Bachelorarbeit

Vergleich von Selektionskriterien für die Anzahl funktionaler Hauptkomponenten

Berger Moritz

Betreuung:
Dr. Sonja Greven
Dr. Jan Gertheiss

23.August 2011

Ludwig-Maximilians-Universität München
Institut für Statistik

Zusammenfassung

Die Hauptkomponentenanalyse ist ein wichtiges Instrument der explorativen Datenanalyse. Grundprinzip ist die Reduktion der Daten auf wenige Hauptkomponenten, die einen möglichst großen Anteil der Gesamtvarianz erklären. Eine der entscheidenden Fragen ist die adäquate Anzahl an Hauptkomponenten. In einer Simulationsstudie werden sieben Selektionskriterien zur Wahl der Anzahl an Hauptkomponenten verglichen.

Betrachtet wird speziell der Fall funktionaler Daten. Bei der Berechnung der Hauptkomponentenscores geht man von der Linearkombination, im multivariaten Fall, zum Integral über. Geht man von einem Modell mit zusätzlichen Messfehlern aus, so eignet sich als bessere Alternative das PACE-Verfahren („principal component analysis through conditional expectation“), bei dem die Hauptkomponentenscores durch bedingte Erwartungen berechnet werden. Der Simulationstudie liegt dieses Schätzverfahren zugrunde.

Die Auswertung der Simulationsstudie zeigt, dass drei der vier vorgestellten Selektionskriterien, die klassisch im Fall multivariater Daten Anwendung finden, auch zur Selektion funktionaler Hauptkomponenten geeignet sind. Desweiteren eignet sich ein cAIC-Kriterien, das für die Selektion zufälliger Effekte in gemischten Modellen entwickelt wurde. Diskutiert werden außerdem zwei Kriterien, die speziell für den Fall spärlicher funktionaler Daten entwickelt wurden. Wie sich zeigt, sind diese im betrachteten einfachen funktionalen Fall ungeeignet.

Inhaltsverzeichnis

1 Einleitung **1**

2 Hauptkomponentenanalyse **5**

2.1 Hauptkomponentenanalyse für multivariate Daten 5

2.2 Funktionale Hauptkomponentenanalyse (FPCA) 7

2.2.1 Lösung des Eigenwertproblems 7

2.2.2 Definition einer optimalen empirischen Orthonormalbasis 9

2.2.3 Varimax-Rotation 11

2.2.4 Glättung der Hauptkomponenten 11

2.3 FPCA durch Bedingte Erwartung 15

2.3.1 Modell mit Messfehlern 15

2.3.2 Schätzung aller Modellkomponenten 16

2.3.3 Durchführung der PACE 17

3 Kriterien zur Wahl der Anzahl der Hauptkomponenten **21**

3.1 Erklärung eines Anteils der Gesamtvarianz 22

3.2 Varianz der Hauptkomponenten (Kaiser-Kriterium) 22

3.3 Scree-Test 23

3.4 Test nach Bartlett 24

3.5 Bestimmung von k durch Kreuzvalidierung 25

3.6 Bestimmung von k durch das cAIC 26

3.7 cAIC für gemischte Modelle 27

4 Simulationsstudie **29**

4.1 Simulationsaufbau 29

4.2 Ausgewählte Ergebnisse 32

4.3 Auswertung der Selektionskriterien 37

4.3.1 Erklärung eines Anteils der Gesamtvarianz 37

4.3.2 Varianz der Hauptkomponenten (Kaiser-Kriterium) 39

4.3.3 Scree-Test 40

4.3.4 Test nach Bartlett 42

4.3.5 cAIC nach [Yao et al., 2005] 44

4.3.6 cAIC für gemischte Modelle 45
4.3.7 Bestimmung von k durch Kreuzvalidierung 47
4.3.8 Vergleich und Bewertung der Kriterien 48
4.3.9 Vergleich der verschiedenen Szenarien 51

5 Anwendung **53**

6 Resümee **57**

Literaturverzeichnis **60**

A Verfügbare Dateien **61**

B Auszüge des R-Codes und Outputs **63**

1 Einleitung

Was sind funktionale Daten? Ein typisches Beispiel für die Art von Daten, die in dieser Arbeit behandelt werden, zeigt Abbildung 1.

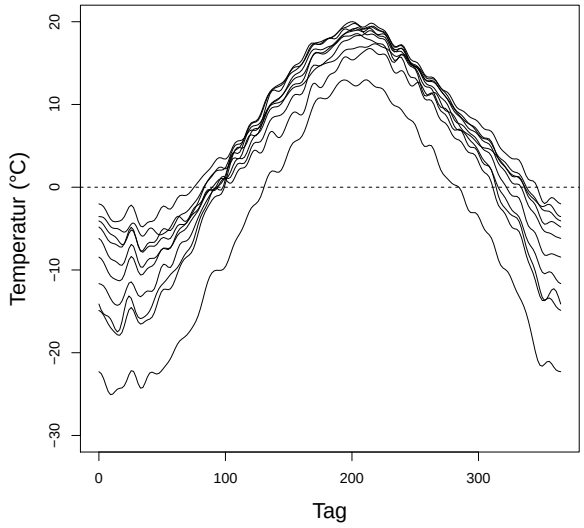


Abbildung 1: Messungen der Durchschnittstemperaturen zehn kanadischer Städte. Registriert wurden alle 365 Tage eines Jahres.

Abgebildet sind die Durchschnittstemperaturen zehn verschiedener Orte in Kanada. Gemessen wurden alle 365 Tage eines Jahres. Jede der Messungen besteht zwar nur aus diskreten Werten, aber über das Jahr hinweg verändert sich die Temperatur nicht sprunghaft und daher weisen die einzelnen Messungen einen glatten Verlauf auf. Aus diesem Grund kann man jede der Messungen als Funktion und nicht als Vektor von Beobachtungen an diskreten Zeitpunkten auffassen [Ramsay and Silverman, 2005]. Somit erhält man als Daten 10 funktionale Beobachtungen $\text{Temperatur}_i(\text{t})$, $i = 1, ..., 10$.
Von Interesse bei der Analyse sind nicht Temperaturen an einzelnen Tagen, sondern beispielsweise die Amplitude und die Steigung der Temperaturen. Man stellt sich die Frage, wie und wie schnell sich die Temperaturen ändern. In Betracht gezogen werden dazu die erste und zweite Ableitung der Temperaturkurven. Auch dies zeigt den Sinn, die Messungen nicht als Vektoren von Beobachtungen aufzufassen, sondern als Funktionen [Ramsay and Silverman, 2005].
Generell muss das betrachtete Interval nicht wie in diesem Fall ein Zeitintervall T sein. Die Punkte, an denen die Beobachtungen gemessen werden, müssen nicht gleich-

abständig sein und können sich auch je nach Messung unterscheiden.

Ein sehr nützliches Instrument zur Analyse funktionaler Daten ist die Hauptkomponentenanalyse. Sie gehört zu den explorativen Methoden der Statistik und ist ursprünglich ein Instrument der multivariaten Datenanalyse. Eingeführt wurde die Hauptkomponentenanalyse zunächst durch Pearson (1901) und weiterentwickelt durch Hotelling (1933) [Jolliffe, 2002].
Die Grundidee der Hauptkomponentenanalyse ist eine Datenreduktion bzw. Dimensionsreduktion. Die Dimension der betrachteten Daten soll auf möglichst wenige Hauptkomponenten reduziert werden, ohne zu viel Information aus den Daten zu verlieren. Die Hauptkomponenten sollen einen möglichst großen Teil der Varianz der Daten erklären. Die bestimmten Hauptkomponenten sind unkorreliert und absteigend geordnet, d.h. die erste Hauptkomponeten hat die größte Varianz, die zweite Hauptkomponente die zweitgrößte Varianz usw. [Jolliffe, 2002].
Eine wichtige Frage, die sich bei der Bestimmung der Hauptkomponenten stellt, ist die Anzahl der relevanten Hauptkomponenten. Die Überlegung, wie stark der Datensatz reduziert werden soll, ohne zu viel Information zu verlieren, ist von entscheidender Bedeutung. Die Frage, welche Anzahl an Hauptkomponenten bei der funktionalen Hauptkomponentenanalyse selektiert werden, ist Gegenstand dieser Arbeit. In einer Simulationsstudie werden verschiedene Kriterien zur Selektion der Anzahl funktionaler Hauptkomponenten verglichen.

Die vorliegende Arbeit ist folgendermaßen aufgebaut:

Kapitel 2 bietet eine Einführung in die Hauptkomponentenanalyse. Betrachtet wird der Fall multivariater Daten und zwei Ansätze der funktionalen Hauptkomponentenanalyse. Theoretische und mathematische Grundlagen sollen vermittelt werden.

Die verschiedenen Kriterien zur Selektion einer Anzahl an Hauptkomponenten beinhaltet Kapitel 3. In Betracht gezogen werden einerseits klassisch für den Fall multivariater Daten verwendete Kriterien, als auch speziell für den Fall funktionaler Daten entwickelte Methoden.

Kapitel 4 beinhaltet die eigentliche Simulationsstudie. Alle in Kapitel 3 vorgestellten Kriterien werden anhand von simulierten Daten ausgewertet und analysiert.

Als letztes enthält Kapitel 5 eine kleine Anwendung. Hier werden die Selektionskriterien auf den vorgestellten Wetterdatensatz angewendet und die empfohlene Anzahl an funktionalen Hauptkomponenten anhand der Simulationsergebnisse diskutiert.

Alle statistischen Analysen, die in der Arbeit vorgestellt werden, wurden mit der Software R durchgeführt [R Development Core Team, 2011]. Der Anhang enthält eine Aufstellung aller erzeugten Source-Dateien („.r“), der darin enthaltenen Funktionen und der Dateien, in denen die Ergebnisse gespeichert sind („.RData“).

In den mathematischen Formeln und Ausdrücken in der Arbeit sind Vektoren klein und fett markiert (z.B. $\boldsymbol{\phi}$) und Matrizen groß und fett markiert (z.B. $\boldsymbol{R}$), um diese von Skalaren und Funktionen zu unterscheiden.

2 Hauptkomponentenanalyse

Dieses Kapitel bietet eine Einführung in die Theorie und die Mathematik der Hauptkomponentenanalyse. Betrachtet wird zunächst die Hauptkomponentenanalyse für den Fall multivariater Daten. Einige der Selektionskriterien für die Anzahl funktionaler Hauptkomponenten, die in Kapitel 3 vorgestellt werden, werden klassisch für den multivariaten Fall verwendet. Abschnitt 2.1 gibt hierzu einen kurzen Einblick.
Anschließend werden zwei Lösungsansätze der funktionalen Hauptkomponentenanalyse vorgestellt. Der klassische Ansatz (Abschnitt 2.2), bei dem die Hauptkomponentenscores durch numerische Integration berechnet werden [Ramsay and Silverman, 2005] und ein weiterer Ansatz (Abschnitt 2.3), bei dem man die Lösung über bedingte Erwartungen erhält [Yao et al., 2005]. Dieser Ansatz lautet auch „principal component analysis through conditional expectation“ (PACE) und wurde speziell für den Fall spärlicher funktionaler Daten entwickelt. Von Interesse ist auch der Vergleich und die Eignung der beiden unterschiedlichen Verfahren.

2.1 Hauptkomponentenanalyse für multivariate Daten

Ausgangspunkt ist eine Datenmatrix $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$,wobei $\boldsymbol{X} \in \mathbb{R}^{nxp}$. n bezeichnet die Anzahl der Beobachtungen, p die Anzahl an betrachteten Variablen.

Grundlage der mathematischen Betrachtung sind **zentrierte** Daten. Es gilt: $\tilde{x}_{ij} = x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. Für Varianz und Kovarianz gilt dann folglich:

$$Var_{\tilde{x}_j} = \frac{1}{n-1}\sum_{i=1}^{n} \tilde{x}_{ij}^2 \qquad \text{und} \qquad Cov_{\tilde{x}_j\tilde{y}_j} = \frac{1}{n-1}\sum_{i=1}^{n}[\tilde{x}_{ij}][\tilde{y}_{ij}] \tag{2.1}$$

Allgemein ist die Hauptkompontenanalyse eine schrittweise Prozedur. Gesucht sind jeweils normierte Hauptkomponenten $\boldsymbol{\phi}$, die jeweils die Varianz von $\boldsymbol{\xi} = \boldsymbol{\phi}^T\boldsymbol{x}$ maximieren [Ramsay and Silverman, 2005, S. 148/149].

1. Gesucht ist der Vektor $\boldsymbol{\phi}_1 = (\phi_{11}, ...\phi_{p1})^T$, für welche die **Linearkombinationen**

$$\xi_{i1} = \sum_j \phi_{j1}x_{ij} = \boldsymbol{\phi}_1^T\boldsymbol{x}_i, \quad i = 1, ..., n,\ j = 1, ..., p \tag{2.2}$$

   die **größte** Varianz annehmen, d.h. $\text{Var}(\boldsymbol{\xi}_1) = \frac{1}{n-1}\sum_i \xi_{i1}^2 \rightarrow \max$

$\frac{1}{n-1}\sum_i \xi_{i1}^2$ entspricht aufgrund der Betrachtung zentrierter Daten genau der Varianz. $\boldsymbol{\phi}_1$ bezeichnet die erste Hauptkomponente. Sie erklärt die stärkste und wichtigste Art der Variabilität.

Die Bedingung für die Lösung von $\boldsymbol{\phi}_1$ lautet: $\sum_{j=1}^{p} \phi_{j1}^2 = \|\boldsymbol{\phi}_1\|^2 = 1$. Dies wird festgelegt, um sicherzustellen, dass das Problem wohldefiniert ist. Ohne die Bedingung könnte die Varianz beliebig groß aufgeblasen werden. Durch die Bedingung der Wohldefiniertheit sind die Hauptkomponenten jedoch nicht einmalig festgelegt. Ändert man beispielsweise alle Vorzeichen des Vektors $\boldsymbol{\phi}$, so bleibt die Varianz trotzdem dieselbe.

2. Man bestimmt die weiteren Hauptkomponenten, bis maximal zur Anzahl p.
   Im m-ten Schritt berechnet man die m-te Hauptkomponente $\boldsymbol{\phi}_m$ mit Komponenten ϕ_{jm} und $\xi_{im} = \boldsymbol{\phi}_m^T \boldsymbol{x}_i$, sodass die Varianz der ξ_{im} maximal wird.

Die Bedingungen für die Lösung der $\boldsymbol{\phi}_m$ lauten: $\|\boldsymbol{\phi}_m\|^2 = 1$ und $\sum_j \phi_{jk}\phi_{jm} = \boldsymbol{\phi}_k^T \boldsymbol{\phi}_m = 0, k < m$. Zweiteres stellt sicher, dass jede neu bestimmte Hauptkomponente auf den vorher bestimmten Hauptkomponenten senkrecht steht. Somit nimmt der Anteil der Varianz, den $\boldsymbol{\phi}_m$ erklärt, in jedem Schritt ab.

Die Lösung des Maximierungsproblems ist äquivalent zur Bestimmung der Eigenwerte und Eigenvektoren der Varianz-Kovarianz $\mathbf{V}$ von $\mathbf{X}$. Das Kriterium zur Maximierung der Varianz kann dargestellt werden durch [Ramsay and Silverman, 2005, S. 152/153]:

$$\max \frac{1}{n}\sum (\boldsymbol{\phi}^T \boldsymbol{x})^2 = \max \frac{1}{n} \boldsymbol{\phi}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\phi} = \max \boldsymbol{\phi}^T \boldsymbol{V} \boldsymbol{\phi} \tag{2.3}$$

Es gilt die bekannte Bedingung $\boldsymbol{\phi}^T \boldsymbol{\phi} = 1$. Betrachtet werden in (2.3) die Matrix $\mathbf{X} \in \mathbb{R}^{nxp}$, welche die x_{ij} beinhaltet, der Vektor $\boldsymbol{\phi} \in \mathbb{R}^{px1}$ und die Matrix $\mathbf{V} \in \mathbb{R}^{pxp}$. Da es für die Berechnung keinen Unterschied macht, wird hier bei der Varianz nur durch n und nicht durch $n-1$ geteilt.

Das Maximierungsproblem wird nun gelöst durch das Eigenwertproblem:

$$\boldsymbol{V}\boldsymbol{\phi} = \lambda \boldsymbol{\phi} \tag{2.4}$$

Man erhält Eigenwert-Eigenvektor Paare $(\lambda_j, \boldsymbol{\phi}_j)$. Alle Eigenvektoren $\boldsymbol{\phi_j}$ stehen aufeinander senkrecht. Da der Mittelwert jeder Spalte von $\mathbf{X}$ von den jeweiligen Beobachtungen abgezogen wird, hat $\mathbf{X}$ maximal Rang $n-1$. Die Matrix $\mathbf{V}$ hat somit maximal

$\min\{p, n-1\}$ Eigenwerte λ_j, die von Null verschieden sind. Für die m-te Hauptkomponente gilt, dass $\text{var}(\boldsymbol{\xi}_m) = \lambda_m$, der m-größte Eigenwert von $\boldsymbol{V}$.

2.2 Funktionale Hauptkomponentenanalyse (FPCA)

Definiert man sich die Hauptkomponentenanalyse analog zum multivariaten Fall (Kapitel 2.1), so betrachtet man nun anstelle von Werten x_{ij} Funktionen $x_i(t)$. Der Index j im multivariaten Kontext wird nun durch den stetigen Index t ersetzt. Bei der Analyse erhält man als Hauptkomponenten Funktionen $\phi_j(t)$ und Hauptkomponentenscores $\xi_i = \int \phi x_i = \int \phi(t)x_i(t)\text{dt}$ [Ramsay and Silverman, 2005, S.149/150]. Das so definierte Skalarprodukt durch das Integral entspricht der Definition der Linearkombination im diskreten Fall.

Man bestimmt analog zum multivariaten Fall $\phi_1(s)$, indem man die Varianz $\frac{1}{n}\sum_i \xi_{i1}^2 = \frac{1}{n}\sum_i (\int \phi_1 x_i)^2$ maximiert. Auch hier geht man von zentrierten Daten aus, d.h. $\tilde{x}_i(t) = x_i(t) - \frac{1}{n}\sum_{i=1}^n x_i(t)$.

Die Bedingung der Wohldefiniertheit der Lösung lautet nun: $\int \phi_1(t)^2\text{dt} = \|\phi_1\|^2 = 1$. Die zusätzliche Nebenbedingung der Orthogonalität zur Bestimmung der weiteren Hauptkomponenten lässt sich schreiben als $\int \phi_k(t)\phi_m(t)dt = 0, k < m$.

2.2.1 Lösung des Eigenwertproblems

Geht man im Fall funktionaler Daten ebenso von zentrierten Daten aus, so lässt sich die Varianz-Kovarianz-Funktion $v(s,t)$ definieren durch [Ramsay and Silverman, 2005, S.153]:

$$v(s,t) = \frac{1}{n}\sum_{i=1}^{n} x_i(s)x_i(t) \tag{2.5}$$

Es ist zu beachten, dass zur Definition von $v(s,t)$ wieder n anstelle von $n-1$ verwendet wurde.

Analog zum multivariaten Fall erhält man ein Eigenwertproblem, das sich folgendermaßen formulieren lässt:

$$\int v(s,t)\phi(t)dt = \lambda\phi(s) \tag{2.6}$$

Die linke Seite in (2.6) ist eine Integraltransformation von ϕ, definiert durch:

$$V\phi = \int v(\cdot, t)\phi(t)dt \tag{2.7}$$

Durch den sogenannten Kovarianz-Operator V kann für das Eigenwertproblem direkt geschrieben werden:

$$V\phi = \lambda\phi \tag{2.8}$$

ϕ ist nun eine Eigenfunktion. Es gilt ebenfalls, dass $\text{var}(\xi_m) = \lambda_m$. Die Darstellung (2.8) ähnelt sehr dem multivariaten Eigenwertproblem (2.4). Der Unterschied liegt jedoch in der Anzahl der möglichen Eigenwert-Eigenfunktions-Paare. Theoretisch ist die maximale Anzahl an Eigenfunktionen die Anzahl der Funktionswerte $x(t)$ und somit unbegrenzt. Definiert man sich jedoch in der Praxis n Funktionen $x(t)$ auf einem Gitter $t = 1, ..., T$, so kann man höchstens $min(n, T)$ Eigenfunktionen bestimmen [Ramsay and Silverman, 2005, S.154].

Um $(\lambda, \phi(t))$ explizit zu berechnen, diskretisiert man die Funktionen $x_i(t)$, die im Interval T liegen, in k Werte s_j mit jeweils gleichem Abstand. Man führt das Problem auf den diskreten Fall zurück, denn man erhält eine Datenmatrix $\boldsymbol{X} \in \mathbb{R}^{nxk}$. Betrachtet man die Singulärwertzerlegung $\mathbf{UDW}^T$ von $\mathbf{X}$ so gilt für die Kovarianzmatrix $\mathbf{V}$ [Ramsay and Silverman, 2005, S.161]:

$$n\boldsymbol{V} = \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{W}\boldsymbol{D}^2\boldsymbol{W}^T \tag{2.9}$$

Aus (2.9) erhält man als Eigenwerte $diag(\boldsymbol{D}^2)$ und als Eigenvektoren die Spalten von $\mathbf{U}$, bezeichnet mit $\mathbf{u}$. Diese bilden ein Orthonormalsystem.
Um die Vektoren wieder zu Funktionen zurück zu transformieren, definiert man sich den Vektor $\tilde{\boldsymbol{\phi}}$, der k Werte $\phi(s_j)$ enthält. Dann gilt approximativ:

$$\int v(s_j, s)\phi(s)\text{ds} \approx \frac{|T|}{k}\sum v(s_j, s_k)\tilde{\boldsymbol{\phi}}_k \tag{2.10}$$

wobei $v(s_j, s_k)$ die Elemente der Kovarianzmatrix $\mathbf{V}$ sind. Somit hat das funktionale Eigenwertproblem folgende diskrete Form:

$$\frac{|T|}{k}\boldsymbol{V}\tilde{\boldsymbol{\phi}} = \lambda\tilde{\boldsymbol{\phi}} \tag{2.11}$$

Unter der Bedingung $\frac{|T|}{k}\|\tilde{\boldsymbol{\phi}}\|^2 = 1$, gilt für die diskrete Approximierung $\tilde{\boldsymbol{\phi}} = \frac{|T|}{k}^{-\frac{1}{2}}\boldsymbol{u}$. Die Funktion ϕ erhält man aus $\tilde{\boldsymbol{\phi}}$ durch geeignete Interpolation.
Die Wahl der Interpolationsmethode hat keinen großen Einfluss auf das Ergebnis, falls man die Werte s_j nur nah genug beieinander wählt [Ramsay and Silverman, 2005, S.161].

2.2.2 Definition einer optimalen empirischen Orthonormalbasis

Alternativ kann man die Hauptkomponentenanalyse auch durch folgendes Problem motivieren [Ramsay and Silverman, 2005, S.151/152]: Man sucht k orthonormale Funktionen ξ_m, sodass die Basisentwicklung jeder Funktion bezüglich dieser Basisfunktionen, die ursprüngliche Funktion möglichst genau approximiert. Die Basisentwicklung ist von der Form

$$\hat{x}_i(t) = \sum_{m=1}^{k} \xi_{im}\phi_m(t) \tag{2.12}$$

mit $\xi_{im} = \int x_i\phi_m$. Man betrachtet das Fehler-Kriterium:

$$\text{PCASSE} = \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 = \int [x(s) - \hat{x}(s)]^2 ds \tag{2.13}$$

Dies wird auch als **globaler Fehler** bezeichnet. Die Minimierung des globalen Fehlers entspricht der Maximierung des Varianzkriteriums. Die Basisfunktionen ϕ_m werden in diesem Kontext auch als empirische Orthonormalfunktionen bezeichnet.

Mathematische Lösung des Eigenwertproblems

Betrachtet man wieder das Problem der Lösung des funktionalen Eigenwertproblems (2.6), so kann man sich auch die Idee der Basisentwicklung zu nutze machen. Um das Eigenwertproblem zu diskretisieren bzw. in Matrix-Form auszudrücken, nimmt man an, dass jede Funktion x_i als Linearkombination bekannter Basisfunktionen dargestellt werden kann [Ramsay and Silverman, 2005, S.162], d.h.

$$x_i(t) = \sum_{m=1}^{k} c_{im}\gamma_m(t) \quad \text{bzw.} \quad \boldsymbol{x} = \boldsymbol{C\gamma} \tag{2.14}$$

Betrachtet werden in (2.14) der Vektor $\boldsymbol{x} \in \mathbb{R}^{n\text{x}1}$, der Vektor $\boldsymbol{\gamma} \in \mathbb{R}^{k\text{x}1}$, ein Vektor aus Basisfunktionen, und die Matrix $\boldsymbol{C} \in \mathbb{R}^{n\text{x}k}$.

Ein Beispiel ist das Fourier-Basis-System, bei dem gilt [Forster, 2008]:

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{k}[a_m cos(m\omega x) + b_m sin(m\omega x)] \tag{2.15}$$

mit $a_m = \frac{1}{\pi}\int_0^{2\pi} f(x)cos(m\omega x)$ und $b_m = \frac{1}{\pi}\int_0^{2\pi} f(x)sin(m\omega x)$.
Geht man von der Basisentwicklung (2.14) aus, so kann man die Varianz-Kovarianz Funktion schreiben als:

$$v(s,t) = \frac{1}{n}\boldsymbol{\gamma}(s)^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\gamma}(t) \tag{2.16}$$

Definiert man die Hilfsmatrix $\boldsymbol{W} = \int \boldsymbol{\gamma}\boldsymbol{\gamma}^T$ und nimmt an, dass $\phi(t) = \boldsymbol{\gamma}(t)^T\boldsymbol{b}$, d.h. die $\phi(t)$ ebenfalls durch die bekannten Basisfunktionen $\gamma(t)$ entwickelt werden, führt das zu [Ramsay and Silverman, 2005, S.162]:

$$\begin{aligned}\int v(s,t)\phi(t)\mathrm{dt} &= \int \frac{1}{n}\boldsymbol{\gamma}(s)^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\gamma}(t)\boldsymbol{\gamma}(t)^T\boldsymbol{b}\ \mathrm{dt} \\ &= \boldsymbol{\gamma}(s)^T\frac{1}{n}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{W}\boldsymbol{b}\end{aligned} \tag{2.17}$$

Das Eigenwertproblem (2.6) sieht schließlich wie folgt aus:

$$\boldsymbol{\gamma}(s)^T\frac{1}{n}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{W}\boldsymbol{b} = \lambda\boldsymbol{\gamma}(s)^T\boldsymbol{b} \tag{2.18}$$

Da (2.18) für alle s gelten muss, erhält man die Matrix-Gleichung:

$$\frac{1}{n}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{W}\boldsymbol{b} = \lambda\boldsymbol{b} \tag{2.19}$$

Durch Umformungen erhält man für die bekannte Bedingung der Wohldefiniertheit der Eigenfunktionen $\|\phi\| = 1$: $\boldsymbol{b}^T\boldsymbol{W}\boldsymbol{b} = 1$. Ähnlich gilt für die Orthogonalität zweier Eigenfunktionen ϕ_m und ϕ_k: $\int \phi_k\phi_m = \boldsymbol{b}_k^T\boldsymbol{W}\boldsymbol{b}_m = 0$.
Hat man $\mathbf{b}$ explizit berechnet, so erhält man daraus die Hauptkomponenten $\phi(t)$. Im Fall einer Orthonormalbasis, z.B. der Fourier-Basis, gilt $\boldsymbol{W} = \boldsymbol{I}$ und man erhält ein vereinfachtes Lösungsproblem. Nimmt man an, dass die beobachtete Funktion x_i ihre Basisentwicklung selbst ist, so gibt es folglich n Basisfunktionen. Das impliziert klarerweise, dass $\boldsymbol{C} = \boldsymbol{I}$ und man löst das Eigenwertproblem der Matrix $\frac{1}{n}\boldsymbol{W}$ [Ramsay and Silverman, 2005, S.163].

2.2.3 Varimax-Rotation

Im Abschnitt 2.2.2 wurde die Hauptkomponentenanalyse als Suche nach k orthonormalen Basisfunktionen ϕ_m motiviert, die die Funktionen x_i so aufspannen, dass der globale Fehler (2.13) minimiert wird. Allgemein bedeutet dies nicht, dass es nur eine eindeutige orthonormale Menge an Funktionen ϕ_m gibt, die dies leistet. Eine Methode, die man auf die berechneten Eigenfunktionen anwenden kann, ist die sogenannte Varimax-Rotation [Ramsay and Silverman, 2005, S. 156-158]. Gesucht ist eine andere Orthonormalbasis $\boldsymbol{\psi} = (\psi_1, ..., \psi_k)^T$, die die originalen Kurven mindestens genauso gut approximiert wie $\boldsymbol{\phi}$. Die Orthonormalbasis ist definiert durch:

$$\boldsymbol{\psi} = \boldsymbol{T}\boldsymbol{\phi} \tag{2.20}$$

wobei $\mathbf{T}$ eine orthonormale Matrix ist und gilt $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{I}$.
Sei $\boldsymbol{B} \in \mathbb{R}^{k\text{x}n}$, die Matrix der ersten k Hauptkomponenten $\xi_1, ..., \xi_k$. Die korrespondierende Matrix $\boldsymbol{A}$ der rotierten Basisfunktionen ist dann gegeben durch:

$$\boldsymbol{A} = \boldsymbol{T}\boldsymbol{B} \tag{2.21}$$

Die Strategie, die orthonormale Rotationsmatrix $\boldsymbol{T}$ zu bestimmen, ist die Varianz der Werte a_{mj}^2 zu maximieren. Für die Quadratsumme gilt:

$$\sum_m \sum_j a_{mj}^2 = \text{spur}(\boldsymbol{A}^T\boldsymbol{A}) = \text{spur}(\boldsymbol{B}^T\boldsymbol{T}^T\boldsymbol{T}\boldsymbol{B}) = \text{spur}(\boldsymbol{B}^T\boldsymbol{B}) \tag{2.22}$$

Die Quadratsumme bleibt also dieselbe, unabhängig von der Wahl der Rotation.
Geometrisch ist $\boldsymbol{\psi}$ eine Rotation von $\boldsymbol{\phi}$.
Sinnvoll ist die Anwendung der Varimax-Rotation bei der Frage, ob es alternative Basisfunktionen zu ϕ_m gibt, die einfacher zu interpretieren sind. Es kann dann jedoch nicht mehr sichergestellt werden, dass ψ_1 die größte Varianzkomponente beschreibt.

2.2.4 Glättung der Hauptkomponenten

In vielen Bereichen der Statistik, wie z.B. der Regressionsanalyse, spielt die Glattheit der geschätzten Funktionen eine wichtige Rolle. Speziell im Fall funktionaler Daten wird an die Beobachtungen eine gewisse Glattheitsanforderung gestellt. Smoothing-Splines und lokale polynomiale Glättung sind zwei Ansätze, wie man Glattheit von

Funktionen erhalten kann. Diese spielen bei der PACE (Abschnitt 2.3) eine wichtige Rolle, wenn es darum geht, die Glattheit der geschätzten Eigenfunktionen zu erzeugen. Im gerade betrachteten Fall der funktionalen Hauptkomponentenanalyse wird das Konzept der Glättung durch einen Penalisierungsterm umgesetzt. Die Idee dahinter ist, dass man die Glättung direkt in die Hauptkomponentenanalyse mit aufnimmt. Ein alternativer Ansatz wäre die Daten zunächst zu glätten, z.B. durch oben genannte Methoden, und dann eine Hauptkomponentenanalyse durchzuführen.

Als Maß für die Rauheit verwendet man [Ramsay and Silverman, 2005, S.177]:

$$PEN_2(\phi) = \|D^2\phi\|^2 = \int \phi''(t)^2 dt \tag{2.23}$$

Bisher wurden die Hauptkomponenten ermittelt, indem man die Varianz $\frac{1}{n}\sum_i \xi_{i1}^2 = \frac{1}{n}\sum_i (\int \phi_1 x_i)^2$ maximiert. Dies führte zum Eigenwertproblem (2.6). Das Ziel der regularisierten Hauptkomponentenanalyse ist, zugleich die Rauheit der Hauptkomponenten $PEN_2(\phi)$ zu kontrollieren. Als neues Kriterium betrachtet man die „penalisierte“ Varianz [Ramsay and Silverman, 2005, S.177]:

$$\text{PCAPSV}(\phi) = \frac{var \int \phi x_i}{\|\phi\|^2 + \lambda \times \text{PEN}_2(\phi)} \tag{2.24}$$

$\lambda \geq 0$ bezeichnet den Glättungsparameter.

Falls $\lambda \to 0$, gilt $\text{PCAPSV}(\phi) \to \text{var} \int \phi x_i$. Die zu bestimmende Hauptkomponente bleibt also wie im unpenalisierten Fall.

Falls $\lambda \to \infty$, gilt $\phi = a$ für den periodischen Fall und $\phi = a + bt$ für den nichtperiodischen Fall, mit $a, b =$ konstant. Die zu bestimmende Hauptkomponente wird also zu einer Konstanten bzw. einer Geraden mit Steigung. Im Allgemeinen heißt eine Funktion periodisch, falls gilt: $f(t + p) = f(t)$, mit Periode p.

Die Hauptkomponenten werden bestimmt durch die Maximierung von PCAPSV (ϕ_j). Die bekannten Bedingungen für die Lösung lauten nun $\|\phi_j\|^2 = 1$ für die Wohldefiniertheit, und $\int \phi_j(t)\phi_k(t)dt + \int D^2\phi_j(t)D^2\phi_k(t)dt = 0$,für $k = 1, ..., j-1$ für die Orthogonalität. In letzterer fließt zusätzlich die zweite Ableitung von $\phi_j(t)$ mit ein [Ramsay and Silverman, 2005, S.177/178].

Bestimmung des Glättungsparameters

Die Wahl des Glättungsparameters λ hat einen entscheidenden Einfluss auf das Ergebnis der Glättung. Zur Bestimmung von λ wird hier das Verfahren der Kreuzvalidierung angewendet [Ramsay and Silverman, 2005, S.178/179]. Um ein Kreuzvalidierungsmaß zu berechnen, betrachtet man x_i und die zugehörige Basisentwicklung $\hat{x}_i(t) = \sum_{m=1}^{k} \xi_{im}\phi_m(t)$. Geometrisch interpretiert, ist $\hat{x}_i$ die Projektion von x_i in den Unterraum, der von $\phi_1, ..., \phi_k$ aufgespannt wird. Man definiert die Residualkomponente $\zeta_k = x_i - \hat{x}_i$, also die Komponente von x_i orthogonal zu $\phi_1, ..., \phi_k$ und als Maß für die Wirksamkeit von $\phi_1, ..., \phi_k$ wählt man

$$E\|\zeta_k\|^2 \tag{2.25}$$

Das allgemeine Vorgehen zur Bestimmung des Glättungsparameters ist Folgendes:

1. Lege ein Gitter von Werten fest, aus dem λ gewählt wird.
2. Bestimme $\phi_j^{[i]}(\lambda)$, als Schätzung für ϕ_j aus allen Beobachtungen **außer** x_i (leaving-one-out Kreuzvalidierung).
3. Definiere $\zeta_k^{[i]}(\lambda)$ als die Komponente von x_i, die orthogonal zum, von $\{\phi_j^{[i]}(\lambda) : j = 1, ..., k\}$, aufgespannten Unterraum ist.
4. Berechne $\mathrm{CV}_k(\lambda) = \sum_{i=1}^{n} \|\zeta_k^{[i]}(\lambda)\|^2$ und daraus $\mathrm{CV}(\lambda) = \sum_{k=1}^{\infty} \mathrm{CV}_k(\lambda)$. Man beschränkt sich also nicht auf eine Anzahl k Hauptkomponenten.

Man wählt dasjenige λ, für das $\mathrm{CV}(\lambda)$ minimal wird. In der Praxis mit n Datenkurven, die auf einem Gitter $t = 1, ..., T$ definiert sind, kann man höchstens $min(n, T)$ Hauptkomponenten bestimmen und die Summe wird bei $k = min(n, T)$ abgebrochen.

Praktische Bestimmung der regularisierten Hauptkomponenten

Bei der expliziten Berechnung der regularisierten Hauptkomponenten geht man ebenfalls von der Idee der Basisentwicklung, wie in Abschnitt 2.2.2, aus. Betrachtet wird zunächst der Fall, dass $x_i(t)$ **periodisch** ist [Ramsay and Silverman, 2005, S.179/180]:

$x_i(t) = \boldsymbol{c}_i^T \boldsymbol{\gamma}(t)$ ist durch eine Fourier-Basis entwickelt, wobei $\boldsymbol{\gamma}(t)$ den Vektor aus Basisfunktionen darstellt, und $\mathbf{V}$ ist die Varianz-Kovarianz-Matrix der Vektoren $\boldsymbol{c}_i$. Außer-

dem soll gelten, dass $\phi_m(t) = \boldsymbol{y}_m^T \boldsymbol{\gamma}(t)$, d.h. eine mögliche Hauptkomponente lässt sich durch dieselben Basisfunktionen $\boldsymbol{\gamma}(t)$ wie $x_i(t)$ darstellen.
Eine spezielle Eigenschaft der Fourier-Basis-Funktionen ist, dass $D^2\gamma_m = -\omega_m^2\gamma_m$ und somit $D^2\phi(s) = -\sum_m \omega_m^2 y_m \gamma_m(s)$. Da die γ_m orthonormal sind, gilt:

$$\|D^2\phi\|^2 = \sum_m \omega_m^4 y_m^2 \tag{2.26}$$

Sei $\mathbf{S}$ die Diagonalmatrix mit Einträgen $S_{mm} = (1 + \lambda\omega_m^4)^{-\frac{1}{2}}$, dann entspricht $\mathbf{S}$ einem Glätter und man erhält:

$$\text{PCAPSV}(\xi) = \frac{\boldsymbol{y}^T \boldsymbol{V} \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{S}^{-2} \boldsymbol{y}} \tag{2.27}$$

Die „penalisierte“ Varianz (2.27) entpricht der allgemeinen Darstellung (2.24). Die Bedingung der Orthogonalität lässt sich ausdrücken durch $\boldsymbol{y}_j^T \boldsymbol{S}^{-2} \boldsymbol{y}_k = 0$,für $k = 1, ..., j-1$.
Es lässt sich ableiten, dass das Eigenwertproblem $\boldsymbol{V}\boldsymbol{y} = \rho \boldsymbol{S}^{-2}\boldsymbol{y}$ zu lösen ist, welches sich folgendermaßen darstellen lässt:

$$(\boldsymbol{S}\boldsymbol{V}\boldsymbol{S})(\boldsymbol{S}^{-1}\boldsymbol{y}) = \rho(\boldsymbol{S}^{-1}\boldsymbol{y}) \tag{2.28}$$

$\boldsymbol{S}\boldsymbol{V}\boldsymbol{S}$ ist die Varianz-Kovarianz-Matrix der Vektoren $\boldsymbol{S}\boldsymbol{c}_i$, d.h. der Fourier-Koeffizienten der Originaldaten, geglättet durch den Glätter $\mathbf{S}$.
Die Eigenvektoren und Eigenwerte von $\boldsymbol{S}\boldsymbol{V}\boldsymbol{S}$ zu bestimmen entspricht genau der ungeglätteten Hauptkomponentenanalyse der geglätteten Daten $\boldsymbol{S}\boldsymbol{c}_i$.
Sei nun $\mathbf{u}$ ein Eigenvektor von $\mathbf{SVS}$, so ist $\boldsymbol{S}\boldsymbol{u}$ eine Lösung von (2.28). Da die Bedingung $\|\boldsymbol{y}\|^2 = 1$ erfüllt sein soll, setzt man $\boldsymbol{y} = \frac{\boldsymbol{S}\boldsymbol{u}}{\|\boldsymbol{S}\boldsymbol{u}\|}$.
Die Hauptkomponente ϕ lässt sich schließlich aus $\boldsymbol{y}$ einfach berechnen.

Nun betrachte man den Fall, dass die aufzuspannenden Funktionen $x_i(t)$ **nicht periodisch** sind [Ramsay and Silverman, 2005, S.181]:
Als Basisfunktionen $\boldsymbol{\gamma}(t)$ wählt man in diesem Fall z.B. B-Splines oder orthogonale Polynome bestimmten Grades.
Wie im periodischen Fall nimmt man an, dass die $x_i(t) = \boldsymbol{c}_i^T \boldsymbol{\gamma}(t)$ durch passende Basisfunktionen entwickelt sind. $\mathbf{V}$ ist die Varianz-Kovarianz-Matrix der Vektoren $\boldsymbol{c}_i$.

Definiert man $\boldsymbol{J}$ als Matrix mit Elementen $\int \gamma_j \gamma_k$ und $\boldsymbol{K}$ als Matrix mit Elementen $\int D^2\gamma_j D^2\gamma_k$, so erhält man für die penalisierte Varianz (2.24):

$$\text{PCAPSV}(\xi) = \frac{\boldsymbol{y}^T \boldsymbol{J}\boldsymbol{V}\boldsymbol{J}\boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{J}\boldsymbol{y} + \lambda \boldsymbol{y}^T \boldsymbol{K}\boldsymbol{y}} \tag{2.29}$$

2.3 FPCA durch Bedingte Erwartung

In diesem Kapitel wird ein weiterer Ansatz der Hauptkomponentenanalyse vorgestellt. Dieser nennt sich "principal component analysis through conditional expectation"(PACE) [Yao et al., 2005].

2.3.1 Modell mit Messfehlern

In Abschnitt 2.2.2 wurde die funktionale Hauptkomponentenanalyse motiviert, als die Suche nach k orthonormalen Funktionen ϕ_m, sodass die Basisentwicklung jeder Funktion bezüglich dieser Basisfunktionen, die ursprüngliche Funktion möglichst genau approximiert. Die Basisentwicklung war von der Form:

$$x_i(t) = \mu(t) + \sum_{m=1}^{k} \xi_{im}\phi_m(t) \tag{2.30}$$

mit $\xi_{im} = \int x_i\phi_m$, wobei bisher davon ausgegangen wurde, dass $\mu(t) = 0$. Nun gelte allgemein: $\mathbb{E}X(t) = \mu(t)$.
Zusätzlich gelte, dass eine Folge (ϕ_m) von stetigen Eigenfunktionen und eine Folge (λ_m) nicht steigender Eigenwerte existiert, so dass sich die Varianz-Kovarianz-Funktion v(s,t) durch die folgende unendliche Summe ausdrücken lässt:

$$v(s,t) = \sum_{m=1}^{\infty} \lambda_m \phi_m(s)\phi_m(t) \tag{2.31}$$

Es wird nun angenommen, dass zusätzliche Messfehler existieren und das Modell (2.30) wird um unkorrelierte Messfehler erweitert.
Die Erweiterung führt zu folgendem Gesamtmodell [Yao et al., 2005]:

$$\begin{aligned} y_i(t) &= x_i(t) + \varepsilon_i(t) \\ &= \mu(t) + \sum_{m=1}^{\infty} \xi_{im}\phi_m(t) + \varepsilon_i(t), \quad t = 1, ..., T \end{aligned} \tag{2.32}$$

Die ξ_{im} sind unkorrelierte Zufallsvariablen mit $\mathbb{E}(\xi_{im}) = 0$ und $\text{var}(\xi_{im}) = \mathbb{E}((\xi_{im})^2) = \lambda_m$ und für den Messfehler ε gilt, dass $\mathbb{E}(\varepsilon_i) = 0$ und $\text{var}(\varepsilon_i) = \sigma^2$. Die ξ_{im} sind unabhängig von den $\varepsilon_i(t)$.

2.3.2 Schätzung aller Modellkomponenten

Als erster Schritt, werden alle Komponenten des Modells (2.32) geschätzt. Man nimmt an, dass Mittelwert-, Varianz-Kovarianz- und Eigenfunktionen glatte Funktionen sind [Yao et al., 2005]. Man verwendet lokale lineare Glättung zur Funktions- und Oberflächenschätzung. Eine gute Einführung zum Prinzip der Lokalisierung bei der nonparametrischen Regression findet man in [Fahrmeir et al., 2007].

Das Prinzip der **lokalen polynomialen Glättung** entspricht der lokalen Berechnung eines linearen Modells mithilfe der KQ-Methode. Für die Funktion f(x) legt man ein polynomiales Modell vom Grad m zugrunde, d.h. $\mathbb{E}(f(x_i)) = \beta_0 + \sum_{s=1}^{m} \beta_s (x_i - x)^s$.
Für festes x und festgelegtes m erhält man folgendes Minimierungsproblem:

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{s=1}^{m} \beta_s (x_i - x)^s))^2 \omega_\lambda(x, x_i) \underset{\beta}{\longrightarrow} \min \tag{2.33}$$

Es gilt: $\hat{f}(x) = \hat{\beta}_0 = (1, 0, 0, ...0)\hat{\boldsymbol{\beta}}$.
Klassiker für die Gewichtsfunktion ω_λ sind Kernfunktionen, d.h. stetige, um 0 symmetrische Funktionen, für die gilt: $\int K(u)du = 1$.

$$\omega_\lambda(x, x_i) = K\left(\frac{x - x_i}{\lambda}\right) \tag{2.34}$$

Als Lösung für das Minimierungsproblem (2.33) erhält man den KQ-Schätzer

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}_x^T \boldsymbol{W}_x \boldsymbol{Z}_x)^{-1} \boldsymbol{Z}_x^T \boldsymbol{W}_x \boldsymbol{y} \qquad \text{, mit} \tag{2.35}$$

$$\boldsymbol{Z}_x = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^m \\ 1 & x_2 - x & \cdots & (x_2 - x)^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^n \end{pmatrix} \qquad \boldsymbol{W}_x = \begin{pmatrix} K\left(\frac{x_1 - x}{\lambda}\right) & & 0 \\ & \ddots & \\ 0 & & K\left(\frac{x_n - x}{\lambda}\right) \end{pmatrix}$$

Hat man $\hat{f}(x)$ für alle x auf einem sinnvoll gewählten Gitter berechnet, so erhält man eine glatte Funktion. Die Kernfunktion steuert, wie viele benachbarte x_i jeweils in die Schätzung mit aufgenommen werden.

Zur Schätzung von Mittelwert-, Varianz-Kovarianz- und Eigenfunktionen wird hier lokal linear geglättet, d.h. es wird ein Polynom 1. Grades verwendet. Im ersten Schritt wird die Mittelwertfunktion $\hat{\mu}(t)$ aus allen Beobachtungen geschätzt. Anschließend wird die Varianz-Kovarianzfunktion $\hat{v}(s,t) = (y(t) - \hat{\mu}(t))(y(s) - \hat{\mu}(s))$ ebenfalls aus allen Daten geschätzt, wobei $\hat{\mu}(t)$ der geschätzte Mittelwert des vorherigen Schrittes ist.
Den zu schätzenden Messfehler $\hat{\sigma}^2$ stellt man sich als Zusatzterm auf der Diagonalen von $v(s,t)$ vor. Der Messfehler wirkt sich hier also nur auf die Varianzen und nicht auf die Kovarianzen aus, d.h.

$$cov(y(s), y(t)|s,t) = cov(x(t), x(s)) + \sigma^2 \mathbb{1} \tag{2.36}$$

$v(s,t)$ wird nochmal glatt geschätzt und zwar ohne die Werte auf der Diagonalen. Nur $v(s,t), s \neq t$, wird in die Glättung mit aufgenommen. [Yao et al., 2005] verwenden ebenfalls leaving-one-out Kreuzvalidierung zur Wahl des Glättungsparameters für diesen Glättungsschritt. $\hat{\sigma}^2$ ergibt sich letztendlich als Differenz dieser beiden Schätzungen. Wie bisher erhält man die Eigenfunktionen $\hat{\phi}_m$ und korrespondierende Eigenwerte $\hat{\lambda}_m$ über die Lösung des Eigenwertproblems:

$$\int \hat{v}(s,t)\hat{\phi}_m(s)ds = \hat{\lambda}_m \hat{\phi}_m(t) \tag{2.37}$$

Wie auch schon im bisherigen Ansatz der Hauptkomponentenanalyse (Abschnitt 2.2), gelten für die Lösung die Bedingungen, dass $\int \hat{\phi}(t)^2 dt = 1$ und $\int \hat{\phi_k}(t)\hat{\phi_m}(t)dt = 0$ für $k < m$. Die Eigenfunktionen ϕ_m sind also wohldefiniert und stehen aufeinander senkrecht [Yao et al., 2005].

2.3.3 Durchführung der PACE

Hat man die Eigenfunktionen $\phi_m(t)$ bestimmt, so werden die Gewichte ξ_{im} des Modells (2.32), welche bisher als Hauptkomponentenscores definiert waren, durch bedingte Erwartungen berechnet. Diese Alternative wird kurz **motiviert**:

Bisher wurde definiert: $\xi_{im} = \int (x_i(t) - \mu(t))\phi_m(t)dt$. ξ_m so zu berechnen, funktioniert

gut, falls die Dichte des Gitters der Beobachtungen, hier t=1,...,T, ausreichung groß ist.

Beispielsweise für den Fall spärlicher Daten ist dies nicht so. Speziell für diesen Fall wurde die Methode über den bedingten Erwartungswert entwickelt. Hier werden die $y_i(t)$ nur an wenigen diskreten Zeitpunkten beobachtet. Diese können auch zwischen den Subjekten variieren. Man beobachtet dann y_{ij}, als j-te Beobachtung zum Zeitpunkt t_{ij}, $j = 1, ..., n_i$. Das Integral in der Definition der ξ_m würde dann übergehen zur Summe. Die Lösung für ξ_m würde dann lauten: $\hat{\xi}_{im}^S = \sum_{j=1}^{n_i} = y_{ij} - \hat{\mu}(t_{ij}))\hat{\phi}_m(t_{ij})(t_{ij} - t_{i,j-1})$, mit $t_{i0} = 0$ [Yao et al., 2005].

Falls nur sehr wenige Beobachtungen pro Subjekt vorliegen, liefert $\hat{\xi}_{im}^S$ keine gute Näherung für ξ_{im}. Wenn zudem Messfehler vorliegen, wie im Modell (2.32) angenommen, ergeben sich verzerrte Scores. Berücksichtigt man dies, so motiviert das die alternative PACE-Methode.

Aus dem zugrundeliegenden Modell (2.32) nimmt man an, dass die ξ_{im} und $\epsilon_i(t)$ multivariat normalverteilt sind. Sei $\tilde{\boldsymbol{y}}_i = (y_{i1}, ..., y_{iT})^T$ der Vektor der Beobachtungen y_i, $\boldsymbol{\mu} = (\mu_1, ..., \mu_T)^T$ der Vektor der Mittelwerte und $\boldsymbol{\phi}_m = (\phi_{m1}, ..., \phi_{mT})^T$, dann ist die beste Prädiktion für die ξ_{im}, unter Normalverteilung, die bedingte Erwartung [Yao et al., 2005]:

$$\hat{\xi}_{im} = \hat{E}[\xi_{im}|\tilde{\boldsymbol{y}}_i] = \hat{\lambda}_m \hat{\boldsymbol{\phi}}_m^T \hat{\boldsymbol{\Sigma}}_{y_i}^{-1}(\tilde{\boldsymbol{y}}_i - \hat{\boldsymbol{\mu}}) \tag{2.38}$$

Für das (s,t)-te Element von $\hat{\boldsymbol{\Sigma}}_{y_i}^{-1}$ gilt, dass $(\hat{\boldsymbol{\Sigma}}_{y_i}^{-1})_{s,t} = \hat{v}(s,t) + \hat{\sigma}^2 \mathbb{1}_{s,t}$.

Verwendet man in der Praxis zur Approximation der $x_i(t)$ k Eigenfunktionen, so lautet die Schätzung für die i-te Beobachtung:

$$\hat{x}_i^k(t) = \hat{\mu}(t) + \sum_{m=1}^{k} \hat{\xi}_{im}\hat{\phi}_m(t) \tag{2.39}$$

Im Allgemeinen kann man festhalten, dass diese bedingte Methode einfach ist, die beste Prädiktion unter Normalverteilung liefert und sowohl im Fall von Messfehlern als auch spärlichen Beobachtungen funktioniert [Yao et al., 2005].

Sowohl die klassische Methode, bei der die Hauptkomponentenscores durch Integration berechnet werden, die in Abschnitt 2.2 vorgestellt wurde, als auch die PACE führen zum gleichen Ergebnis. Ein entscheidender Unterschied ist vor allem der Ansatz der Glättung. Der klassische Ansatz nimmt einen Penalisierungsterm λ mit ins Modell auf. Dieser steuert die zu minimierende Varianz (Abschnitt 2.2.4). Bei der PACE hingegen werden die Modellkomponenten von vornherein, vor der Schätzung der Eigenfunktionen, durch Lokalisierung glatt schätzt.
Die in der Simulation (Kapitel 4) verwendeten Daten sind nach dem Modell (2.32) gebildet. Alle für die Auswertungen der Selektionskriterien (Kapitel 3) benötigten Modellkomponenten werden, wie in Abschnitt 2.3.2 beschrieben, geschätzt.

3 Kriterien zur Wahl der Anzahl der Hauptkomponenten

Der Hauptgegenstand dieser Arbeit ist die Wahl der Anzahl der Hauptkomponenten. Wie bereits dargestellt, sollen die Daten bei der Hauptkomponentenanalyse auf möglichst wenige Hauptkomponenten reduziert werden. Zwei Überlegungen sind dazu anzustellen:

1. Wünschenswert ist, dass die Anzahl k der relevanten Hauptkomponenten sehr viel kleiner ist, als die ursprüngliche Anzahl an betrachteten Variablen. Beobachtet man im Fall funktionaler Daten n Funktionen an T Zeitpunkten, so ist dies $min(n, T)$.

2. Die Hauptkomponenten sollen einen möglichst hohen Informationsgehalt mit sich liefern, also einen möglichst großen Anteil der Gesamtvarianz erhalten. Wählt man zu wenige Hauptkomponenten, so geht womöglich sehr viel Information über die Daten verloren.

Der Konflikt zwischen Einfachheit des Modells und Informationsverlust spielt auch in anderen Bereichen der Statistik, wie zum Beispiel der Regressionsanalyse, eine wichtige Rolle. Die kritische Auswahl der Anzahl an Hauptkomponenten ist von entscheidender Bedeutung.
Im Folgenden werden einige Auswahlkriterien vorgestellt, mit Hilfe derer man die Anzahl der Hauptkomponenten festlegen kann.
Die Kriterien, die in Abschnitt 3.1 bis 3.4 beschrieben sind, werden klassisch für den Fall multivariater Daten angewendet. Von Interesse ist in dieser Arbeit zu überprüfen, ob und gegebenenfalls wie gut sich diese Kriterien verwenden lassen, wenn man eine Hauptkomponentenanalyse für funktionale Daten durchführt.
Die Kriterien in Abschnitt 3.2 bis 3.4 basieren auf der standardisierten Datenmatrix $\boldsymbol{Z}$. Die diskrete, auf dem Gitter T definierte, Datenmatrix $\boldsymbol{Y} \in \mathbb{R}^{nxT}$ mit n Beobachtungen y_i, $i = 1, ..., n$ an den Gitterpunkten $t = 1, ..., T$ wird zur standardisierten Datenmatrix $\boldsymbol{Z} = (z_i) \in \mathbb{R}^{nxT}$, durch [Fahrmeir et al., 1996, S. 662]:

$$z_i = \frac{y_i - \bar{y_i}}{\sqrt{n-1}s_i} \qquad \text{mit} \qquad \bar{y_i} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad , \quad s_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y_i})^2 \tag{3.1}$$

Die empirische Korrelationsmatrix lässt sich aus $\boldsymbol{Z}$ dann einfach berechnen als:

$$\boldsymbol{R} = \boldsymbol{Z}^\top \boldsymbol{Z} \tag{3.2}$$

Es ist zu bemerken, dass die Korrelationsmatrix der Varianz-Kovarianz-Matrix standardisierter Daten entspricht. Man geht generell bei der Hauptkomponentenanalyse zu standardisierten Daten über, da die Ergebnisse skalenabhängig sind und daher sehr unterschiedlich ausfallen können. Bisher wurde bei der Vorstellung der Hauptkomponentenanalyse (Kapitel 2) nur von der Varianz-Kovarianz-Matrix gesprochen.

3.1 Erklärung eines Anteils der Gesamtvarianz

Eine erste Möglichkeit die Anzahl der Hauptkomponenten festzulegen, ist eine intuitiv plausible Faustregel [Jolliffe, 2002, S.112/113]:
Verwende soviele Hauptkomponenten, bis ein beliebig festgelegter Anteil c% der Gesamtvarianz p der Daten durch sie erklärt ist. Sei s_k^2 die Varianz der k-ten Hauptkomponente, so gilt für den Anteil c_m% der ersten m Komponenten $c_m = \frac{100}{p} \sum_{k=1}^{m} s_k^2$ und die Anzahl k der Hauptkomponenten ist definiert durch:

$$k = \min \left\{ m | \sum_{k=1}^{m} s_k^2 \geqq \frac{c \cdot p}{100} \right\} \tag{3.3}$$

3.2 Varianz der Hauptkomponenten (Kaiser-Kriterium)

Das zweite Kriterium bezieht sich auf die empirische Korrelationsmatrix $\boldsymbol{R}$. Berücksichtigt werden alle Hauptkomponenten mit Eigenwerten $\lambda_m \geqq 1$. Man wählt also die Hauptkomponenten aus, die mindestens die Varianz 1 erklären. Wären alle Elemente in den x_i unabhängig, dann wären die Hauptkomponenten identisch zu den ursprünglichen Variablen, im Fall funktionaler Daten die Beobachtungszeitpunkte t, und hätten alle Varianz 1. Somit enthält eine Hauptkomponente mit Varianz kleiner 1 weniger Information als die ursprünglichen Variablen und es ist daher nicht Wert, sie zu behalten [Jolliffe, 2002, S.114]. Für die Anzahl an Hauptkomponenten k gilt also:

$$k = \max \{ m | \lambda_m \geqq 1 \} \tag{3.4}$$

3.3 Scree-Test

Betrachtet man ein Diagramm der Eigenwerte $\lambda_1, ..., \lambda_{min(n,T)}$ von $\boldsymbol{R}$ aus den vorliegenden Daten (Scree-Plot), so lässt sich meist ein deutlicher Knick ausmachen, bis zu dem die Eigenwerte relativ große Werte annehmen und danach flach abfallen (Beispiel: siehe Abbildung 2).

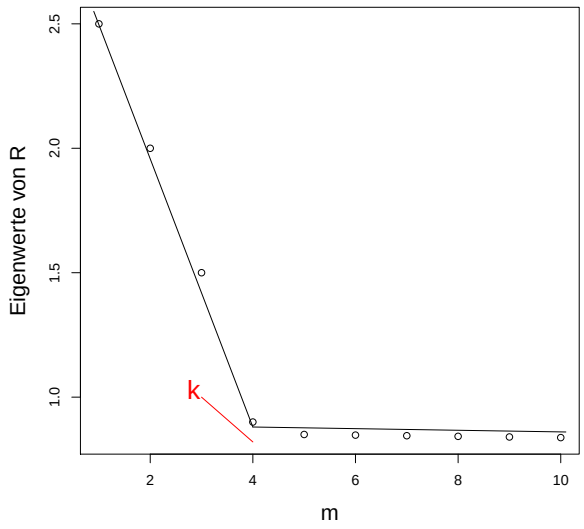


Abbildung 2: Beispielhaftes Diagramm der Eigenwerte von $\boldsymbol{R}$ *(Scree-Plot).*

Eigenwerte von empirischen Korrelationsmatrizen normalverteilter, unabhängiger Zufallszahlen zeigen typischerweise einen flach und gleichmäßig abfallenenden Verlauf. Man nimmt daher an, dass Hauptkomponenten mit Eigenwerten nach dem Knick nur noch zufällig sind und wählt die Knickstelle k als Kriterium. Das Kriterium beruht auf der Eigenschaft normalverteilter Daten [Fahrmeir et al., 1996, S.669].

Die Knickstelle im Scree-Plot für jeden Datensatz in einer Simulation graphisch zu bestimmen, wäre zu aufwendig. Mathematisch soll die Knickstelle über die Methode der sogenannten „Optimalen Koordinaten“ bestimmt werden. Diese Methode ist eine der mathematischen Lösungen, die in der Funktion *nScree* aus dem Packet „nFactors“ implementiert ist [Raiche and Magis, 2010].
$\lambda_1, ..., \lambda_p$ seien die Eigenwerte von $\boldsymbol{R}$, mit $p = min(n, T)$. Man betrachtet jeweils ein lineares Regressionsmodell zwischen λ_p und λ_{i+1} , $i = 1, ..., p-1$, also zwischen dem letzten und dem (i+1)-ten Eigenwert. Als „Optimale Koordinaten“ bezeichnet man die extrapolierten Koordinaten an der Stelle i durch das Regressionsmodell mit λ_{i+1}.

Insgesamt gibt es $p-2$ dieser Regresssionsmodelle.
Seien nun $\beta_0^{(i+1)}$ und $\beta_1^{(i+1)}$ die Koeffizienten des Regressionsmodells mit λ_{i+1} und

$$\hat{\lambda}^{(i)} = \beta_0^{(i+1)} + i\beta_1^{(i+1)} \tag{3.5}$$

der extrapolierte Eigenwert an der Stelle i, so betrachet man folgende Bedingungen:

1. $\lambda_i > \hat{\lambda}^{(i)}$
2. $\lambda_i > 1$

Solange beide Bedingungen wahr sind, betrachtet man jeweils ein weiteres Regressionsmodell. Sobald eine der Bedingungen nicht mehr wahr ist, wählt man als Anzahl der zu verwendenden Hauptkomponenten $k = i$.
Die zweite Bedingung entpricht dem Kriterium 3.2. Ist also zunächst Bedingung 2 nicht mehr wahr, so entsprechen sich bei dieser mathematischen Auswertung das Kaiser-Kriterium und der Scree-Test.

3.4 Test nach Bartlett

Der Test nach Bartlett prüft die Signifkanz der Hauptkomponenten. Wie auch der Scree-Test (Abschnitt 3.3) setzt dieser Test die Normalverteilung der Daten voraus [Fahrmeir et al., 1996, S.670]. Es wird getestet, ob sich die $p-k$ kleinsten Eigenwerte $\lambda_{k+1}, ..., \lambda_p$ von $\boldsymbol{R}$ noch signifikant unterscheiden und man deshalb noch weitere Hauptkomponenten berücksichtigen sollte. $p = min(n, T)$ bezeichnet die maximale Anzahl an Hauptkomponenten. Die Extraktion der Hauptkomponenten erfolgt schrittweise. Die Nullhypothese zur Bestimmung von k lautet:

$$H_0 : \lambda_{k+1} = ... = \lambda_p \tag{3.6}$$

Als Teststatistik für $0 < k < p-1$ verwendet man:

$$U_k = (n-1)[-\ln(|\boldsymbol{R}|) + \ln(\prod_{i=1}^{k} \lambda_i) + (p-k)\ln(\lambda)]\,, \tag{3.7}$$

$$\text{mit} \quad \lambda = \frac{p - \sum_{i=1}^{k} \lambda_i}{p-k}$$

U_k ist annähernd χ^2-verteilt mit $d_k = \frac{1}{2}(p-k+2)(p-k-1)$ Freiheitsgraden.
H_0 wird abgelehnt und die Extraktion einer weiteren Hauptkomponente erscheint als sinnvoll, wenn gilt:

$$U_k > \chi^2(d_k, 1-\alpha) \tag{3.8}$$

$\chi^2(d_k, 1-\alpha)$ ist das $(1-\alpha)$-Quantil der Chi-Quadrat-Verteilung mit d_k Freiheitsgraden. α bezeichnet das Signifikanzniveau.
Für kleine Stichprobengrößen n bringt $n^* = n - k - \frac{1}{6}(2(p-k)+1+\frac{2}{p-k})$ in U_k eine bessere Anpassung an die χ^2-Verteilung.

Die beiden nächsten Kriterien zur Bestimmung der Anzahl k an Hauptkomponenten in Abschnitt 3.5 und 3.6 sind speziell für den Fall funktionaler Daten und basieren auf dem Modell (2.32) [Yao et al., 2005].

3.5 Bestimmung von k durch Kreuzvalidierung

Das folgende Kriterium bezieht sich auf das Modell (2.32). Es soll geprüft werden, wie gut die Datenmatrix $\boldsymbol{Y}$ durch die Wahl von jeweils m Hauptkomponenten geschätzt wird. [Yao et al., 2005] schlägt vor, die $y_i(t)$ jeweils durch die leaving-one-out Kreuzvalidierung zu schätzen, d.h. $\hat{y}_i^{(-i)}(t)$ ist die geschätzte Kurve nachdem aus dem Datensatz die i-te Beobachtung herausgenommen wurde. Aufgrund des hohen Rechenaufwands wird hier das Kriterium basierend auf einer 5-fach Kreuzvalidierung vorgestellt. Dieses wird auch in der Simulation in Kapitel 4 verwendet. Eine Einführung hierzu bietet [Hastie et al., 2009, Kap. 7].
Bei der 5-fach Kreuzvalidierung wird die Datenmatrix $\boldsymbol{Y}$ in fünf gleich große Teile geteilt. Sei $\kappa : \{1, ..., n\} \longmapsto \{1, ..., 5\}$, eine Funktion, die jeder Beobachtung i einen Index 1 bis 5 zuordnet. κ teilt $\boldsymbol{Y}$ also in fünf Teile. Bezeichne nun $\hat{y}^{(-d)}$, die geschätzten Kurven nachdem der d-te Teil der Beobachtungen aus dem Datensatz herausgenommen wurde, $d = 1, ..., 5$, so gilt:

$$\hat{y}_i^{(-d)}(t) = \hat{\mu}^{(-d)}(t) + \sum_{m=1}^{k} \hat{\xi}_{im}^{(-d)} \hat{\phi}_m^{(-d)}(t) \tag{3.9}$$

$\hat{\xi}_{im}^{(-d)}$ wird über den bedingten Erwartunswert (2.38) berechnet. $\hat{\mu}^{(-d)}$ ist die geschätzte Mittelwertsfunktion und $\hat{\phi}_m^{(-d)}$ die geschätzten Eigenfunktionen nach Herausnahme des d-ten Teils der Beobachtungen.
Der zu berechnende Kreuzvalidierungsscore mit k Hauptkomponenten lautet schließlich [Yao et al., 2005]:

$$CV(k) = \sum_{d=1}^{5} \sum_{j=1}^{T} \left\{ y_d(j) - \hat{y}_d^{(-d)}(j) \right\}^2 \tag{3.10}$$

Man wählt dasjenige k für welches der Kreuzvalidierungs-Score (CV) **minimal** wird und somit die geschätzte Datenmatrix $\hat{y}_i$ am besten die wahre Datenmatrix $y_i(t)$ approximiert.

3.6 Bestimmung von k durch das cAIC

Eine weitere Möglichkeit die Anzahl der Hauptkomponenten k zu bestimmen ist das folgende cAIC-Kriterium [Yao et al., 2005].
Geht man von den geschätzten Gewichten $\hat{\xi}_{im}$ aus und betrachtet das Modell (2.32) mit m Hauptkomponenten, so gilt für den bedingten Erwartungswert der $\hat{y}_i$:

$$\mathbb{E}(\hat{y}_i|\hat{\xi}_{im}) = \hat{\mu}_i + \sum_{m=1}^{k} \hat{\xi}_{im}\hat{\phi}_{im} \tag{3.11}$$

Eine log-Likelihood L ist dann gegeben durch:

$$\begin{aligned} \hat{L}(k) = \sum_{i=1}^{n} \Bigg\{ & -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\hat{\sigma}^2) \\ & -\frac{1}{2\hat{\sigma}^2} \left(y_i^T - \underbrace{\hat{\mu}_i + \sum_{m=1}^{k} \hat{\xi}_{im}\hat{\phi}_{im}}_{\mathbb{E}(\hat{y}_i|\hat{\xi}_{im})} \right)^T \times \left(y_i^T - \underbrace{\hat{\mu}_i + \sum_{m=1}^{k} \hat{\xi}_{im}\hat{\phi}_{im}}_{\mathbb{E}(\hat{y}_i|\hat{\xi}_{im})} \right) \Bigg\} \end{aligned} \tag{3.12}$$

Man definiert schließlich AIC(k) = -$\hat{L}(k) + k$.
Man wählt dasjenige k für welches das AIC **minimal** wird.

Das letzte Kriterium, das in Abschnitt 3.7 vorgestellt wird, basiert auf der Idee gemischter Modelle und ist entnommen aus dem Paper „On the behaviour of marginal and conditional AIC in linear mixed models“ [Greven and Kneib, 2010].

3.7 cAIC für gemischte Modelle

Lineare gemischte Modelle sind aussagekräftige Instrumente der Inferenzstatistik und finden eine breite Anwendung, wie z.B. in der funktionalen Datenanalyse. Gemischte Modelle bieten Flexibilität und eignen sich für komplexe Modelle großer Datensätze. Die erwünschte Flexibilität und die Komplexität des Modells machen eine Modelwahl zunehmend wichtig. Das beinhaltet die Selektion von zufälligen Effekten, beispielsweise solchen, die die Heterogenität zwischen Subjekten modellieren. Das Paper [Greven and Kneib, 2010] behandelt die Eigenschaften des Akaike-Informations-Kriteriums, AIC, für die Selektion von zufälligen Effekten.
Betrachtet wird das lineare gemischte Modell:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{3.13}$$

$\boldsymbol{X}$ und $\boldsymbol{Z}$ sind bekannte Designmatrizen mit vollem Rang, $\boldsymbol{\beta}$ beinhaltet feste Parameter und man nimmt an, dass $\mathbf{b}$ und $\boldsymbol{\varepsilon}$ unabhängig und normalverteilt sind.
In [Greven and Kneib, 2010] wird als Möglichkeit zur Modellselektierung, vor allem der zufälligen Effekte b, das bedingte Akaike-Informations-Kriterium (cAIC), basierend auf der bedingten Verteilung $\boldsymbol{y}|\boldsymbol{b} \sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b},\, \sigma^2\mathbb{1}_n\right)$, mit $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{1}_n$ vorgestellt.
Betrachtet man das hier zugrunde liegende Modell (2.32), so kann man dieses auch als lineares gemischtes Modell auffassen. Sei $\boldsymbol{y} \in \mathbb{R}^{T\text{x}1}$, $\boldsymbol{\Phi} \in \mathbb{R}^{T\text{x}k}$ die Designmatrix aus bekannten Basisfunktionen, $\boldsymbol{\xi} \in \mathbb{R}^{k\text{x}1}$ und $\boldsymbol{\varepsilon} \in \mathbb{R}^{T\text{x}1}$, so lautet (2.32) mit k Hauptkomponenten in Matrixschreibweise:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{3.14}$$

Es wird wieder angenommen, dass $\boldsymbol{\xi}$ und $\boldsymbol{\varepsilon}$ unabhängig und normalveteilt sind. 3.14 entspricht genau dem zweiten Teil $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$ des Modells (3.13). Betrachtet man also ein Kriterium zur Selektion von zufälligen Effekten in einem linearen gemischten Modell, so

kann man dies auch verwenden, um Modelle mit verschieden vielen Hauptkomponenten zu vergleichen.

Die bedingte log-Likelihood L ist hier gegeben durch:

$$\hat{L} = -\frac{1}{2\hat{\sigma}^2}\left(\boldsymbol{y} - \boldsymbol{\Phi}\hat{\boldsymbol{\xi}}\right)^T\left(\boldsymbol{y} - \boldsymbol{\Phi}\hat{\boldsymbol{\xi}}\right) \tag{3.15}$$

Beim AIC-Kriterium (Abschnitt 3.6) wird k, die Anzahl an betrachteten Hauptkomponenten, als Zahl der Freiheitsgrade verwendet. [Greven and Kneib, 2010] stellt ein cAIC vor, bei dem die effektiven Freiheitsgrade alternativ geschätzt werden. Dieses korrigierte cAIC [Greven and Kneib, 2010, Kap. 4.2] ist ebenfalls in die Simulationsstudie (Kapitel 4) mit aufgenommen.

4 Simulationsstudie

4.1 Simulationsaufbau

Die sieben Selektionskriterien zur Wahl der Anzahl der Hauptkomponenten bei der funktionalen Hauptkomponentenanalyse, die in Kapitel 3 vorgestellt wurden, werden nun in einer Simulationsstudie auf ihre Güte getestet und verglichen.

Die Datensätze sind nach dem in Abschnitt 2.3.1 vorgestellten Modell (2.32) gebildet. $y_i(t) = \mu(t) + \sum_{m=1}^{N} \xi_{im}\phi_m(t) + \varepsilon_i(t)$. Für alle Szenarien gelte, dass $\mu(t) = 0$, d.h. der Mittelwert jeder Spalte von $\boldsymbol{Y} \in \mathbb{R}^{nxT}$ ist bis auf Abweichungen durch den Messfehler ε gleich Null.

Die Funktionen $y_i(t)$ werden an einem gleichabständigen Gitter $t \in \tau = \left\{\frac{x}{T}, x = 1, ..., T\right\}$ beobachtet.

Für die Simulationsstudie wurde eine Grundeinstellung aller Paramter des Modells (2.32), die von Interesse sind, gewählt und anschließend jeder der Parameter zweimal varriert.

Das Modell beinhaltet folgende relevante Parameter:

1. n: Anzahl der Beobachtungen y_i

   n $\in \{50, 100, 200\}$

2. T: Anzahl der Gitterpunkte, an denen die $y_i(t)$ beobachtet werden

   T $\in \{50, 100, 200\}$

3. σ^2: Varianz des Messfehlers ε_i

   $\sigma \in \{0.01, 0.05, 0.1\}$

4. N: Anzahl der verwendeten Eigenfunktionen

   N $\in \{2, 4, 8\}$

5. λ_m^2: Varianz der normalverteilten ξ_m

   Für die Varianzen λ_m^2 gilt: $\lambda_1 > \lambda_2 > ... > \lambda_N$

   $\lambda_m = \frac{0.25}{2^{x-1}}$ („e“), $\lambda_m = \frac{1}{x}$ („h“), $x = 1, ..., N$ bzw. $\lambda_m \in [2, 1]$ („l“)

6. ϕ_m: Verwendete Eigenfunktionen

   Als Eigenfunktionen werden trigonometrische Funktionen („trig“) bzw. orthogonale Polynome („pol“) verwendet.

Für die Grundeinstellung gilt:

$$n = 200,\ T = 100,\ \sigma = 0.05,\ N = 4,\ \lambda_m = \text{„h“},\ \phi_m = \text{„trig“} \tag{4.1}$$

Variiert man jeden Parameter zweimal und die verwendeten Eigenfunktionen einmal, so erhält man insgesamt zwölf Szenarien mit folgenden Parameterkombinationen (siehe Tabelle 1).

Szenario	1	2	3	4	5	6	7	8	9	10	11	12
n	200	50	100	200	200	200	200	200	200	200	200	200
T	100	100	100	50	200	100	100	100	100	100	100	100
σ	0.05	0.05	0.05	0.05	0.05	0.01	0.1	0.05	0.05	0.05	0.05	0.05
N	4	4	4	4	4	4	4	2	8	4	4	4
λ	„h“	„h“	„h“	„h“	„h“	„h“	„h“	„h“	„h“	„e“	„l“	„h“
ϕ	„trig“	„trig“	„trig“	„trig“	„trig“	„trig“	„trig“	„trig“	„trig“	„trig“	„trig“	„pol“

Tabelle 1: Übersicht über die Parameterkombinationen der ersten zwölf betrachteten Szenarien in der Simualtion.

Die, in Tabelle 1 unterstrichenen Werte, werden im jeweiligen Szenario verändert. Für jedes der betrachteten Szenarien werden 100 Datensätze generiert. Diese sind in der Datei *Datensaetze.RData* gespeichert (siehe Anhang A).

Für N=4 sind die Werte von λ in Abbildung 3 dargestellt. Die λ_m werden zweimal variiert. „l“ steht für linear. Die Werte fallen zwischen 2 und 1 linear ab. Sie sind eindeutig von Null verschieden. „e“ steht für exponentiell und meint, dass die Funktionswerte einer Exponentialfunktion folgen, sie fallen von 0.25 exponentiell ab. „h“ steht für hyperbolisch. In diesem Fall liegen die Funktionswerte auf einer Hyperbel und fallen von der 1 hyperbolisch ab. Sie nähern sich nicht so schnell der Null, wie bei den exponentiellen Werten „e“.

Die Eigenfunktionen ϕ für N=4 sind in Tabelle 2 aufgelistet. Im ersten Fall werden die trigonometrischen Funktionen *sin* und *cos* verwendet. Im zweiten Fall wird eine

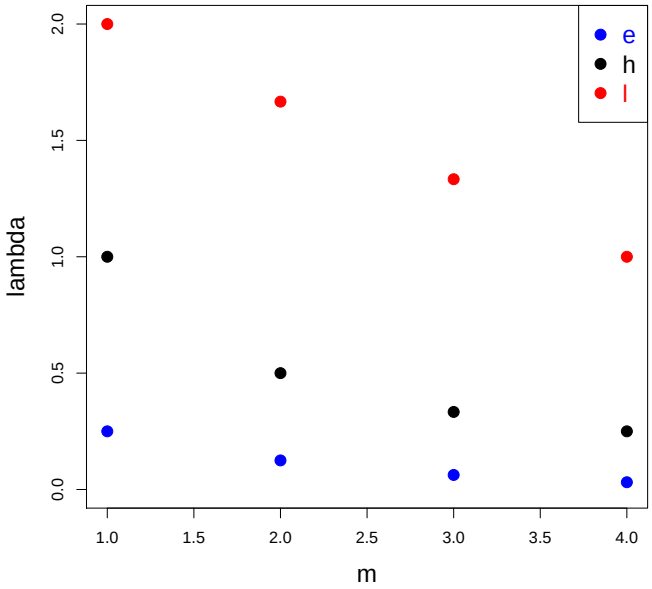


Abbildung 3: Werte für die Standardabweichungen λ_m für N=4. Betrachtet werden je drei Variationen „e“, „h“ und „l“.

ϕ	„trig “	„pol “
ϕ_1	$\sin(2\pi t)$	1
ϕ_2	$\cos(2\pi t)$	$\sqrt{3}(2t-1)$
ϕ_3	$\sin(4\pi t)$	$\sqrt{5}(6t^2-6t+1)$
ϕ_4	$\cos(4\pi t)$	$\sqrt{7}(20t^3-30t^2+12t-1)$

Tabelle 2: Die verwendeten Eigenfunktionen ϕ für N=4. Betrachtet werden trigonometrische Funktionen („trig“) und orthogonale Polynome („pol“).

Form der sogenannten „Legendre-Polynome“ verwendet, nämlich die „verschobenen“ Legendre-Polynome. Diese sind auf dem Interval (0,1) definiert und bilden ein orthogonales Funktionssystem [E.Weisstein, 2011]. Implementiert sind die Polynome in R im Packet „orthopolynom“ [Novomestky, 2009].

Die Eigenfunktionen ϕ_m werden jeweils normiert, um die Orthonormalität sicherzustellen. In beiden Fällen liegt somit, wie gefordert, ein orthonormales Basissystem vor.

Da bei der Bestimmung der Anzahl zu verwendender Hauptkomponenten die Anzahl der im wahren Modell enthaltenen Eigenfunktionen (N) und die Größe der im Modell enthaltenen Varianz der Gewichte ξ_m (λ_m^2) wohl eine entscheidende Rolle spielen, sollen

alle Kombinationsmöglichkeiten dieser beiden Paramter betrachtet werden. Dies sind alle Tupel $\{N, \lambda\}$, wobei $N \in \{2, 4, 8\}$, $\lambda \in \{e, h, l\}$. Bei konstant halten aller anderen Parameter sind das insgesamt 9 Szenarien. Zusätzlich zu den 12 Szenarien aus Tabelle 1 werden also noch die, in Tabelle 3 aufgelisteten Szenarien, betrachtet. Die Szenarien 7 bis 11 und 13 bis 16 enthalten alle Kombinationen der Paramter N und λ.

Szenario	13	14	15	16
n	200	200	200	200
T	100	100	100	100
σ	0.05	0.05	0.05	0.05
$\underline{\text{N}}$	2	2	8	8
$\underline{\lambda}$	„e“	„l“	„e“	„l“
ϕ	„trig“	„trig“	„trig“	„trig“

Tabelle 3: Szenarien 13 bis 16 der Simulation. Variiert werden die Anzahl N an Eigenfunktionen und die Varianzen der Gewichte ξ_m (λ_m^2).

Für alle 16 Szenarien aus Tabelle 1 und 3 wurden die Kriterien 3.1 bis 3.7 und die daraus resultierende Anzahl zu verwendender Hauptkomponenten bestimmt. Die Ergebnisse, Interpretationen und der Vergleich der Kriterien werden im Folgenden (Abschnitt 4.3) dargestellt. Zunächst werden noch einige allgemeine Ergebnisse präsentiert.

4.2 Ausgewählte Ergebnisse

Bevor genauer auf die Auswertungen der Selektionskriterien aus Kapitel 3 eingegangen wird, werden in diesem Abschnitt einige andere Ergebnisse aus der Analyse vorgestellt. Detailiert betrachtet werden die Szenarien 1 und 12 aus Tabelle 1.

Abbildung 4 zeigt für Szenario 1 die ersten fünf Beobachtungen y_i für zwei Wiederholungen. Im Modell mit aufgenommen sind in Abbildung 4 die vier trigonometrischen Funktionen (siehe erste Spalte aus Tabelle 2). Vor der Generierung der Datensätze mit der Software R wurde jeweils ein Seed gesetzt, um die Ergebnisse reproduzierbar zu machen.

Für Szenario 12 erhält man analog folgende Beobachtungen (siehe Abbildung 5). Abgebildet sind in Abbildung 5 wieder die ersten fünf Beobachtungen für zwei Wiederholungen. Im Modell mit aufgenommen sind diesmal die vier orthogonalen "Legendre-

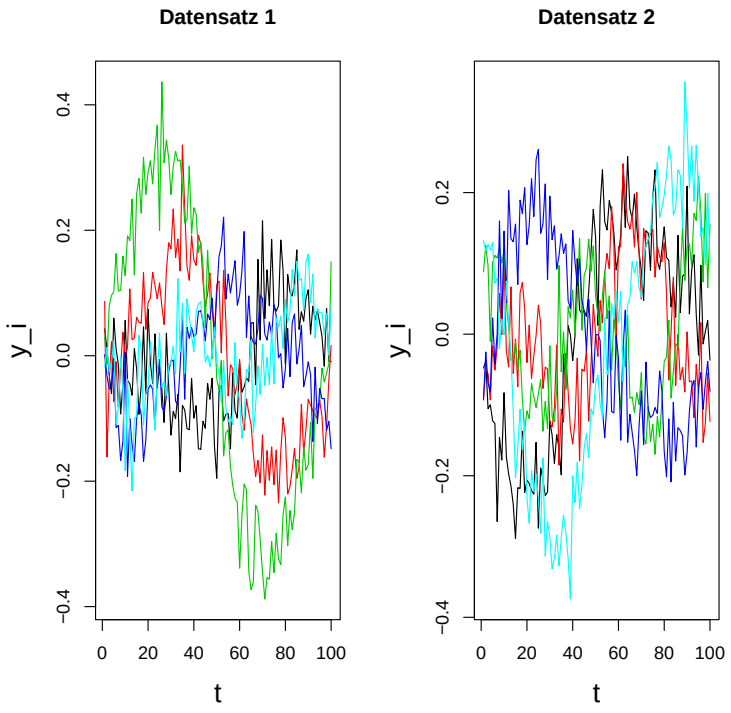


Abbildung 4: Die ersten fünf Beobachtungen y_i zweier Datensätze für Szenario 1 der Simulation.

Polynome" (siehe zweite Spalte aus Tabelle 2).

Die im Folgenden präsentierten Schätzergebnisse wurden durch die Funktion LFPCAg geschätzt. Der zugehörige R-Code ist in der Datei *LFPCAg_newsigma.r* verfügbar (siehe Anhang A). Das Schätzverfahren, welches die Funktion anwendet, ist identisch zur Schätzung der Modellkomponenten, wie in Abschnitt 2.3.2 beschrieben. Es besteht lediglich ein Unterschied. Die Glättung der Modellkomponenten geschieht nicht wie beschrieben durch lokale polynomiale Glättung, sondern durch penalisierte Splines. Eine gute Einführung hierzu bietet [Fahrmeir et al., 2007].

Bei **penalisierten Splines** nimmt man zunächst an, dass man die Funktion f(x) als Linearkombination bekannter Basisfunktionen darstellen kann, d.h.

$$f(x_i) = \sum_{j=1}^{m} \underbrace{\delta_j}_{\text{unbekannt}} \underbrace{B_j(x_i)}_{\text{bekannt}} \tag{4.2}$$

m ist die Anzahl verwendeter Basisfunktionen.
Meist gebräuchliche Basisfunktionen sind B-Splines. In Matrixschreibweise lautet das

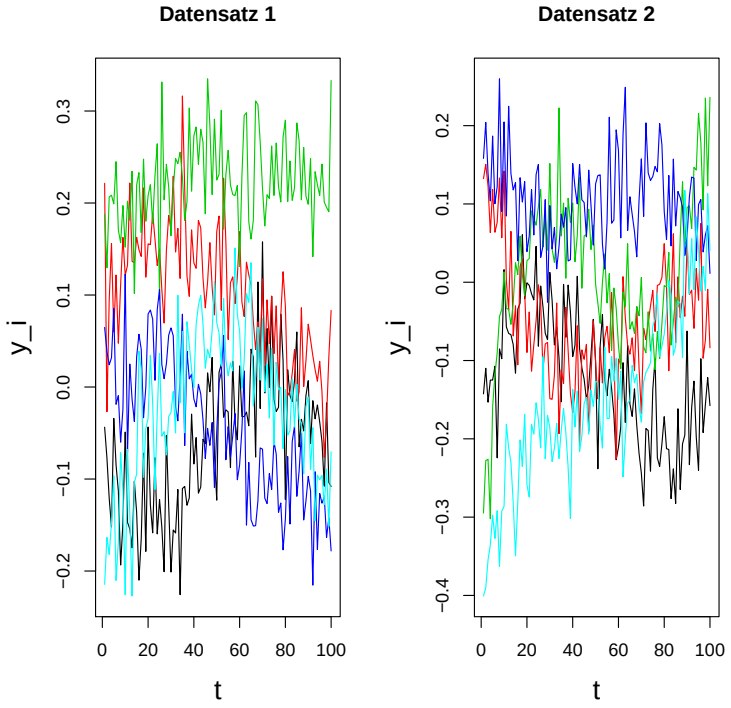


Abbildung 5: Die ersten fünf Beobachtungen y_i zweier Datensätze für Szenario 12 der Simulation.

Modell: $f = \mathbf{B}\boldsymbol{\delta}$. Penalisiert man aufeinanderfolgende Koeffizienten, so erhält man für die Berechnung des nonparametrischen Regressionsmodells folgendes Minimierungsproblem:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\delta_j B_j(x_i))^2 + \lambda \sum_{j=d+1}^{m}(\Delta^d \delta_j)^2 \underset{\delta}{\longrightarrow} \min \tag{4.3}$$

Δ^d steht für die Differenz d-ter Ordnung. $\sum_{j=d+1}^{m}(\Delta^d\delta_j)^2 = \boldsymbol{\delta}^T\mathbf{D}_d^T\mathbf{D}_d\boldsymbol{\delta} = \boldsymbol{\delta}^T\mathbf{K}_d\boldsymbol{\delta}$. Als Lösung des Minimierungsproblems (4.3) erhält man:

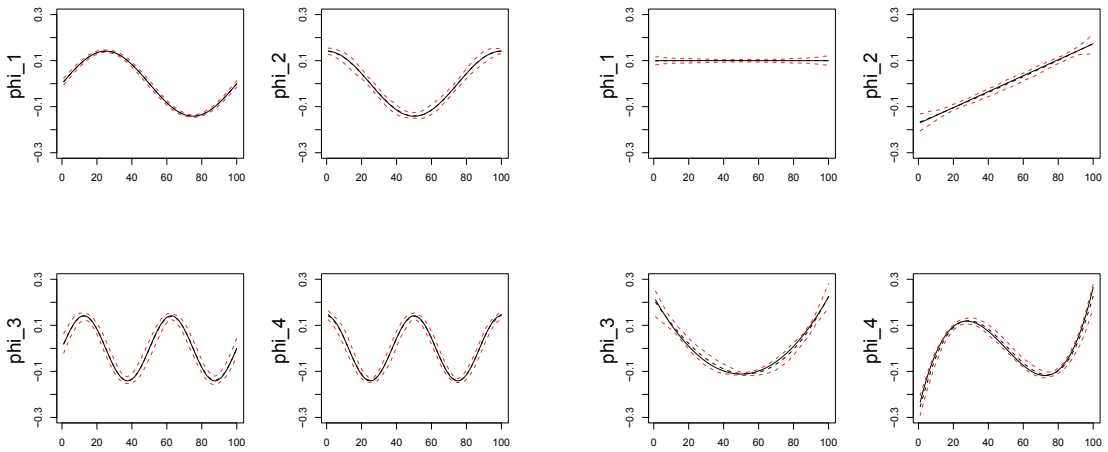
$$\hat{\delta} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{K}_d)^{-1}\mathbf{B}^T\boldsymbol{y} \tag{4.4}$$

Die Lösung entspricht einem KQ-Schätzer mit additivem Term. Im Fall, dass $\lambda = 0$ ist, erhält man den klassischen KQ-Schätzer $\hat{\delta} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\boldsymbol{y}$.

Der Funktionsaufruf lautet:

$$LFPCAg(Y = Y, groups = matrix(NA, n, 0), Zvars = list(),$$
$$NPC = m, smooth = T, bf = 15, smoothalg = "gamGCV")$$

Geschätzt wurden jeweils m Eigenfunktionen (*NPC=m*) für Datensatz Y. *smooth=T*, bedeutet, dass die Varianz-Kovarianz-Matrix glatt geschätzt wird. *smoothalg=„gamGCV“* legt die verwendete Glättungsmethode fest. Zur Glättung wurden 15 Basisfunktionen verwendet (*bf=15*). Die Argumente *groups* und *Zvars* beziehen sich auf eine komplexere longitudinale Datenstruktur mit zusätzlichen Gruppierungsvariablen, die hier nicht betrachtet wird. Sie wurden als leere Argumente übergeben.
Abbildung 6 zeigt das Ergebnis der Schätzung der Eigenfunktionen ϕ_m.



(a) Szenario 1 (b) Szenario 12

Abbildung 6: Wahre und geschätzte Eigenfunktionen ϕ_m. Abgebildet sind die vier wahren Eigenfunktionen (schwarze, durchgezogene Linie), der Mittelwert der geschätzten Eigenfunktionen der 100 Simulationen (schwarze, gestrichelte Linie) und punktweise das 5%- und 95%-Quantil der geschätzten Funktionen (rote, gestrichelte Linien).

Abgetragen sind jeweils die vier wahren Eigenfunktionen (schwarze, durchgezogene Linie), der Mittelwert der geschätzten Eigenfunktionen der 100 Simulationen (schwarze, gestrichelte Linie) und punktweise das 5%- und 95%-Quantil der geschätzten Funktionen (rote, gestrichelte Linien). Für Szenario 1 (Abbildung 6a) sind es die normierten

Funktionen sin(2πt), cos(2πt), sin(4πt) und cos(4πt), für Szenario 12 (Abbildung 6b) die normierten Funktionen 1, $\sqrt{3}(2t-1)$, $\sqrt{5}(6t^2-6t+1)$ und $\sqrt{7}(20t^3-30t^2+12t-1)$. Für alle Funktionen gilt, dass die geschätzten Kurven sehr nahe an die wahren Funktionen herankommen und die Variabiltät sehr gering ist.
Abbildung 7 zeigt das Ergebnis der Schätzung der Eigenwerte (λ_m^2).

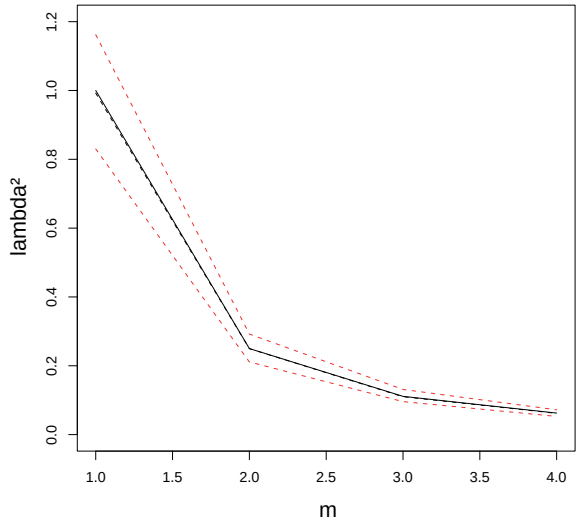


Abbildung 7: Wahre und geschätzte Eigenwerte λ_m^2 für Szenario 1. Abgebildet sind die wahren Eigenwerte (schwarze, durchgezogene Linie) der Mittelwert der geschätzten Eigenwerte der 100 Simulationen (schwarze, gestrichelte Linie) und punktweise das 5%- und 95%-Quantil der geschätzten Eigenwerte (rote gestrichelte Linien).

Geschätzt wurden vier, zu den Eigenfunktionen aus Abbildung 6 gehörende, Eigenwerte. Im Modell 2.32 sind dies die Varianzen der Gewichte bzw. Hauptkomponentenscores ξ_m. Zu sehen sind die wahren, ins Modell aufgenommenen Werte λ_m^2 (schwarze, durchgezogene Linie) der Mittelwert der geschätzten Eigenwerte der 100 Simulationen (schwarze, gestrichelte Linie) und punktweise das 5%- und 95%-Quantil der geschätzten Eigenwerte (rote gestrichelte Linien). Die Ergebnisse für Szenario 1 und Szenario 12 sind für $\lambda =$„h“ nahezu identisch und daher in Abbildung 7 nur für Szenario 1 abgetragen.
Auch wie schon in Abbildung 6 kann man feststellen, dass die geschätzten Werte sehr nahe an die wahren Werte herankommen und die Streuung sehr gering ist.
Insgesamt funktioniert also die Schätzung der Eigenfunktionen und Eigenwerte mit der Funktion *LFPCAg* sehr gut. Es reichen 15 Basisfunktionen zur Schätzung aus. Die Zahl

der Basisfunktionen (*bf*) und die die Art der Glättung (*smoothalg*) steuern sehr stark die Laufzeit der Funktion *LFPCAg*. Auf dem verwendeten Server konnte die Laufzeit durch Veränderung dieser Parameter von 40 min auf 1 min reduziert werden, ohne, wie an den Ergebnissen sichtbar, die Güte der Schätzung zu verschlechtern. Die Einstellung *bf=15* und *smoothalg=„gamGCV“* beim Funktionsaufruf von *LFPCAg* wurde für die komplette Simulation gewählt.

4.3 Auswertung der Selektionskriterien

In diesem Kapitel werden die Ergebnisse der Simualtion dargestellt. Gespeichert sind diese in der Datei *Ergebnisse.RData* (siehe Anhang A). Berechnet wurden alle 16 Szenarien, die in Abschnitt 4.1 vorgestellt wurden. Die Funktion *noPC* ("number of Principal Components") berechnet die Kriterien jeweils für einen Datensatz. Der zugehörige R-Code ist in der Datei *NoPC.r* verfügbar (siehe Anhang A). Für jedes Szenario wurden jeweils 100 Datensätze generiert. Das Kriterium durch Kreuzvalidierung (Abschnitt 3.5) wurde zunächst weggelassen und gesondert berechnet, da es, wie später diskutiert, für den hier betrachteten funktionalen Fall nicht funktioniert. Es wurde aufgrund des sehr hohen Rechenaufwands nur für zwei Szenarien beispielhaft berechnet. Anhand dessen wird die Problematik ersichtlich (siehe Abschnitt 4.3.7).
Die Auswertung Abschnitt 4.3 umfasst drei Teile:

1. Einzelne Betrachtung der Selektionskriterien (Abschnitt 4.3.1 bis 4.3.7).
2. Vergleich und Bewertung der multivariaten Kriterien und der cAIC-Kriterien (Abschnitt 4.3.8).
3. Vergleich der verschiedenen Szenarien (Abschnitt 4.3.9).

4.3.1 Erklärung eines Anteils der Gesamtvarianz

Im Output der Funktion *noPC* ist die Anzahl an Hauptkomponenten, die einen Anteil der Gesamtvarianz erklären (siehe Abschnitt 3.1), die Variable **nErkVar**. In der Simulation sollen so viele Hauptkomponenten bestimmt werden bis 95% der Gesamtvarianz erkärt sind.
Die Ergebnisse für die Szenarien **1** bis **5** und **10** sind sehr ähnlich. Tabelle 4 beinhaltet die absoluten Häufigkeiten der bestimmten Anzahl an Hauptkomponenten. Im Modell

sind jeweils N=4 Eigenfunktionen enthalten. Jeweils in ca. 80% der Fälle werden nur drei Hauptkomponenten gewählt, ansonsten korrekterweise vier.

nErkVar	3	4
Szenario 1	84	16
Szenario 2	82	18
Szenario 3	74	26
Szenario 4	85	15
Szenario 5	88	12
Szenario 12	84	16

Tabelle 4: Absolute Häufigkeiten der Anzahl an Hauptkomponenten, die bei Szenarien 1 bis 5 und 12 95% der Varianz erklären.

Man sieht also, dass die Veränderung der Paramter n und T das Ergebnis nicht beeinflussen. Auch für Szenario 12, welches als Eigenfunktionen die „Legendre-Polynome“ besitzt, sind die Ergebnisse identisch.
Verändert man die Fehlervarianz σ, so werden wieder jeweils drei oder vier Hauptkomponenten gewählt, jedoch im Verhältnis **90/10** für $\sigma = 0.01$ (Szenario 6) und im Verhältnis **64/36** für $\sigma = 0.1$ (Szenario 7). Wird das sogenannte *signal-to-noise ratio* durch Verkleinerung/Vergrößerung der Fehlervarianz größer/kleiner, werden tendenziell weniger/mehr Hauptkomponenten ausgewählt.
Tabelle 5 stellt die Ergebnisse der Szenarien bei Veränderung der Parameter N und λ dar.

		N		
		2	4	8
	„l“	2	4	7
λ	„h“	2	3	5
	„e“	2	3	3

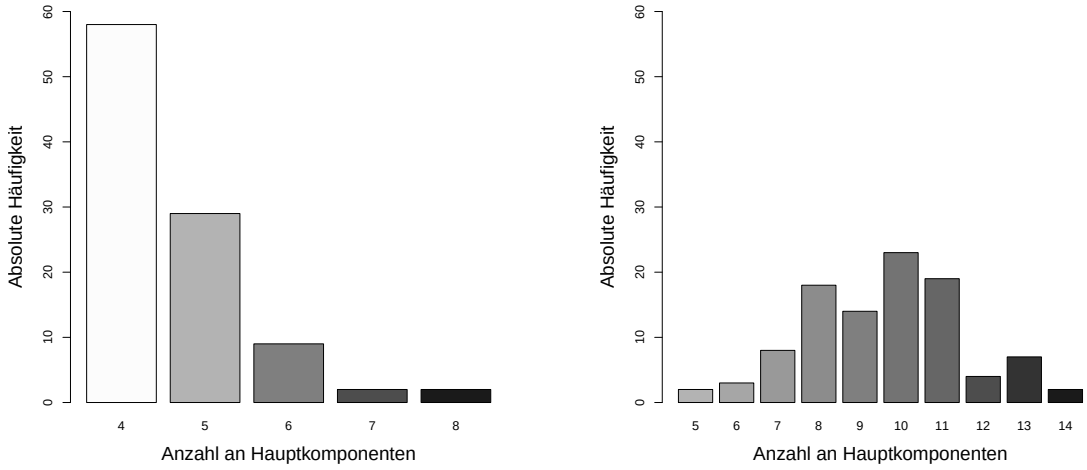
Tabelle 5: Anzahl an Hauptkomponenten, die 95% der Varianz erklären. Variiert werden die Paramter N und λ.

Für $N = 2$ werden jeweils korrekt zwei Hauptkomponenten ausgewählt. Mit steigender Anzahl N und kleiner werdenden Werten für λ werden die Ergebnisse schlechter.

Für $N = 8$ und kleinen Werten λ („e“) werden lediglich drei Hauptkomponenten ausgewählt.
In Tabelle 5 ist der Übersichtlichkeit halber jeweils nur die am häufigsten gewählte Anzahl an Hauptkomponenten aufgelistet, da nur in zwei der betrachteten Szenarien die Anzahl nicht in allen Fällen identisch ist. Grün markiert sind die Szenarien, bei denen die Anzahl korrekt spezifiert wurde, orange markiert sind díe Szenarien, bei denen die Anzahl sehr nahe an die wahre Anzahl herangeht und die eindeutig falsch spezifizierten Szenarien sind rot markiert.
Insgesamt wählt dieses Kriterium tendenziell zu wenige Hauptkomponenten. Im Fall großer Werte für λ und einer kleinen Anzahl an Eigenfunktionen im Modell funktioniert die Selektion sehr gut.

4.3.2 Varianz der Hauptkomponenten (Kaiser-Kriterium)

Im Output der Funktion *noPC* ist die Anzahl an Hauptkomponenten nach dem Kaiser-Kriterium (siehe Abschnitt 3.2), die Variable **nKaiser**. Betrachtet man alle Kriterien, so schneidet das Kaiserkriterium insgesamt sehr gut ab. Bei **9** der **16** betrachteten Szenarien wird in 100% der Fälle die Anzahl zu verwendender Hauptkomponenten korrekt spezifiert. Es treten lediglich einzelne Szenarien auf, bei denen das Kriterium schlecht oder gar nicht funktioniert.
In Abbildung 8 sind die Anzahl an Hauptkomponenten nach dem Kaiser-Kriterium für Szenario 2 und 5 abgetragen.
Die Balken in der Graphik sind farblich abgestuft. Die Abstufung geht von **weiß** (Anzahl an HK richtig spezifiert) über **grau** (Anzahl an HK falsch spezifiert) bis hin zu **schwarz** (deutlich falsche Spezifikation der HK). Diese Abstufung gilt für alle weiteren Graphiken, insbesondere der Balkendiagramme, die während der Auswertung vorgestellt werden.
Bei beiden Szenarien sind jeweils vier Eigenfunktionen im wahren Modell vorhanden. Bei Szenario 2 (Abbildung 8a) sind n=50 und T=100, d.h. $\mathbf{n<T}$. Bei Szenario 5 (Abbildung 8b) sind n=200 und T=200, d.h. $\mathbf{n=T}$. Im Fall multivariater Daten gilt meist, dass $\mathbf{p<<n}$. Somit lässt sich erklären, warum das multivariate Kaiser-Kriterium in diesen Fällen nicht so gut funktioniert.
Ein weiterer Fall, bei dem das Kaiser-Kriterium nicht funktioniert, ist Szenario 7. Hier



(a) Szenario 2

(b) Szenario 5

Abbildung 8: Anzahl an Hauptkomponenten nach dem Kaiser-Kriterium.

gilt für die Fehlervarianz $\sigma = 0.1$. Im wahren Modell sind wieder N=4 Eigenfunktionen vorhanden. Die Anzahl bestimmter Hauptkomponenten ist deutlich zu groß (siehe Tabelle 6). Bei hoher Fehlervarianz sind also folglich wesentlich mehr Eigenwerte größer 1.

nKaiser	7	8	9	10	11	12	13	14
abs. Häufigkeit	1	3	15	18	36	19	7	1

Tabelle 6: Anzahl an Hauptkomponenten nach dem Kaiser-Kriterium für Szenario 7.

Als letztes wird noch der Fall $\lambda =$„e“ analysiert. Auch hier funktioniert die Selektion anhand des Kaiserkriteriums nicht. $\lambda =$„e“ bedeutet, dass die Varianzen der Gewichte ξ sehr kleine Werte annehmen. Dies ist bei Szenario 10, 13 und 15 der Fall. Abbildung 9 zeigt beispielhaft die Auswertung des Kaiser-Kriteriums für Szenario 15.
Im wahren Modell sind N=8 Eigenfunktionen vorhanden. Ausgewählt werden zwischen 29 und 34 Hauptkomponenten. Auch in diesem Fall sind also deutlich mehr Eigenwerte der Korrelationsmatrix größer 1.

4.3.3 Scree-Test

Im Output der Funktion *noPC* ist die Anzahl an Hauptkomponenten, die durch den Scree-Test bestimmt wird (siehe Abschnitt 3.3), die Variable **nScree**. Wie auch beim

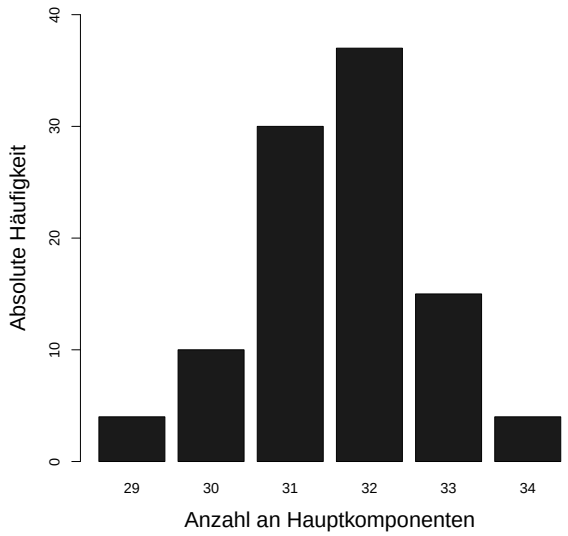


Abbildung 9: Anzahl an Hauptkomponenten nach dem Kaiser-Kriterium für Szenario 15. Im Modell vorhanden sind N=8 Eigenfunktionen.

Kaiser-Kriterium bezieht sich der Scree-Test auf die Größe der Eigenwerte der Korrelationsmatrix. Bei der mathematischen Bestimmung des Knicks des Scree-Plots über die „Optimalen Koordinaten“, wie in Kapitel 3.3 erläutert, stimmen die beiden Kriterien überein, falls das zweite Abbruchkriterium, dass $\lambda_i > 1$ ist, als erstes nicht mehr erfüllt ist. Diese Tatsache spiegelt sich auch in den Ergebnissen wider. Es zeigt sich ein sehr ähnliches Bild wie beim Kaiser-Kriterium (Abschnitt 4.3.2).

Für **9** der **16** Szenarien gilt ebenfalls, dass die Anzahl der selektierten Hauptkomponenten in allen Fällen mit der wahren Anzahl übereinstimmt. Ausnahmen gelten wieder für die Szenarien 2 (n<T), 5 (n=T=200) und 7 ($\sigma = 0.1$). Für die Szenarien 2 und 5 erhält man das analoge Ergebnis wie beim Kaiserkriterium (siehe Abbildung 8). Das Ergebnis für Szenario 7 ist in Abbildung 10 zu sehen. Im wahren Modell sind N=4 Eigenfunktionen vorhanden. Der Scree-Test wählt hier in den meisten Fällen deutlich zu viele Hauptkomponenten aus.

Änlich wie für das Kaiser-Kriterium zeigen sich auch die Ergebnisse der Szenarien 10, 13 und 15, bei denen gilt $\lambda =$„e“. Analog zu Abbildung 9 zeigt Abbildung 11 beispielhaft das Ergebnis des Scree-Tests für Szenario 15. Im Modell sind N=8 Eigenfunktionen vorhanden. Man sieht, dass die Werte sehr stark um den wahren Wert 8 streuen. Nur in 10 von 100 Fällen werden korrekterweise acht Hauptkomponenten ausgewählt.

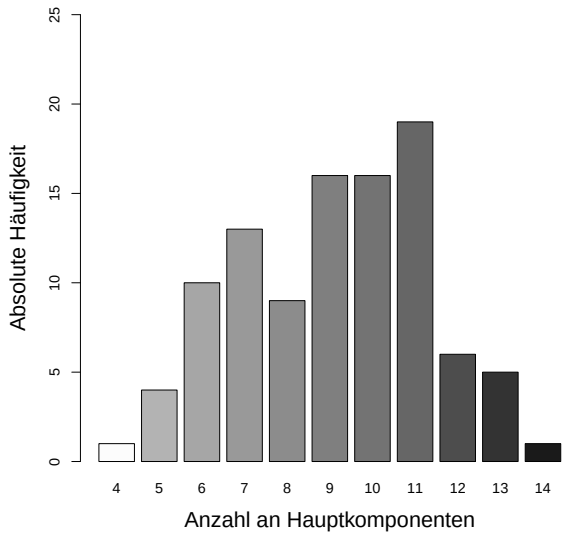


Abbildung 10: Anzahl an Hauptkomponenten nach dem Scree-Test für Szenario 7. Im Modell vorhanden sind N=4 Eigenfunktionen.

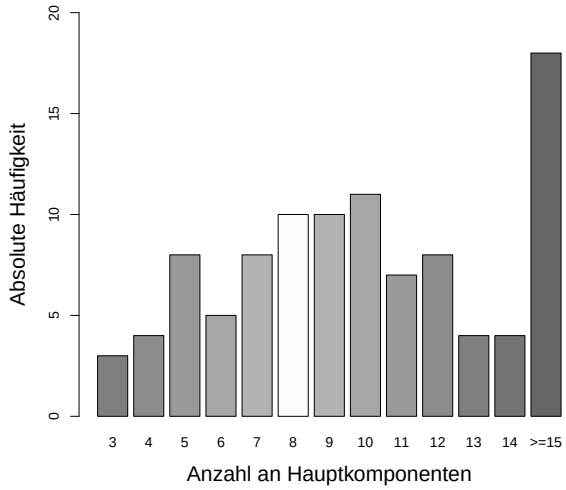


Abbildung 11: Anzahl an Hauptkomponenten nach dem Scree-Test für Szenario 15. Im Modell vorhanden sind N=8 Eigenfunktionen.

4.3.4 Test nach Bartlett

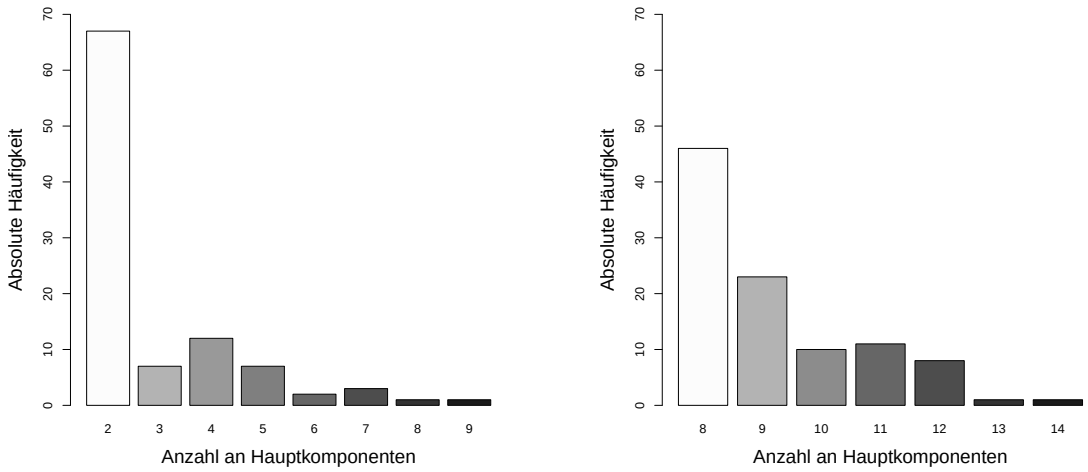
Im Output der Funktion *noPC* ist die durch den Signifikanztest nach Bartlett (siehe Abschnitt 3.4) bestimmte Anzahl an Hauptkomponenten, die Variable **nBartlett**. Die Teststatistik U_k enthält die Determinante der Korrelationsmatrix $\boldsymbol{R}$. Diese ist nur ungleich Null, falls $\boldsymbol{R}$ vollen Rang hat, d.h. n>T gilt. Für die Szenarien 2, 3 und 5 ist der

Signifikanztest nach Bartlett also nicht durchführbar.
Für alle anderen Szenarien ist der Test zwar berechenbar, funktioniert jedoch nicht. Tabelle 7 zeigt auswahlsweise die absoluten Häufigkeiten der bestimmten Anzahl an Hauptkomponenten des Bartlett-Tests für Szenario 6, 9 und 12.

nBartlett	14	15	16	17	18	19	>20
Szenario 6 (N=4)	2	0	0	1	2	2	93
Szenario 9 (N=8)	0	3	1	3	4	9	80
Szenario 14(N=2)	0	0	0	1	0	2	97

Tabelle 7: Absolute Häufigkeiten der Anzahl an Hauptkomponenten nach dem Bartlett-Test für Szenario 6, 9 und 14.

Die drei Szenarien sind Beispiele dafür, dass der Bartlett-Test nicht funktioniert. Der Test bestimmt immer deutlich zu viele Hauptkomponenten. Auch bei Szenario 6, dem Szenario mit kleiner Fehlervarianz ($\sigma = 0.01$) ist das Ergebnis nicht besser.
Es gibt vereinzelte Szenarien für die der Bartlett-Test eingeschränkt funktioniert. Betrachtet werden auswahlsweise Szenario 13 und 16 (siehe Abbildung 12).



(a) Szenario 13 (b) Szenario 16

Abbildung 12: Anzahl an Hauptkomponenten nach dem Bartlett-Test.

Bei Szenario 13 gilt N=2 und $\lambda =$„e“. Ausgewählt werden in 67% der Fälle korrekterweise zwei Hauptkomponenten. Bei Szenario 16 gilt N=8 und $\lambda =$„l“. Ausgewählt werden in 46% der Fälle korrekterweise acht Hauptkomponenten. Das Ergebnis erschließt

sich nicht, da bei Szenario 14, bei dem auch gilt $\lambda =$„l“, mehr als 17 Hauptkomponenten bestimmt werden.

Insgesamt zeigt die Studie, dass der Signifikanztest nach Bartlett sehr schlecht auf den Fall funktionaler Daten angewendet werden kann.

Ein Versuch die Rohdaten $\boldsymbol{Y}$ beispielsweise durch die beiden schon vorgestellten Ansätze der Lokalisierung oder penalisierter Splines zu glätten und danach den Signifikanztest nach Bartlett anzuwenden, scheitert, da es bei der Glättung zu einem Rangabfall kommt und daher wieder die Determinante von $\boldsymbol{R}$ gleich Null ist. Es gibt keine Möglichkeit zu testen, ob der Signifikanztest mit geglätteten Daten besser funktionieren würde.

4.3.5 cAIC nach [Yao et al., 2005]

Im Output der Funktion *noPC* ist die Anzahl an Hauptkomponenten nach dem cAIC-Kriterium, welches im Paper [Yao et al., 2005] vorgestellt wird (siehe Abschnitt 3.6), die Variable **nAIC1**. Betrachtet man alle Ergebnisse, so funktioniert dieses AIC insgesamt nicht gut. Bei **11** der **16** betrachteten Szenarien wird die Anzahl der Hauptkomponenten deutlich überschätzt. Ausgewählt ist in Tabelle 8 das Szenario 9 als Beispiel.

nAIC1	7	8	9	10	11	12	13
abs. Häufigkeit	1	0	0	15	59	22	3

Tabelle 8: Anzahl an Hauptkomponenten nach dem cAIC-Kriterium nach [Yao et al., 2005] für Szenario 9.

Bei Szenario 9 sind im wahren Modell N=8 Eigenfunktionen enthalten. In 99 von 100 Fällen werden zehn oder mehr Hauptkomponenten durch das Kriterium bestimmt. Farblich grün gekennzeichnet sind in der Tabelle die korrekte Anzahl an Eigenfunktionen im Modell, rot gekennzeichnet ist die maximal ausgewählte Zahl an Eigenfunktionen.

Bei insgesamt **4** der **16** Szenarien wird die Anzahl an Hauptkomponenten eher unterschätzt. Als Beispiel wird Szenario 16 betrachtet (siehe Tabelle 9).

Bei Szenario 16 sind ebenso N=8 Eigenfunktionen enthalten. In 62 von 100 Fällen werden sieben oder weniger Hauptkomponenten durch das Kriterium bestimmt.

nAIC1	≤3	4	5	6	7	8	9	10	11	12	13
abs. Häufigkeit	14	10	2	5	31	0	0	16	1	10	11

Tabelle 9: Anzahl an Hauptkomponenten nach dem cAIC-Kriterium nach [Yao et al., 2005] für Szenario 16.

Auffallend ist bei allen Szenarien, dass unabhängig davon, ob die wahre Anzahl an Eigenfunktionen im Modell eher unterschätzt oder eher überschätzt wird, die wahre Anzahl selbst nie explizit ausgewählt wird.
Von allen Szenarien sticht als einziges Szenario 15 heraus (siehe Abbildung 13).

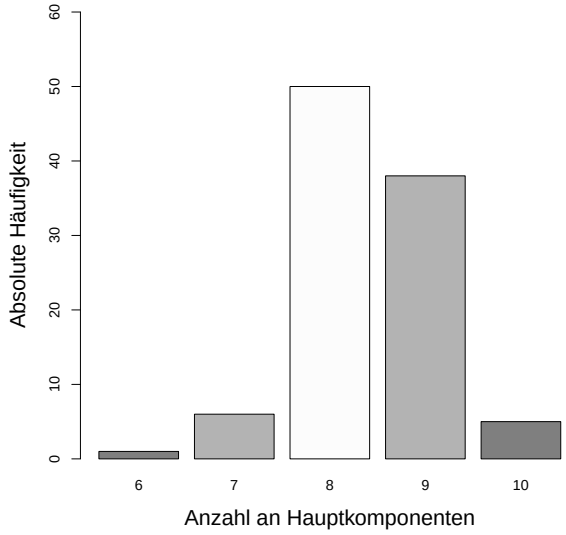


Abbildung 13: Anzahl an Hauptkomponenten nach dem cAIC-Kriterium nach [Yao et al., 2005] für Szenario 15. Im Modell vorhanden sind N=8 Eigenfunktionen.

Auch bei Szenario 15 sind N=8 Eigenfunktionen im wahren Modell enthalten. In 50 von 100 Fällen werden korrekterweise acht Eigenfunktionen spezifiert. Ansonsten streuen die Werte nur leicht um den wahren Wert 8. Dies ist der einzige Fall, bei dem das cAIC-Kriterium nach [Yao et al., 2005] gut funktioniert.

4.3.6 cAIC für gemischte Modelle

Im Output der Funktion *noPC* ist die Anzahl an Hauptkomponenten nach dem cAIC-Kriterium für gemischte Modelle (siehe Abschnitt 3.7), die Variable **nAIC2**. Betrachtet man alle Auswertung für dieses Kriterium, so kann man eindeutige Tendenzen ausma-

chen. Die Güte der Schätzung hängt stark vom Paramter N ab, also der Zahl der Eigenfunktionen im Modell. Mit steigendem N wird die Schätzung deutlich schlechter. Für die Kriterien 8, 13 und 14 werden korrekterweise zwei Hauptkomponenten selektiert.

Für die neun Kriterien, bei denen N=4 Eigenfunktionen im Modell enthalten sind, wird die Anzahl immer leicht unterschätzt. In fast allen Fällen werden drei statt vier Hauptkomponenten ausgewählt.

Für die Kriterien 9, 15 und 16, bei denen N=8 Eigenfunktionen im Modell enthalten sind, wird die Anzahl immer sehr stark unterschätzt. Abbildung 14 zeigt beispielhaft die Auswertung für Szenario 16. Die Ergebnisse streuen relativ gleichmäßig zwischen 1 und 7.

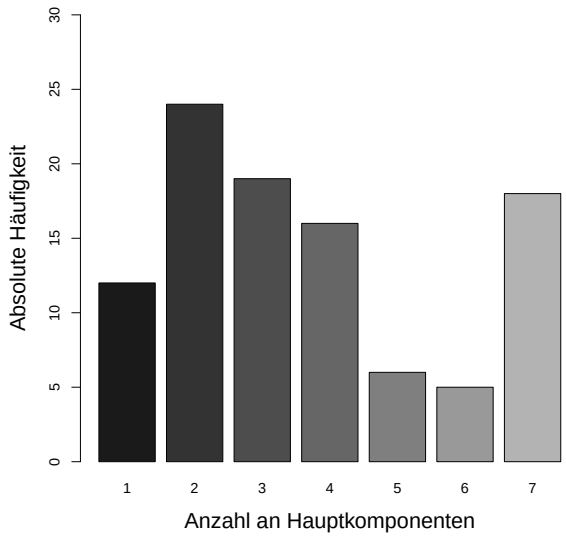


Abbildung 14: Anzahl an Hauptkomponenten nach dem cAIC-Kriterium für gemischte Modelle für Szenario 16. Im Modell vorhanden sind N=8 Eigenfunktionen.

Beim cAIC für gemischte Modelle fällt zuletzt Szenario 12 auf, für das gilt, dass ϕ =„poly“, also als Eigenfunktionen die „Legendre-Polynome“ verwendet wurden. Es ist das einzige Szenario mit N=4 Eigenfunktionen bei dem das cAIC für gemischte Modelle immmer korrekterweise vier Hauptkomponenten selektiert.

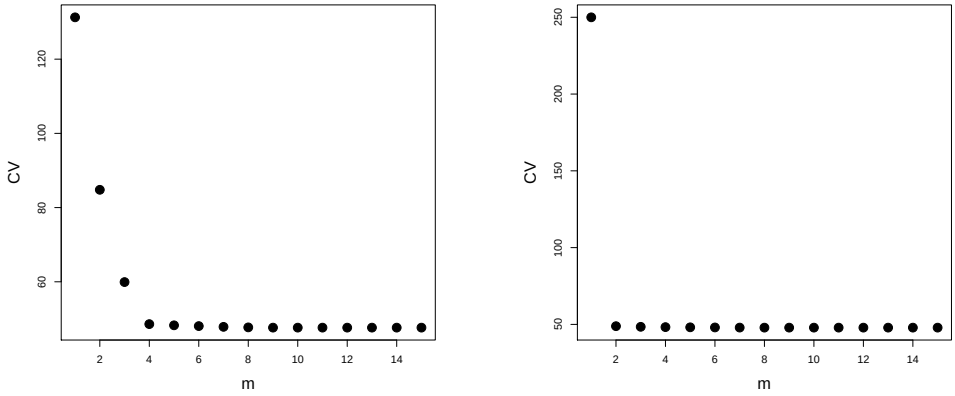
4.3.7 Bestimmung von k durch Kreuzvalidierung

Als letztes, der in Kapitel 3 vorgestellten Kriterien, wird das Kriterium durch Kreuzvalidierung analysiert. Bei der Berechnung der Anzahl an Hauptkomponenten durch die Funktion *noPC* wurde dieses Kriterium zunächst außen vor gelassen. Es wurde anschließend gesondert nur für die Szenarien 1 und 14 berechnet. Gespeichert sind diese in der Datei *ErgebnisseCV.RData*. Dies hat zwei Gründe:

1. Aufgrund der Kreuzvalidierung und der daraus resultierenden Anzahl zu berechnender Modelle durch die Funktion *LFPCAg*, ist die Laufzeit für die Auswertung deutlich höher, als die der restlichen Simulation.

2. Wie sich herausstellt,erhält man bei der Berechnung der CV-Werte kein eindeutiges Minimum. Das Kriterium ist für den hier betrachteten Fall funktionaler Daten nicht geeignet.

Berechnet wurden für die beiden Szenarien jeweils Modelle mit einschließlich 15 Eigenfunktionen und daraus jeweils der Wert für CV. Diese sind in Abbildung 15 abgetragen.



(a) Szenario 1 (b) Szenario 14

Abbildung 15: Die berechneten CV-Werte für die Modelle mit bis zu 15 Eigenfunktionen. Aus den Werten lässt sich kein Minimum ablesen.

Aus der Abbildung wird ersichtlich, dass die Werte kein eindeutiges Minimum erreichen. Die Selektierung der Anzahl k funktionaler Hauptkomponenten, wie in Abschnitt 3.5 beschrieben, funktioniert im betrachteten funktionalen Fall nicht.

Das Problem liegt bei der Bestimmung der $\hat{\xi}_{im}^{(-d)}$. Diese sollen, wie bei [Yao et al., 2005] beschrieben, durch den bedingten Erwartungswert (2.38) berechnet werden. Berechnet man ein Modell mithilfe der Funktion *LFPCAg* ohne den d-ten Teil der Beobachtungen, so erhält man die geschätzten Eigenfunktionen $\hat{\phi}_m^{(-d)}$,die geschätzten Eigenwerte $\hat{\lambda}_m^{(-d)}$ und die geschätzte Kovarianzmatrix $\hat{\mathbf{\Sigma}}^{(-d)}$. Diese hängen nicht von i ab. Zur Schätzung von $\hat{y}_i^{(-d)}$ werden die $\hat{\xi}_{im}^{(-d)}$ für den d-ten Teil der Daten dann berechnet durch:

$$\hat{\xi}_{im}^{(-d)} = \hat{E}[\xi_{im}^{(-d)}|\tilde{\boldsymbol{y}}_i] = \hat{\lambda}_m^{(-d)}\hat{\boldsymbol{\phi}}_m^{(-d)^T}\hat{\mathbf{\Sigma}}_{y_i}^{(-d)-1}\tilde{\boldsymbol{y}}_i \tag{4.5}$$

$\tilde{\boldsymbol{y}}_i = (y_{i1}, ..., y_{iT})^T$ stellt den Vektor der Beobachtungen y_i dar. Somit wird in (2.38) zur Schätzung der $\hat{\xi}_{im}^{(-d)}$ der Teil der Daten verwendet, der vorher aus dem Datensatz herausgenommen wurde. Das ist der Grund warum die Kreuzvalidierung so nicht funktioniert. Das Ergebnis wird immer besser je mehr Eigenfunktionen verwendet werden. Die Schätzung $\hat{y}_i^{(-d)}$ ist perfekt, falls $m = min(n, T)$, die maximale Anzahl an Hauptkomponenten, verwendet wird.
Im Paper [Yao et al., 2005] wird das Kriterium durch Kreuzvalidierung speziell für den Fall spärlicher Daten vorgeschlagen, also für den Fall, dass von einer Datenkurve y_i nicht zu jedem Zeitpunkt t Messungen vorhanden sind. Alternativ könnte man also einige Punkte aus den Datenkurven weglassen und darüber die CV-Werte berechnen. Das wäre jedoch sehr aufwendig und rechenintensiv.
Ein weiterer Ansatz die CV-Werte für eine Auswertung zu nutzen, ist eine ähnliche Betrachtung wie beim Scree-Test. Betrachtet man Abbildung 15a, so sieht diese einem Scree-Plot wie in Abbildung 2 sehr ähnlich. Eine Überlegung wäre, so viele Hauptkomponenten ins Modell aufzunehmen bis die CV-Werte nicht mehr stark abnehmen. Als Kriterium könnte man die Knickstelle im Graphen der CV-Werte verwenden.

4.3.8 Vergleich und Bewertung der Kriterien

Nachdem die Ergebnisse der einzelnen Kriterien vorgestellt und analysiert wurden, werden die Kriterien verglichen und ihre Nützlichkeit bewertet.

Betrachtet man die **multivariaten Kriterien**, so kann man zunächst sagen, dass der Bartlett-Test, wie schon in Abschnitt 4.3.4 erläutert, sehr schlecht für den hier betracheteten Fall funktionaler Daten geeignet ist. Bei der Selektierung von funktionalen

Hauptkomponenten sollte er nicht zu Rate gezogen werden.
Sehr gut geeignet sind das Kaiserkriterium und der Scree-Test. Bei der Analyse des Scree-Tests in Abschnitt 4.3.3 wurde bereits eingehend die Parallelität dieser beiden Kriterien erläutert. Sowohl die Szenarien bei denen die Selektion fehlerfrei funktioniert als auch die Szenarien bei denen sie weniger gut funktioniert sind identisch. Für kleine Werte von λ, also bei den Szenarien wo gilt λ=„e“, ist jedoch ein Unterschied ersichtlich. Abbildung 16 zeigt beispielhaft einen Scree-Plot für Datensatz 90 von Szenario 10 (N=4).

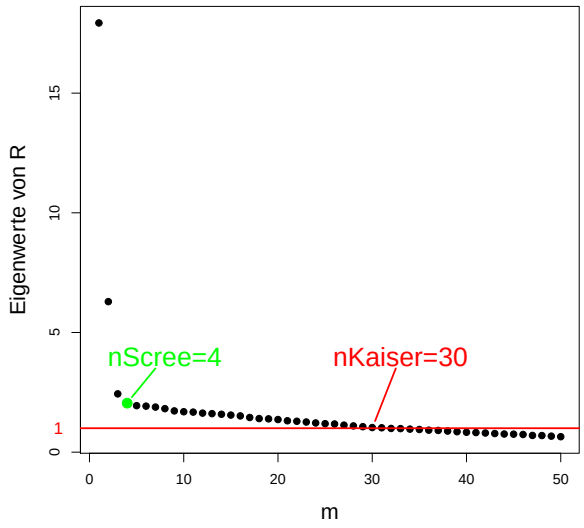


Abbildung 16: Scree-Plot für Datensatz 90 von Szenario 10. Gekennzeichnet sind die Ergebnisse des Scree-Tests und des Kaiser-Kriteriums.

In diesem Fall selektiert der Scree-Test über die „Optimalen Koordinaten“ korrekterweise 4 Hauptkomponenten (nScree=4). Erst Eigenwert 31 ist jedoch kleiner 1 und somit funktioniert das Kaiser-Kriterium nicht (nKaiser=30). Für kleine Werte von λ funktioniert das Kaiser-Kriterium nie (vgl. Abbildung 9), wohingegen die selektierten Werte des Scree-Tests zwar tendenziell zu groß sind, aber um den wahren Wert streuen (vgl. Abbildung 11).
Es ist zu bemerken, dass das Kaiserkriterium und der Scree-Plot die Anzahl an Hauptkomponenten wenn zu hoch schätzen, jedoch niemals unterschätzen.

Vergleicht man die beiden **cAIC-Kriterien**, so sieht man, dass das cAIC für gemischte Modelle besser abschneidet.

- Beim cAIC für gemischte Modelle lassen sich Szenarien ausmachen, bei denen die Selektierung immer gut funktioniert. Die Selektierung ist abhängig vom Parameter N. Mit steigender Zahl an Eigenfunktionen werden leicht bis deutlich zu wenige Hauptkomponenten selektiert.

- Beim cAIC nach [Yao et al., 2005] streuen die Werte sehr stark um den wahren Wert, vor allem weichen sie extrem nach oben ab. Es lässt sich keine Tendenz ausmachen, durch welche Parameter die Güte der Schätzung beeinflusst wird.

Der Hauptunterschied, der beiden cAIC-Kriterien ist die zur Berechnung verwendete Anzahl an Freiheitsgraden. Beim cAIC nach [Yao et al., 2005] sind dies jeweils m, die Zahl an berechneten Eigenfunktionen. Beim cAIC für gemischte Modelle werden diese alternativ berechnet (siehe [Greven and Kneib, 2010]). Um einen Eindruck zu erhalten sind in Tabelle 10 für die Szenarien 12, 13 und 14 die berechneten Freiheitsgrade beispielhaft für einen Datensatz abgetragen.

Szenario	12	13	14
dfAIC1	4	2	2
dfAIC2	785.8	363.8	399.4

Tabelle 10: Freiheitsgrade für die Berechnung der cAIC-Werte beispielhaft für je einen Datensatz der Szenarien 12, 13 und 14.

Berechnet wurden sie jeweils für die Modelle mit der korrekten Anzahl an Eigenfunktionen, also N=2 bzw. N=4. Die Freiheitsgrade sind von deutlich unterschiedlicher Größenordnung. Für das cAIC für gemischte Modelle sind die Werte viel größer. Die Selektierung funktioniert für die drei gewählten Szenarien mit dem cAIC für gemischte Modelle gut, mit dem cAIC nach [Yao et al., 2005] jedoch nicht.

Ein Blick lohnt sich auch auf das multivariate Kriterium der **erklärten Varianz**. Insgesamt schneidet es nicht schlecht ab. Aus Tabelle 5 wird ersichtlich, dass die Güte der Selektierung stark von N und λ abhängt. Je größer N, desto schlechter funktioniert die Selektierung mit Hilfe der erklärten Varianz. Dieser Effekt spiegelt sich, wie schon erwähnt, auch beim cAIC für gemischte Modelle wider. Vergleicht man diese beiden Kriterien sieht man sehr starke Parallelen:

- Beide Kriterien bestimmen für die Szenarien mit N=2 die Anzahl an Hauptkomponenten immer korrekt.
- Für N=4 unterschätzen beide Kriterien die Anzahl an Hauptkomponenten leicht. In den meisten Fällen werden drei Hauptkomponenten selektiert.
- Für N=8 wird die Anzahl an Hauptkomponenten durch beide Kriterien stark unterschätzt. Beim Kriterium der erklärten Varianz hängt die Unterschätzung zusätzlich von λ ab.

Zusammenfassend:

Von den sieben ausgewerteten Kriterien eignen sich vier für die Selektion funktionaler Hauptkomponenten. Nützlich sind drei multivariate Kriterien, nämlich das Kriterium der erklärten Varianz, das Kaiserkriterium und der Scree-Test. Außerdem in Betracht zu ziehen ist das cAIC für gemischte Modelle.
Anhand des Selektionsverhaltens kann man jeweils zwei Kriterien zusammenfassen:

1. **Kaiser-Kriterium** & **Scree-Test**: In den wenigen Fällen in denen die Selektion nicht gut funktioniert, wird die wahre Anzahl an Eigenfunktionen überschätzt.
2. **Erklärung eines Anteils der Varianz** & **cAIC für gemischte Modelle**: In den Fällen in denen die Selektion nicht gut funktioniert, wird die wahre Anzahl an Eigenfunktionen unterschätzt.

4.3.9 Vergleich der verschiedenen Szenarien

Anhand der, für die Selektion funktionaler Hauptkomponenten geeigneter, Kriterien, wie im vorherigen Abschnitt 4.3.8 diskutiert, wird die Simulationsstudie jetzt nicht aus Sicht der einzelnen Kriterien beläuchtet, sondern nochmal aus Sicht der einzelnen Szenarien. Es soll nochmal herausgestellt werden, welche Paramter des Modells (2.32) besonderen Einfluss auf die Güte der Selektion haben.

Insgesamt sehr gut schneiden die Szenarien 8 und 14 ab. Hier spezifizieren alle vier oben genannten Kriterien in 100% der Fälle die korrekte Anzahl an Eigenfunktionen, welche jeweils N=2 ist. Es gilt $\lambda =$„h“ bzw. $\lambda =$„l“.

Bei Szenario 3, welches N=2 Eigenfunktionen enthält und bei dem gilt λ =„e“, also λ sehr kleine Werte annimmt, funktionieren nur das Kriterium der erklärten Varianz und das cAIC für gemischte Modelle. Das Kaiser-Kriterium und der Scree-Plot sind, wie schon bekannt, bei sehr kleinen Werten von λ nicht geeignet.
Ein weiteres Szenario, bei dem alle drei multivariaten Szenarien immer korrekt selektieren, ist Szenario 11. Hier nimmt λ große Werte an. Also λ =„1“. Im wahren Modell sind vier Eigenfunktionen enthalten.
Alles in allem erhält man sehr gute Ergebnisse für Modelle mit wenigen Eigenfunktionen und großen Varianzen der Gewichte ξ.

Insgesamt unbefriedigend schneiden die Szenarien 5, 7 und 10 ab. Hier spezifizieren alle vier oben genannten Kriterien die Anzahl der Eigenfunktionen von N=4, deutlich falsch.

- Bei Szenario 5 gilt, dass n=200 und T=200. Das Kriterium der erklärten Varianz und das cAIC unterschätzen mit N=3 die Zahl an Eigenfunktionen, wohingegen das Kaiser-Kriterium und der Scree-Test die Zahl an Eigenfunktionen deutlich überschätzt. Wie schon erläutert, liegt das schlechte Abschneiden der multivariaten Kriterien daran, dass für multivariate Daten im Normalfall gilt, dass p<<T.

- Bei Szenario 10 gilt, dass λ =„e“. Es ist das Szenario mit N=4 und kleinen Werten für λ. Analog unterschätzen das Kriterium der erklärten Varianz und das cAIC die Zahl an Eigenfunktionen, während das Kaiser-Kriterium und der Scree-Test die Zahl an Eigenfunktionen deutlich überschätzt.

- Szenario 7 enthält eine größere Fehlervarianz. Hier gilt $\sigma = 0.1$. Trotzdem das Kriterium der erklärten Varianz aufgrund der kleineren *signal-to-noise ratio* im Vergleich zu den anderen Szenarien mit N=4 etwas besser abschneidet (siehe Abschnitt 4.3.1), wird die wahre Anzahl an Eigenfunktionen von allen vier Kriterien analog unterschätzt bzw. überschätzt.

Auffallend ist auch, wie schon in Abschnitt 4.3.5 gezeigt, das Szenario 15. Es ist das dritte Szenario mit λ =„e“, hier sind N=8 Eigenfunktionen im wahren Modell. Die drei multivariaten Kriterien und das cAIC für gemischte Modelle spezifizieren die Anzahl an Eigenfunktionen immer falsch. Einzig das cAIC-Kriterium nach [Yao et al., 2005] selektiert in 50 von 100 Fällen korrekterweise acht Hauptkomponenten.

5 Anwendung

Die vorgestellten Selektionskriterien werden in diesem Abschnitt auf einen realen Datensatz angewendet. Anhand der Simulationsergebnisse aus Abschnitt 4.3 soll die empfohlene Anzahl zu verwendender funktionaler Hauptkomponenten diskutiert werden.
Betrachtet werden Wetterdaten aus Kanada. Verfügbar ist dieser Datensatz in R im Packet „fda“ [Ramsay et al., 2011]. Gemessen wurden tägliche Durchschnittstemperaturen innerhalb eines Jahres. Im Datensatz sind 35 Wetterstationen in Kanada enthalten. Die zugrundeliegende Datenmatrix ist $\boldsymbol{Y} \in \mathbb{R}^{35x365}$.
Die durch eine Fourier-Basis (2.15) entwickelten Originaldaten sind in Abbildung 17 abgetragen.

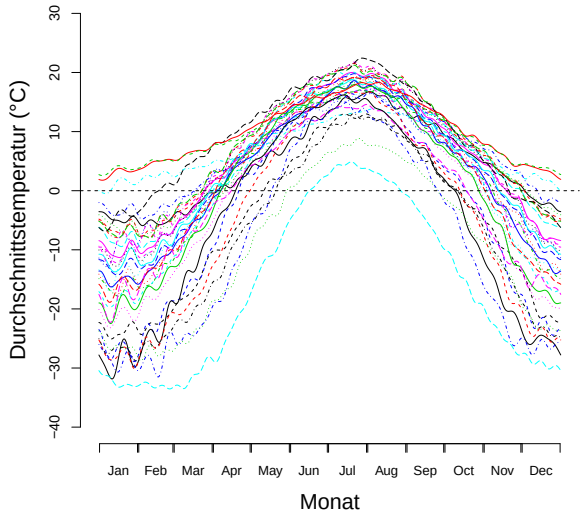


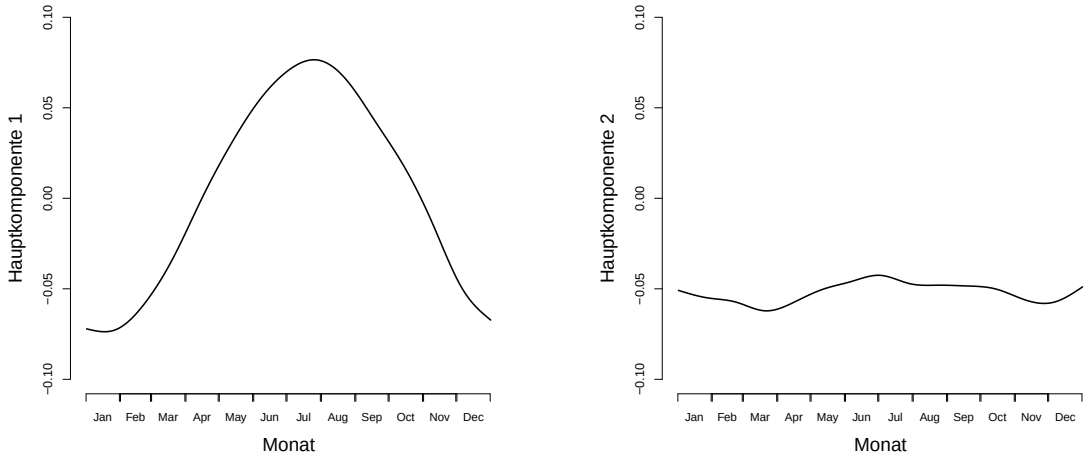
Abbildung 17: tägliche Durchschnittstemperaturen in °C eines Jahres gemessen von 35 Wetterstationen in Kanada.

Dieser Datensatz wurde mit der Funktion *noPC* analysiert. Die Ergebnisse der Selektionskriterien, die sich, legt man die Simulationsstudie zugrunde, zur Selektion funktionaler Hauptkomponenten eignen, sind in Tabelle 11 aufgelistet. Diese Ergebnisse stimmen sehr gut mit der Auswertung der Simulationsstudie überein.
Für den hier betrachteten Datensatz gilt, dass die Anzahl der Beobachtungen wesentlich kleiner ist als die Anzahl an Messzeitpunkten, also n<T. Außerdem gilt, dass $T = 365 >$ 200. Die Anzahl an Messzeitpunkten ist also größer als die größte Anzahl T, die in der Simulation betrachtet wurde.

Kriterium	nErkVar	nKaiser	nScree	nCAIC2
Anzahl an HK	2	5	5	1

Tabelle 11: Ergebnis der Selektionskriterien funktionaler Hauptkomponenten für den kanadischen Wetterdatensatz.

n<T gilt für Szenario 2, $T = 200$ gilt für Szenario 5. Dies sind zwei Fälle, bei denen das Kaiserkriterium und der Scree-Test zu viele Eigenfunktionen selektieren. Die wahre Anzahl von N=4 Eigenfunktionen wird bei Szenario 2 leicht überschätzt (Abbildung 8a) und bei Szenario 5 sogar stark überschätzt (Abbildung 8b).
Das Kriterium der erklärten Varianz selektiert für Szenario 2 und 5 jeweils drei bzw. korrekterweise vier Hauptkomponenten (vgl. Tabelle 4). In den meisten Fällen wird die wahre Anzahl an Eigenfunktionen leicht unterschätzt, in manchen Fällen aber auch korrekt geschätzt.
Das cAIC für gemischte Modelle selektiert für die relevanten Szenarien 2 und 5 jeweils drei von vier Eigenfunktionen und unterschätzt somit leicht die wahre Anzahl.
Für den Wetterdatensatz selektieren das Kaiserkriterium und der Scree-Test jeweils fünf Hauptkomponenten. Das Kriterium der erklärten Varianz und das cAIC für gemischte Modelle zwei bzw. eine Hauptkomponente (vgl. Tabelle 11).



(a) Hauptkomponente 1 (75.6%) (b) Hauptkomponente 2 (23.3%)

Abbildung 18: Die beiden ersten Hauptkomponenten der Analyse des kanadischen Wetterdatensatzes.

Geht man von der Simulationsstudie aus, so überschätzen also das Kaiserkriterium und der Scree-Test die Anzahl an Hauptkomponenten und das cAIC sollte die Anzahl an Hauptkomponenten leicht unterschätzen.
Nimmt man alle Überlegungen zusammen, so lässt sich ableiten, dass zwei Hauptkomponenten zur Analyse dieses Datensatzes ausreichen sollten.
Abbildung 18 zeigt die beiden ersten geglätteten Hauptkomponenten, die man bei der Analyse des Wetterdatensatzes durch die Funktion *LFPCAg* erhält. Insgesamt erklären die beiden Hauptkomponenten 98.9% der Gesamtvarianz. Betrachtet man die Eigenwerte λ, so berechnet sich, dass Hauptkomponente 1 75.6% und Hauptkomponente 2 23.3% der Varianz erklären. Dies bestätigt, dass zwei Hauptkompontenen für die Analyse dieses Datensatzes ausreichen.

6 Resümee

Hauptziel der Arbeit war es, anhand einer Simulationsstudie, die Güte verschiedenster Kriterien zur Wahl der Anzahl funktionaler Hauptkomponenten zu untersuchen. Zugrunde gelegt wurde ein funktionales Modell mit zusätzlichen Messfehlern (2.32). Zur Schätzung der Modellkomponenten insbesondere der Hauptkomponentenscores wurde das PACE-Verfahren angewendet.
In die Simulationstudie wurden insgesamt sieben Selektionskriterien aufgenommen. Vier der Kriterien werden klassisch für den Fall multivariater Daten angewendet:

- Erkläre einen Anteil $c\%$ der Gesamtvarianz
- Kaiser-Kriterium
- Scree-Test
- Signifikanztest nach Bartlett

Zwei der Kriterien wurden speziell für den Fall spärlicher funktionaler Daten entwickelt, vorgestellt im Paper [Yao et al., 2005]:

- CV-Kriterium
- cAIC-Kriterium

Zusätzlich wurde ein weiteres cAIC-Kriterium betrachtete, welches speziell zur Selektion zufälliger Effekte in gemischten Modellen entwickelt wurde.

Die Kernaussagen des Simulationsergebnisses sind Folgende:

1. Das CV-Kriterium für spärliche funktionale Daten lässt sich auf den hier betrachteten Fall einfacher funktionaler Daten nicht anwenden. Anhand der CV-Werte lässt sich kein Minimum bestimmen.

2. Der Signifikanztest nach Bartlett eignet sich für die Selektion funktionaler Hauptkomponenten nicht. Vor allem problematisch bei Anwendung des multivariaten Tests auf funktionale Daten ist die Konstellation der Parameter n und T.

3. Das Kaiser-Kriterium und der Scree-Test schneiden in der Studie am besten ab. Beide multivariate Kriterien lassen sich für die Selektion funktionaler Hauptkomponenten gut anwenden. Da sich beide Kriterien auf die Größe der Eigenwerte der Korrelationsmatrix beziehen, sind die Ergebnisse sehr ähnlich. In wenigen Fällen wird die korrekte Anzahl an Eigenfunktionen im Modell überschätzt, jedoch nie unterschätzt.

4. Das Kriterium der erklärten Varianz ist stark abhängig von der Anzahl N an Eigenfunktionen und der Varianzen der Gewichte ξ. Mit höherer Anzahl N und kleineren Werten λ wird die Güte der Schätzung schlechter.

5. Vergleicht man die beiden cAIC-Kriterien, so sieht man, dass das cAIC für gemischte Modelle wesentlich besser abschneidet. Der Hauptunterschied ist die Anzahl der verwendeten Freiheitsgrade. Das cAIC-Kriteium nach [Yao et al., 2005] überschätzt in den meisten Fällen die Anzahl an Eigenfunktionen deutlich. Beim cAIC für gemischte Modelle lassen sich Szenarien ausmachen, bei denen die Anzahl an Eigenfunktionen korrekt geschätzt wird, wobei sie in den meisten Fällen leicht unterschätzt wird.

6. Ebenso wie das Kriterium der erklärten Varianz, ist auch die Güte des cAIC-Kriteriums stark von der Anzahl N an Eigenfunktionen im Modell abhängig. Je größer N, desto schlechter die Selektion. Von beiden Kriterien wird außerdem die Anzahl an Eigenfunktionen tendenziell unterschätzt, jedoch nicht überschätzt.

Alles in allem sind vier der sieben betrachteten Kriterien für die Selektion der Anzahl funktionaler Hauptkomponenten geeignet. Jeweils zwei können anhand ihres Selektionsverhaltens zusammengefasst werden.
Die Anwendung der Selektionskrterien auf den kanadischen Wetterdatensatz bestätigt die Ergebnisse der Simulation.
Es sollte klar werden, dass alle Selektionskriterien Spielraum für Diskussion und Interpretation lassen und kritisch betrachtet werden sollten. Wie auch im Falle des Wetterdatensatzes ersichtlich, ist es durchaus sinnvoll mehrere Kriterien in die Entscheidung über die verwendete Anzahl an Hauptkomponenten einzubeziehen.

Literatur

[Analytics, 2011a] Analytics, R. (2011a). *doSMP: Foreach parallel adaptor for the revoIPC package*. R package version 1.0-1.

[Analytics, 2011b] Analytics, R. (2011b). *foreach: Foreach looping construct for R*. R package version 1.3.2.

[E.Weisstein, 2011] E.Weisstein (2011). "legendre polynomial.". From MathWorld–A Wolfram Web Resource. Verfügbar online auf http://mathworld.wolfram.com/LegendrePolynomial.html; besucht am 2011-08-22.

[Fahrmeir et al., 1996] Fahrmeir, Hamerle, and Tutz (1996). *Multivariate statistische Verfahren*. Berlin/New York: de Gruyter.

[Fahrmeir et al., 2007] Fahrmeir, Kneib, and Lang (2007). *Regression - Modelle, Methoden und Anwendungen*. Berlin/Heidelberg: Springer.

[Forster, 2008] Forster (2008). *Analysis 1*. Vieweg Studium, 9th edition.

[Greven and Kneib, 2010] Greven and Kneib (2010). On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*.

[Hastie et al., 2009] Hastie, Tibshirani, and Friedman (2009). *The Elements of Statistical Learning*. Springer, 2nd edition.

[Jolliffe, 2002] Jolliffe (2002). *Principal Component Analysis*. New Jork: Springer, 2nd edition.

[Novomestky, 2009] Novomestky, F. (2009). *orthopolynom: Collection of functions for orthogonal and orthonormal polynomials*. R package version 1.0-2.

[R Development Core Team, 2011] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Raiche and Magis, 2010] Raiche, G. and Magis, D. (2010). *nFactors: Parallel Analysis and Non Graphical Solutions to the Cattell Scree Test*. R package version 2.3.2.

[Ramsay and Silverman, 2005] Ramsay and Silverman (2005). *Functional Data Analysis*. New Jork: Springer, 2nd edition.

[Ramsay et al., 2011] Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2011). *fda: Functional Data Analysis*. R package version 2.2.6.

[Yao et al., 2005] Yao, Müller, and Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* ***100***, pages 577–590.

A Verfügbare Dateien

Auf der beigelegten CD befinden sich die Arbeit im PDF-Format und zwei Dateiordner:

- **R**: Beinhaltet den erzeugten R-Code (alle Dateien mit Dateiendung .r) und die gespeicherten Ergebnisse der Simulation (alle Dateien mit Dateiendung .RData).
- **Graphiken**: Beinhaltet alle für die Arbeit erstellten Graphiken (im PDF-Format).

Folgende Übersicht beinhaltet in alphabetischer Reihenfolge eine Auflistung aller verfügbaren **Source-Dateien**, der darin enthaltenen **Funktionen** und der Dateien in denen die **Ergebnisse** der Simulation gespeichert sind:

Source-Dateien

Datei	**Beschreibung**	**Funktionen**
AICdf.r	Berechnung des cAIC für gemischte Modelle	AICdfs, cloglik, cAIC
AICY.r	Berechnung des cAIC nach [Yao et al., 2005]	aicY
Auswertung4.2.r	Berechnung der Datensätze und Auswertung der FPCA für Szenario 1 und 12	
Auswertung4.3.r	Graphiken und Tabellen zur Auswertung der Simulationsergebnisse	
BartlettTest.r	Signifikanztest nach Bartlett	sigTest, bartlett
CV.r	Berechnung und Auswertung des Kriteriums durch Kreuzvalidierung	cv, noPCCV
Datensatz.r	Erstellung aller Datensätze der Simulation	daten
datSpeicherung.r	Speicherung aller Datensätze der Simulation	
Ergebnisse.r	Berechnung der Selektionskriterien für alle 16 Szenarien	berechnung

ErkVar.r	Speicherung der erklärten Varianz eines Modells	erkVar
Kaiser.r	Berechnung des Kaiser-Kriteriums	kaiser
Korrelation.r	Berechnung der Korrelationsmatrix eines Datensatzes	correlation
LFPCAg_newsigma.r	Durchführung der funktionalen Hauptkomponentenanalyse	LFPCAg
NoPC.r	Funktion zur Berechnung aller Selektionskriterien	noPC
ScreeTest.r	Durchführung des Scree-Tests	screeTest
Wetter.r	Auswertung des kanadischen Wetterdatensatzes	
ZGraphiken.r	zusätzliche Graphiken der Arbeit	

RData-Dateien

Datei	**Inhalt**	**Variablen**
Datensaetze.RData	Jeweils 100 Datensätze für jedes der 16 betrachteten Szenarien	Datensaetze
Ergebnisse.RData	Auswertungen der Selektionskriterien für jeweils 100 Datensätze der 16 betrachteten Szenarien	erg
ErgebnisseCV.RData	Berechnete CV-Werte für jeweils 100 Datensätze der Szenarien 1 und 14	ergCV
Wetter.RData	Modellfit durch LFPCAg und Auswertung der Selektionskriterien des kanadischen Wetterdatensatzes	auswertung, pca

Zusätzlich wichtige verwendete Packete sind das Packet „foreach“, welches spezielle Schleifenkonstruktionen ermöglicht [Analytics, 2011b] und das Packet „doSMP“, welches die Parallelisierung der foreach-Schleife ermöglicht [Analytics, 2011a].

B Auszüge des R-Codes und Outputs

Generierung von 100 Datensätzen für Szenario 1 der Simulation:

```
> dat <- daten(1)
> dat$szenario
[1] "n= 200 ; T= 100 ; sigma= 0.05 ; N= 4 ; lambda= h ; Eigenfunktionen= trigo"
```

dat$szenario gibt an, welche Ausprägung die Parameter im betrachteten Modell haben.

Auswertung der Selektionskriterien für den ersten Datensatz von Szenario 1 durch die Funktion *noPC*:

```
> Y <- dat$Y
> npc <- noPC(Y[,,1])
> npc
  nErkVar   nKaiser    nScree nBartlett     nAIC1     nAIC2
        3         4         4        25        10         3
```

Auswertung der Selektionskriterien für zwei Datensätze von Szenario 1 durch die Funktion *berechnung*:

```
> npc <- berechnung(1,2)
> npc
     nErkVar nKaiser nScree nBartlett nAIC1 nAIC2
[1,]       3       4      4        25    10     3
[2,]       3       4      4        24    10     3
```

Als Argumente werden der Funktion *berechnung* das betrachtete Szenario (1) und die Anzahl an Wiederholungen (2) übergeben.

Parallelisierte Auswertung aller 16 Szenarien durch die Funktion *berechnung* ergibt folgenden Output (auszugsweise):

```
> library("foreach")
> library("doSMP")
> w <- startWorkers(workerCount=8)
> registerDoSMP(w)
> erg <- foreach(sz=1:16) %dopar% berechnung(sz)
> erg
$`n= 200 ; T= 100 ; sigma= 0.05 ; N= 4 ; lambda= h ; Eigenfunktionen= trigo`
      nErkVar nKaiser nScree nBartlett nAIC1 nAIC2
 [1,]       3       4      4        25    10     3
 [2,]       3       4      4        24    10     3
 [3,]       3       4      4        21    10     3

  ⋮⋮⋮

$`n= 200 ; T= 100 ; sigma= 0.1 ; N= 4 ; lambda= h ; Eigenfunktionen= trigo`
      nErkVar nKaiser nScree nBartlett nAIC1 nAIC2
 [1,]       3      11      9        14     9     3
 [2,]       3       9      9        13    10     2
 [3,]       3      10      8        10     9     3

  ⋮⋮⋮

$`n= 200 ; T= 100 ; sigma= 0.05 ; N= 8 ; lambda= l ; Eigenfunktionen= trigo`
      nErkVar nKaiser nScree nBartlett nAIC1 nAIC2
 [1,]       7       8      8        12     7     7
 [2,]       7       8      8         8     7     7
 [3,]       8       8      8         8     4     4

  ⋮⋮⋮
```

Eidesstaatliche Erklärung

Hiermit versichere ich, Moritz Berger, die vorliegende Bachelorarbeit selbstständig und lediglich unter Benutzung der angegebenen Quellen und Hilfsmittel verfasst zu haben.

München, den 23.08.2011

Moritz Berger