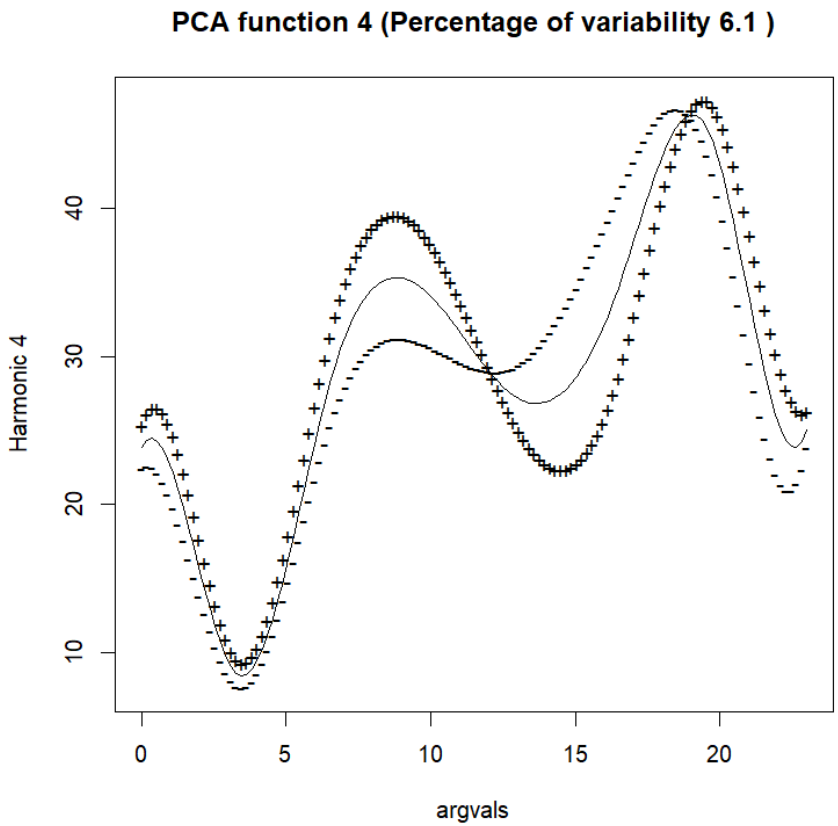
Referenzen

- J. O. Ramsay and B. W. Silverman. Functional Data Analysis. Springer, 2005.
- I. Jolliffe. Principal Component Analysis. Springer, 2 edition, 2002.
- J. Ramsay and B. Silverman. Applied Functional Data Analysis: Methods and Case Studies. Springer, 2002.
- Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2011). fda: Functional Data Analysis. R package version 2.2.6.
- J.-L.Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. Annual Review
- Datensatz: https://archive.ics.uci.edu/ml/datasets/Air+Quality#

Anhangsfolien

- Hauptkomponenten 3,4

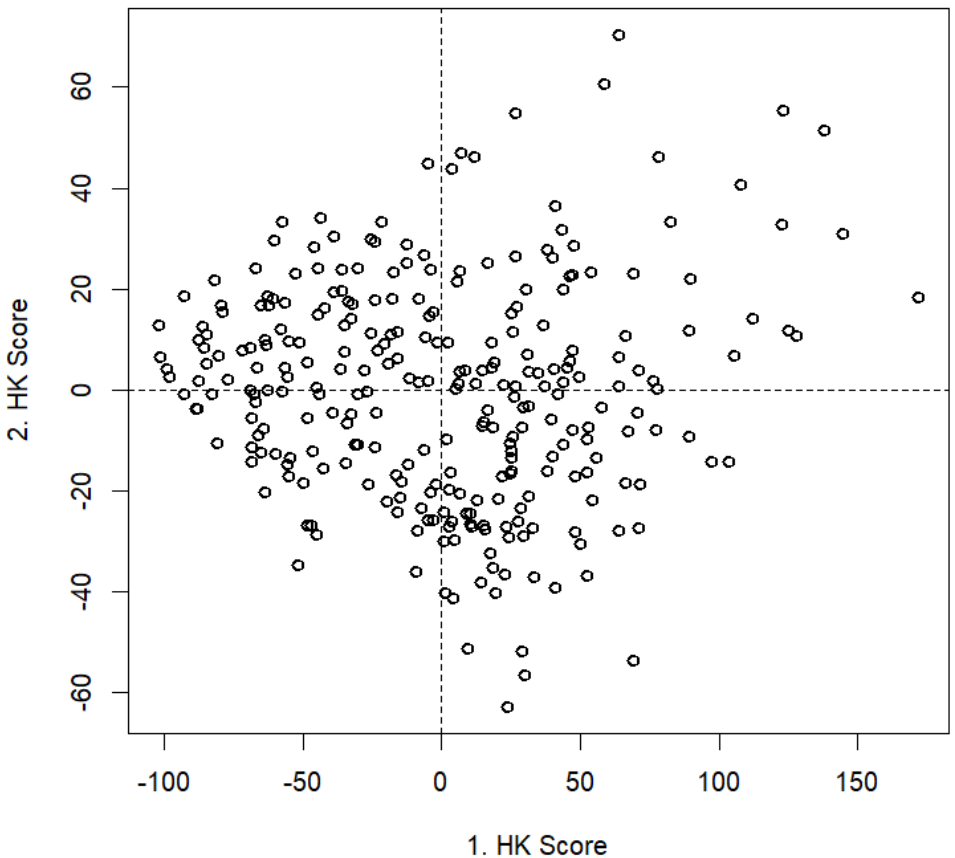




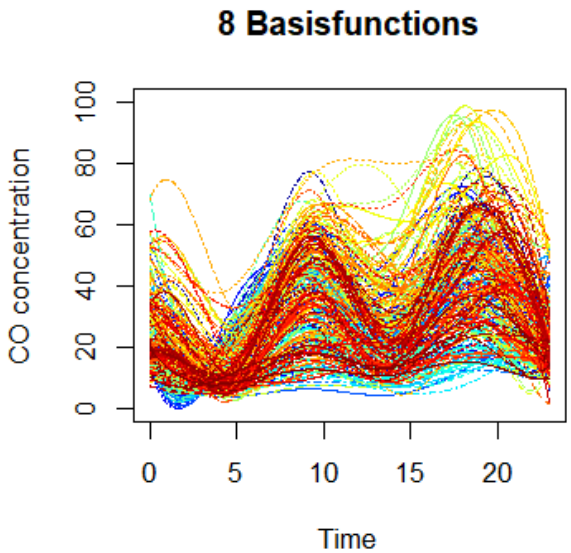
- Interpretation immer schwieriger

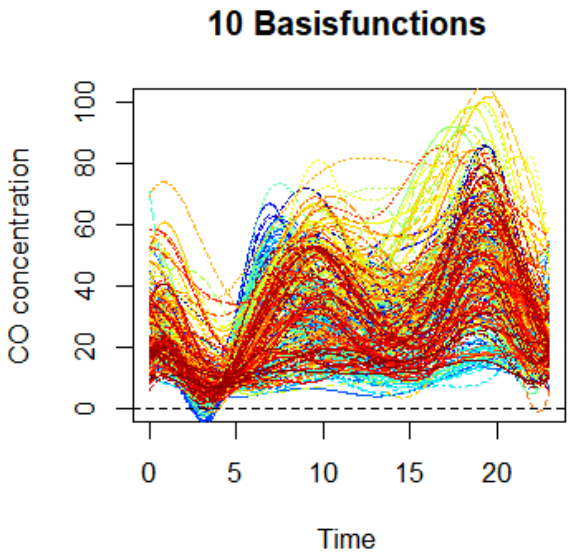
Visualisierung der Hauptkomponenten: Scores

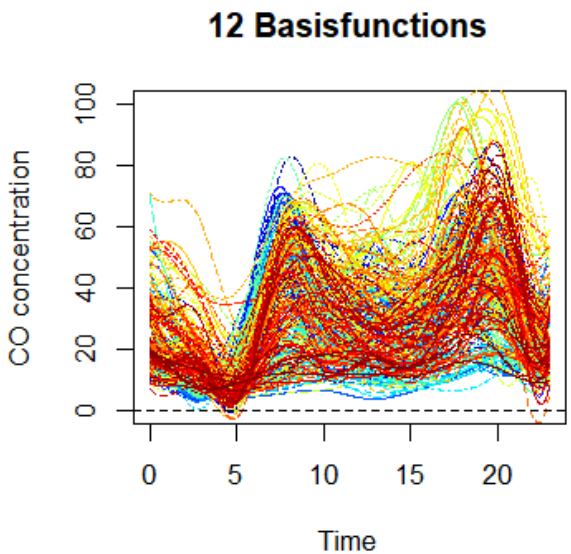
- Hauptkomponentenscores: $\xi_1 vs. \xi_2$

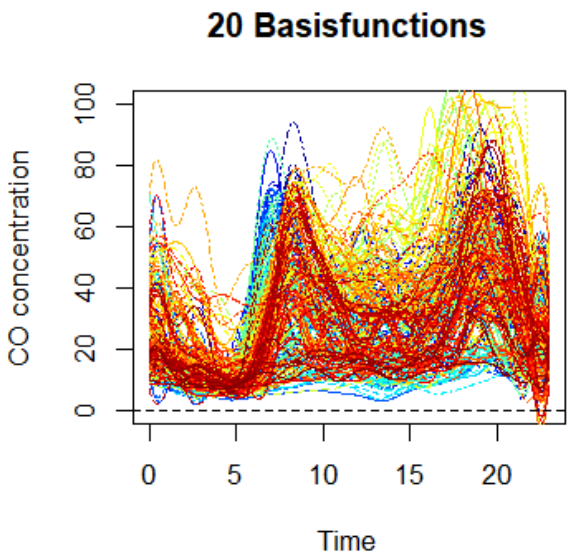


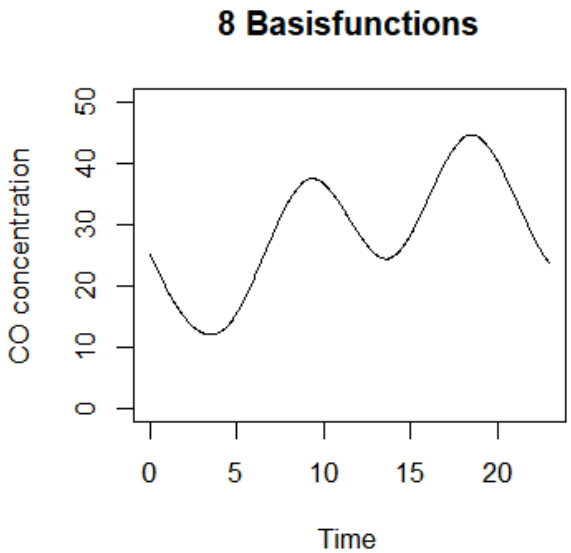
- tendenziell gleich auf Bereiche verteilt
- großes $\xi_1 \rightarrow$ großes ξ_2
- Kurven mit hoher Verschmutzung auch früh und spät stark verschmutzt
- $\rightarrow$ eher schwer interpretierbar
- mgl. Unterschiede zwischen Wochenende/Arbeitstage

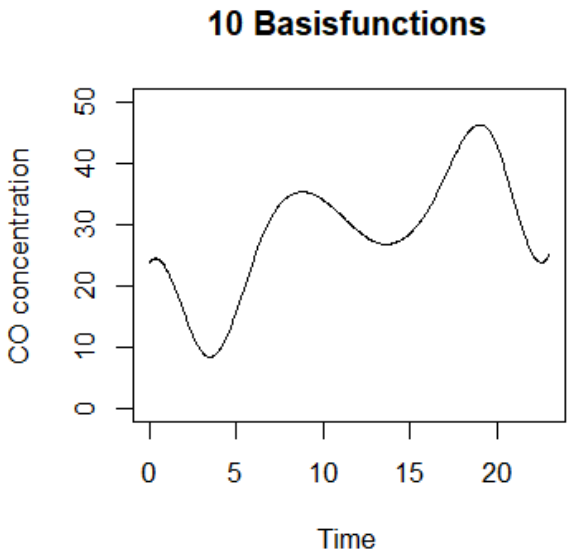

8 Basisfunctions
CO concentration
0
20
40
60
80
100
0
5
10
15
20
Time

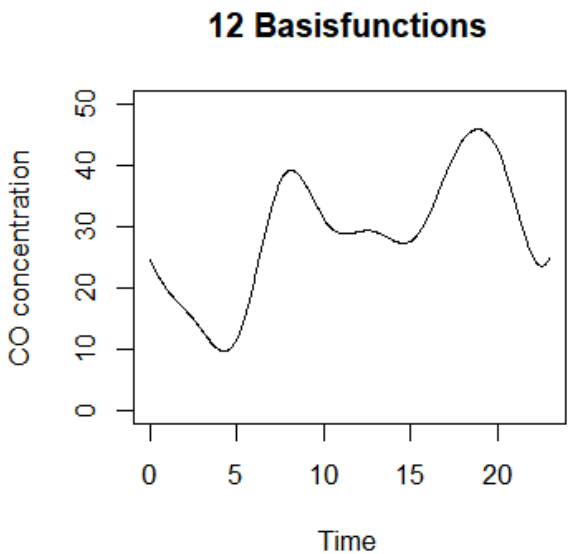

10 Basisfunctions
CO concentration
0
20
40
60
80
100
0
5
10
15
20
Time

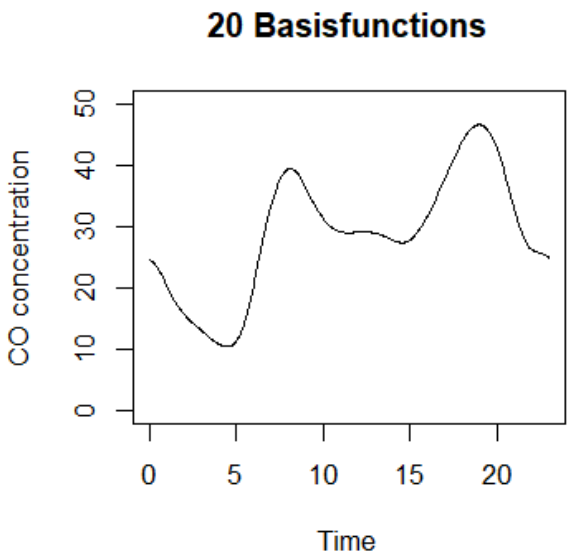

12 Basisfunctions
CO concentration
0
20
40
60
80
100
0
5
10
15
20
Time


20 Basisfunctions
CO concentration
0
20
40
60
80
100
0
5
10
15
20
Time


8 Basisfunctions
CO concentration
0
10
20
30
40
50
0
5
10
15
20
Time


10 Basisfunctions
CO concentration
0
10
20
30
40
50
0
5
10
15
20
Time


12 Basisfunctions
CO concentration
0
10
20
30
40
50
0
5
10
15
20
Time


20 Basisfunctions
CO concentration
0
10
20
30
40
50
0
5
10
15
20
Time

- Definition $\tilde{\phi}$ aus k Werten $\phi(s_j)$

- Dann gilt approximativ:

$$V\phi(s_j) = \int \gamma(s_j, s)\phi(s)ds \approx \frac{T}{n} \sum \gamma(s_j, s_k)\tilde{\phi}_k,$$

- mit Elementen der Kovarianzmatrix $\mathbf{V}$: $\gamma(s_j, s_k)$
- diskrete Form des funktionalen Eigenwertproblems:

$$\frac{T}{k}\mathbf{V}\tilde{\phi} = \lambda\tilde{\phi}$$

- unter $\frac{T}{k}\|\tilde{\phi}\|^2 = 1$
- gilt für die diskrete Approximierung: $\tilde{\phi} = \frac{T}{k}^{-\frac{1}{2}}\boldsymbol{u}$
- Funktion ϕ dann aus $\tilde{\phi}$ durch geeignete Interpolation

- Lösung: Singulärwertzerlegung: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{W}^T$
  - $\mathbf{U}$ ist Nxq und orthogonal: $\mathbf{U}^T\mathbf{U} = \mathbf{I}_q$;
  - $\mathbf{D}$ ist eine qxq Diagonalmatrix mit $diag(D) = d_1 \geq \ldots \geq d_q \geq 0$;
  - $\mathbf{W}$ ist kxq und auch orthogonal: $\mathbf{W}^T\mathbf{W} = \mathbf{I}_q$
- symmetrische Matrix $\mathbf{V} \Rightarrow d_i$ beinhalten alle nichtnegativen Eigenwerte von $\mathbf{X}$
- Auswirkungen auf Kovarianzoperator $\mathbf{V}$:

$$N\boldsymbol{V} = \boldsymbol{X}^T\boldsymbol{X} = (\boldsymbol{W}\boldsymbol{D}^T\boldsymbol{U}^T)(\boldsymbol{U}\boldsymbol{D}\boldsymbol{W}^T) = \boldsymbol{W}\boldsymbol{D}^2\boldsymbol{W}^T$$

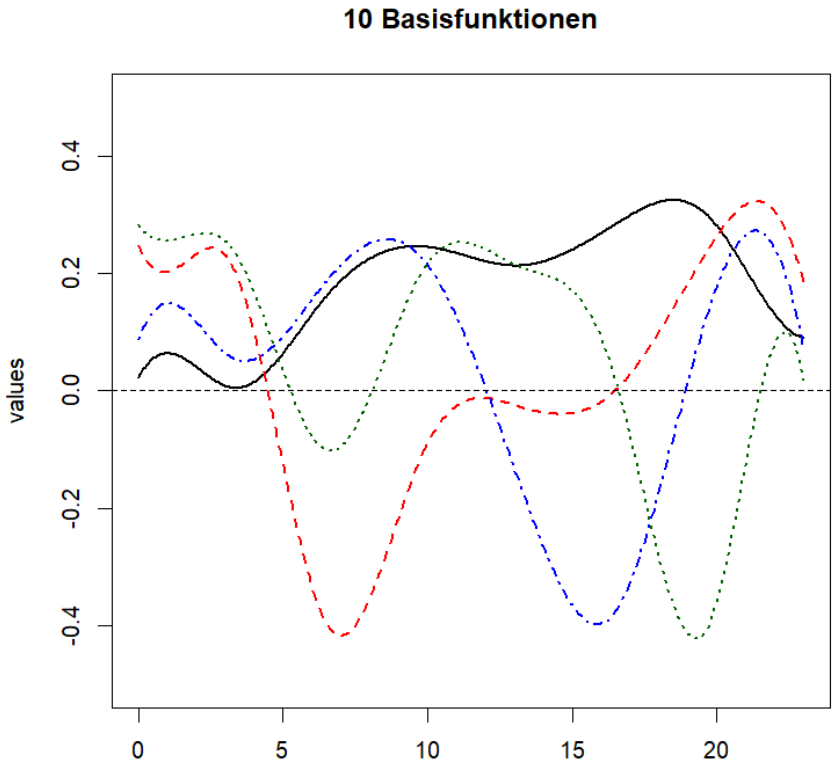
- Eigenwerte von $\mathbf{V}$ dann $diag(\mathbf{D})^2$; Eigenvektoren in Spalten von $\mathbf{W}$

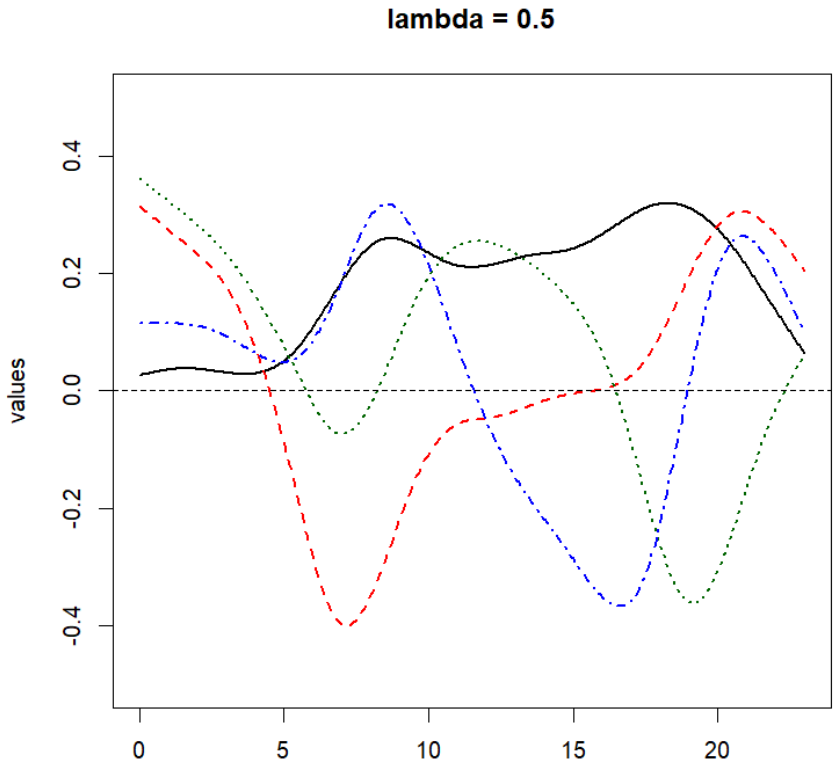
- **maximieren** von **PCAPSV**(ϕ_j) bzgl.
- $\|\phi_j\|^2 = 1$ und einer modifizierten Orthogonalitätsbedingung:

$$\int \phi_j(t)\phi_k(t)dt + \int D^2\phi_j(t)D^2\phi_k(t)dt = 0, \quad k = 1, \ldots, j-1$$

- ergibt j-te Hauptkomponente ϕ_j

- Unterschied zur Glättung vor Hauptkomponentenanalyse





- Anfangsbereich verschieden, danach ähnlich
- Für optimales λ: Leave-one-out Kreuzvalidierung