UNIVERZITA PALACKÉHO V OLOMOUCI
PRÍRODOVEDECKÁ FAKULTA

DIPLOMOVÁ PRÁCA

Funkcionálny regresný model s kompozičnou vysvetľujúcou premennou

Katedra matematickej analýzy a aplikácií matematiky
Vedúci diplomovej práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Paulína Jašková**
Študijný program: B1103 Aplikovaná matematika
Študijný obor: Aplikácia matematiky v ekonómií
Forma štúdia: prezenčná
Rok odovzdania: 2020

BIBLIOGRAFICKÁ IDENTIFIKÁCIA

Autor: Paulína Jašková

Názov práce: Funkcionálny regresný model s kompozičnou vysvetľujúcou premennou

Typ práce: Diplomová práca

Pracovisko: Katedra matematickej analýzy a aplikácií matematiky

Vedúci práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2020

Abstrakt: Konštrukcia regresného modelu pre funkcionálne údaje patrí medzi relatívne dobre preskúmané oblasti funkcionálnej analýzy údajov. V diplomovej práci sa venujeme tvorbe regresného modelu funkcionálnych údajov s kompozičnou vysvetľujúcou premennou, pre ktorú sa príslušná metodika zatiaľ vytvára. Snažíme sa nájsť vzťah medzi reálnou vysvetľovanou premennou a hustotou rozdelenia pravdepodobnosti, ktorá popisuje relatívnu štruktúru funkcionálneho štatistického znaku. Pre účely regresie je najskôr potrebné odhadnúť hustoty ako funkciu a za tým účelom sa zaoberáme možnosťou jej interpolácie B-splajnom. Sústredíme sa na odhad funcionálnych regresných parametrov s využitím simpliciálnej funkcionálnej metódy hlavných komponentov. V praktickej časti túto regresiu aplikujeme na modelovanie závislosti výskytu rakoviny od štruktúry populácie s využitím údajov z vybraných krajín. Na výpočtové účely je použitý software R.

Kľúčové slová: štatistické metódy, funkcionálna analýza, regresná analýza

Počet strán: 85

Počet príloh: 2

Jazyk: slovenský

BIBLIOGRAPHICAL IDENTIFICATION

Author: Paulína Jašková

Title: Functional regression model with compositional predictor

Type of thesis: Master's

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2020

Abstract: The construction of a regression model for functional data is one of the relatively well-researched areas of functional data analysis. However, in this Master thesis we deal with a regression model of functional data with a compositional explanatory variable, for which the respective methodology is still under development. We try to find the relationship between the real explanatory variable and the probability density function, which describes the relative structure of a functional random variable. For the regression purpose it is first necessary to estimate the density as a function and, consequently, we deal with the possibility of its interpolation by a B-spline. We focus on the estimation of functional regression parameters using the simplicial functional principal component analysis. In the practical part, we apply this regression to model the dependence of cancer on the structure of the population using data from selected countries. R software is used for computational purposes.

Key words: statistical methods, functional analysis, regression analysis

Number of pages: 85

Number of appendices: 2

Language: Slovak

Prehlásenie

Prehlasujem, že som diplomovú prácu spracovala samostatne, pod vedením doc. RNDr. Karla Hrona, Ph.D a všetky použité zdroje som uviedla v zozname literatúry.

V Olomouci dňa ....

podpis

Obsah

Poďakovanie

Rada by som poďakovala doc. RNDr. Karlovi Hronovi, Ph.D. za všetok čas, rady a pripomienky, ktorými prispel k tvorbe tejto práce. Zároveň by som chcela poďakovať Mgr. Renate Talskej za pomoc s pochopením témy a spracovaním údajov.

Úvod

Cieľom diplomovej práce je navrhnutie vhodného modelu pre modelovanie vzťahu medzi reálnou vysvetľovanou premennou a hustotou rozdelenia pravdepodobnosti, popisujúcej relatívnu štruktúru (funkcionálneho) štatistického znaku a jeho implementácia v štatistickom softvéri R. Funkcionálne údaje a celkovo funkcionálna analýza údajov sú novšie pojmy pri spracovaní štatistických údajov. Funkcionálne údaje predstavujú celý súbor meraní v čase alebo v rámci nejakej inej spojitej premennej. Súčasťou diplomovej práce preto bude aj zoznámenie sa so základmi funkcionálnej analýzy údajov ako takej.

Práca je delená do šiestich kapitol, ktoré obsahujú niekoľko podkapitol. V prvej kapitole je krátky úvod do funkcionálnej analýzy údajov a predstavený Hilbertov priestor, ktorý budeme ďalej v práci používať. Zároveň sú v tejto kapitole uvedené aj niektoré štatistiky pre tento typ údajov. V druhej kapitole sa budeme zaoberať aproximáciou funkcionálnych údajov. Predstavíme B-splajnovú reprezentáciu a vyhladzovanie pomocou nej, ktorú budeme používať v praktickej časti. Nakoľko v práci nebudeme pracovať s funkcionálnymi údajmi vo všeobecnosti, ale budeme pracovať s hustotami rozdelenia pravdepodobností, tak si v tretej kapitole uvedieme niečo o tomto type pozorovania. Zároveň definujeme Bayesov priestor, ktorý predstavuje iný pohľad na hustoty funkcií ako na funkcionálne údaje nesúce len relatívnu informáciu. Tento priestor je veľmi dôležitý, nakoľko pre takýto typ údajov je podstatné rešpektovanie relatívneho charakteru hustôt. V štvrtej kapitole je uvedená metóda hlavných komponentov pre funkcionálne údaje a špeciálne aj pre husoty. Piata kapitola predstavuje hlavnú časť celej diplomovej práce, regresnú analýzu. Predstavíme si tu rôzne modely pre rozdielne

prediktory a nakoniec regresiu pre hustoty a výpočtové metódy. Posledná časť zahŕňa implementáciu všetkych teoreticky objasnených metód na reálnom príklade. V praktickej aplikačnej časti skúmame, ako veková štruktúra populácie (vyjadrená pomocou hustoty) ovplyvňuje početnosť výskytu daného druhu rakoviny v súbore niekoľkých skúmaných štátov. Budeme modelovať vzťah výskytu rakoviny prostaty na vekovej štruktúre danej krajiny. Spracovanie údajov v celej práci a prevedené jednotlivé analýzy sú uskutočnené pomocou voľne dostupného štatistického softvéru *R*. V celej práci sme použili knižnice *robCompositions*, *compositions* a fda [19].

Kapitola 1

Úvod do funkcionálnej analýzy

Nové typy údajov si vyžadujú nové spôsoby ich analýzy. Hoci funkcionálne údaje ako pojem je relatívne nový, [12] nové spôsoby ich spracovania sú výsledkom rastúcej sofistikácie týchto údajov v behaviorálnych aj prírodných vedách. Technológia získavania údajov sa za posledné roky zdokonalila a umožňuje získavať údaje husto vzorkované v čase, priestore a iných súvislostiach. Tieto údaje zvyčajne odrážajú vplyv určitých hladkých funkcií, o ktorých sa predpokladá, že tvoria pozorovania. Na takéto údaje sa môžu aplikovať klasické viacrozmerné štatistické metódy, nemožno z nich však vyťažiť ďalšie informácie, ktoré vyplývajú z priebehu spomínaných funkcií. Metódy funkcionálnej analýzy údajov (FDA) umožňujú získať ďalšie informácie obsiahnuté vo funkciách a ich deriváciách, ktoré zvyčajne nie sú dostupné tradičnými metódami. Funkčný údaj nepredstavuje samostatné pozorovanie, ale skôr súbor meraní v čase a priestore. Veľa výstupov z oblasti psychológie a behaviorálnych vied je tvorený množinou funkcionálnych údajov. Napríklad psychobiologické funkcie, ktoré sa časom menia, ako sú hladiny hormónov v krvi, emócie alebo počet a intenzita depresívnych epizód, sú len niektoré z mnohých príkladov. Ľudská komunikácia, vrátane hovoreného a posunkového jazyka, hudobné vystúpenia a vnímania, sú ďalšie domény, ktoré sa v čase menia. V prípade, že máme znázornený zvukový záznam, chceme kvantifikovať relatívne úlohy zvukových a obrazových informácií, pri sprostredkovaní emocionálnych a štrukturálnych aspektov hudby. Obdobne môžeme za funkcionálne údaje považovať napríklad rastové krivky detí, priebehy teploty

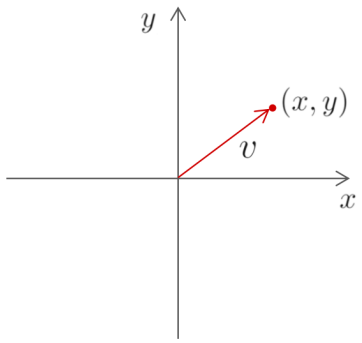
v daných oblastiach, alebo čo nás bude následne zaujímať v kontexte tejto práce, vekové rozdelenie obyvateľstva vo vybraných štátoch.

V tejto kapitole si uvedieme stručný úvod do problematiky FDA a zároveň si predstavíme Hilbertov priestor, ktorý bude podstatný pre ďalšiu analýzu.

1.1. Hilbertové priestory

Pojem Hilbertov priestor zovšeobecňuje pojem euklidovského priestoru na priestory akejkoľvek, dokonca nekonečnej dimenzie. Hilbertov priestor nám umožňuje merať uhly a vzdialenosti medzi prvkami z lineárnych priestorov nekonečnej dimenzie, ako napríklad v priestoroch nekonečných funkcií. V tejto kapitole sme čerpali z [3], [16] a [17].

Uvedieme niektoré vlastnosti euklidovského priestoru, pre jednoduchosť zostaneme v rovine $\mathbf{R}^2$:



Obr. 1.1: Euklidovský priestor $\mathbf{R}^2$

Operácie $(+, \cdot)$ pre vektory $\mathbf{v}_1 = (x_1, y_1)$ a $\mathbf{v}_2 = (x_2, y_2)$:

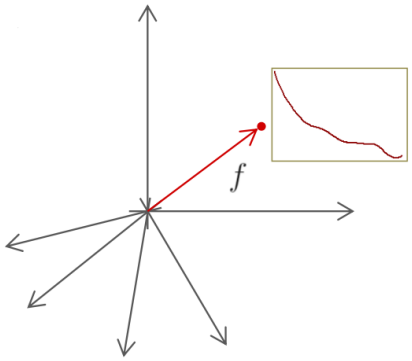
- Súčet dvoch vektorov: $\mathbf{v}_1 + \mathbf{v}_2 = (x_1 + x_2, y_1 + y_2)$.
- Vynásobenie vektoru konštantou c: $c \cdot \mathbf{v}_1 = (c \cdot x_1, c \cdot y_1)$.

Skalárny súčin $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = (x_1 \cdot x_2) + (y_1 \cdot y_2)$:

- Norma vektoru: $||\mathbf{v}_1|| = \sqrt{(x_1^2 + y_1^2)}$.
- Vzdialenosť $d(\mathbf{v}_1 - \mathbf{v}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

- Uhol medzi $\mathbf{v}_1$ a $\mathbf{v}_2$: $\vartheta(\mathbf{v}_1, \mathbf{v}_2) = \arccos \frac{\langle v_1, v_2 \rangle}{||v_1||||v_2||}$.

Euklidovský priestor sa využíva vo viacrozmernej štatistike, zatiaľ čo pri funkcionálnej analýze dát sa využíva Hilbertov priestor. Hilbertov priestor budeme využívať preto, lebo funkcionalne údaje predstavujú body priestoru funkcií, čo je znázornené na obrázku.



Obr. 1.1: Hilbertov priestor

V Hilbertovom priestore vieme využiť veľa techník z lineárnej algebry s využitím skalárneho súčinu a normy. Pre následné použitie si priblížime pojem **lineárny priestor so skalárnym súčinom**.

Definícia 1. Nech H je neprázdna množina. Potom sa algebraická štruktúra $(L, +, \cdot)$ nazýva **lineárny priestor nad R**, ak sú na L definované:

1. operácia sčítanie $(+)$, ktorá priraďuje každému $x, y \in L$ prvok $(x + y) \in L$,
2. operácia násobenie $(\cdot)$, ktorá priraďuje každému $x \in L$ a každému $\alpha \in \mathbf{R}$ prvok $(\alpha \cdot x) \in L$.

Pritom pre každé $x, y, z \in L$ a pre každé $\alpha, \beta \in \mathbf{R}$ platí:

- $x + y = y + x$,
- $x + (y + z) = (x + y) + z$,
- existuje $0 \in L$, že pre každé $x \in L$ platí $x + 0 = x$,
- pre každé $x \in L$ existuje také $(-x) \in L$, že platí $x + (-x) = 0$,

- $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$,
- $(\alpha \cdot \beta) \cdot x = \alpha \cdot (\beta \cdot x)$,
- $1 \cdot x = x$.

Majme lineárny priestor H. Skalárny súčin na tomto lineárnom priestore H je symetrická, pozitívne definitná forma $\langle \cdot, \cdot \rangle : H \times H \to \mathbf{R}$, ktorá vyhovuje nasledujúcim vlastnostiam:

1. $\langle \lambda x + y, z \rangle = \lambda \langle x, z \rangle + \langle y, z \rangle \quad \forall \lambda \in \mathbf{R}, \forall x, y, z \in H$,
2. $\langle x, y \rangle = \langle y, x \rangle \quad \forall x, y \in H$,
3. $\langle x, x \rangle \geq 0 \quad \forall x \in H$,
4. $\langle x, x \rangle = 0 \iff x = 0$.

Norma definovaná pomocou skalárneho súčinu $||x|| = \sqrt{\langle x, x \rangle}$, je norma na tomto priestore, nazývame ju aj norma indukovaná skalárnym súčinom. Vďaka tomu vieme zadefinovať ortogonalitu na priestore H. Dva vektory sú v H ortogonálne, ak platí $\langle x, y \rangle = 0$. Skalárny súčin v tomto prípade indukuje normu a metriku, obdobne ako v prípade euklidovského priestoru, ktorý je najznámejším príkladom lineárneho priestoru so skalárnym súčinom. Nasledujúca definícia je čerpaná z [17].

Definícia 2. Normovaným lineárnym priestorom rozumieme každý lineárny priestor X, ktorý je vybavený normou, čo je funkcionál $|| \cdot || : X \to \mathbf{R}$, splňujúci pre každé $x, y \in X$ a $\alpha \in \mathbf{R}$ podmienky:

- $||x|| \geq 0$, pričom $||x|| = 0$ práve keď $x = 0$,
- $||\lambda \cdot x|| = |\lambda| \cdot ||x||$,
- $||x + y|| \leq ||x|| + ||y||$.

Lineárny priestor so skalárnym súčinom nazývame **úplným**, ak príslušný normovaný lineárny priestor je úplný. Úplný lineárny priestor so skalárnym súčinom nazývame **Hilbertov priestor** H. Nasledujúca definícia je čerpaná z [17].

Definícia 3. Nech H je Hilbertov priestor. O množine $\{x_i, i \in I\}$ prvkov z H hovoríme, že je **úplná**, ak jediný prvok $z \in H$, ktorý je kolmý ku všetkým prvkom uvažovanej množiny, je nulový prvok.

Množinu nenulových prvkov $\{x_i, i \in I\}$ z Hilbertovho priestoru nazveme **ortogonálnu**, ak pre prvky z tejto množiny platí:

$$\forall i \neq j : \langle x_i, x_j \rangle = 0, \qquad i, j \in I.$$

Ak naviac platí, že norma pre každé i je rovná 1, potom túto množinu môžeme nazvať **ortonormálnu**. Samozrejme každú ortogonálnu množinu vieme veľmi jednoducho previesť na ortonormálnu a to tak, že každý prvok podelíme príslušnou normou.

Ortonormálnu bázu v Hilbertovom priestore rozumieme množinu $\{e_i, i \in I\}$ prvkov z H, ak:

1. všetky prvky bázy sú navzájom kolmé a majú jednotkovú dĺžku, t.j. sú ortonormálne,

2. lineárny obal bázy je hustý v H, to znamená že uzáver lineárneho obalu je rovný H.

Separabilným Hilbertovým priestorom rozumieme Hilbertov priestor, na ktorom existuje najviac spočítateľná množina prvkov e_i taká, že ľubovolný prvok x sa dá vyjadriť ako lineárna kombinácia $x = \sum_{i \in I} \alpha_i e_i$.

Vlastnosti:

- Na Hilbertovom priestore dokážeme zadefinovať pojem ortogonálna projekcia a najlepšia aproximácia.

- V každom Hilbertovom priestor H existuje ortonormálna báza $\{u_n\}_{n \in \mathbf{N}}$.

- Ak je H **separabilný** Hilbertov priestor, $\{u_n\}_{n\in\mathbf{N}}$ je ortonormálna báza a $x \in H$, potom

$$x = \sum_{n=1}^{\infty} \langle x, u_n \rangle u_n.$$

Nazývame to aj bázové rozšírenie.

1.1.1. Príklady Hilbertových priestorov

- **Priestory $\mathbf{R}^n$**: Priestory všetkých n-tic, na ktorých je definovaný skalárny súčin daný vzťahom

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i \cdot y_i$$

tvorí Hilbertov priestor konečnej dimenzie.

- **Priestor l^2**: Priestory všetkých nekonečných postupností reálnych čísiel $x = \{x_i\}_{i=1}^{\infty}$, ich súčet druhých mocnín absolútnych hodnôt je konečný, t.j.

$$\sum_{i=1}^{\infty} |x_i|^2 < \infty,$$

spolu so skalárnym súčinom

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i \cdot y_i$$

tvorí Hilbertov priestor nekonečnej dimenzie.

- **Priestor $L^2[a, b]$**: Uvažujeme priestor Lebesgueovsky merateľných funkcií z $[a, b] \to \mathbf{R}$, ktoré sú integrovateľné s kvadrátom na $[a, b]$, t.j. integrál

$$\int_a^b |f(t)|^2 dt$$

existuje a je konečný. Potom na tomto priestore funkcií je definovaný skalárny súčin ako

$$\langle f, g \rangle_2 = \int_a^b f(t)g(t)dt,$$

a tvorí Hilbertov priestor nekonečnej dimenzie.

Môžeme povedať, že priestor $L^2[a,b]$ je priestor integrovateľných funkcií, na ktorom platia nasledujúce vlastnosti:

- Operácia $(+,\cdot)$:
  - Súčet: $(f_1 + f_2)(t) = f_1(t) + f_2(t)$.
  - Vynásobenie konštantou: $(c \cdot f)(t) = c \cdot f(t)$.
- Skalárny súčin $\langle f_1, f_2 \rangle = \int_a^b (f_1(t) \cdot f_2(t))dt$:
  - Norma: $||f||^2 = \int_a^b (f(t))^2 dt$.
  - Vzdialenosť: $||f_1 - f_2||^2 = \int_a^b (f_1(t) - f_2(t))^2 dt$.
  - Uhol medzi funkciami f_1 a f_2: $\vartheta(f_1, f_2) = \arccos \frac{\langle f_1, f_2 \rangle}{||f_1||||f_2||}$.

Tieto vlastnosti sú veľmi dôležité, pretože napríklad daný skalárny súčin nám umožňuje na Hilbertových priestoroch určiť ortogonalitu (kolmosť) jej prvkov.

1.2. Funkcionálne údaje

Vývoj telesnej výšky človeka, časový priebeh úhrnu zrážok, vývoj cien na trhu, to je len niekoľko údajov, ktoré berieme ako funkcionálne údaje. Funkcionálne údaje sú teda entity, ktoré môžu byť popísané pomocou funkcie. Funkcionálna analýza dát spracováva informácie o údajoch, ktoré majú funkcionálny charakter. Ciele funkcionálnej analýzy sú rovnaké ako v každom odvetví štatistiky. Obsahuje nasledujúce ciele:

- Reprezentovať údaje tak, aby nám pomohli v ďalšej analýze.
- Zobraziť údaje tak, aby boli zvýraznené rozdelenie vlastností.
- Sledovať variabilitu v údajoch.
- Porovnať dva alebo viac súborov údajov s ohľadom na určité typy variácií, kde dva súbory údajov môžu obsahovať rozdielne súbory replikácií jednej funkcie alebo rozdielne funkcie pre spoločný súbor replikácií.

V nasledujúcej kapitole si uvedieme niekoľko štatistík pre funkcionálne údaje. Následne si definujeme reprezentáciu pomocou B-splajnov, ktorá je veľmi vhodná pre ďalšie spracovanie údajov.

1.2.1. Súhrnné štatistiky pre funkcionálne údaje

Klasické súhrnné štatistiky pre jednorozmerné údaje vieme odvodiť aj pre funkcionálne. Jedny zo základných štatistík sú miery polohy, ktoré samozrejme existujú aj pre funkcionálne údaje. Miery polohy pre funkcionálne údaje nám poskytujú informáciu o tom, kde sa dané údaje v jednotlivých časových okamihoch koncentrujú. Najčastejšou počítanou charakteristikou polohy je **funkcionálny aritmetický priemer**. Jeho hodnoty sú silne ovplyvňované odľahlými hodnotami, to znamená hodnotami, ktoré sa v čase t líšia od ostatných. Výberový funkcionálny aritmetický priemer pre náhodný výber funkcií $x_1, \ldots, x_N$ rozumieme funkciu premennej t z nejakého intervalu $[a, b]$ z $L^2[a, b]$ a je daný nasledujúcim

vzťahom

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t).$$

Na určenie miery rozptýlenia hodnôt okolo zistenej miery polohy nám slúžia charakteristiky ako rozptyl a smerodajná odchýlka. Pri funkcionálnom náhodnom výbere z $L^2[a,b]$, výberový funkcionálny rozptyl definujeme ako funkciu $s^2(t)$ s hodnotami

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

a výberovú funkcionálnu smerodajnú odchýlku ako odmocninu z výberového rozptylu

$$s(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2}.$$

Ďalšou, veľmi dôležitou charakteristikou je kovariančná funkcia. Táto funkcia umožňuje zistiť závislosť medzi nameranými hodnotami. Kovariančná funkcia zahrňuje závislosť jednotlivých hodnôt naprieč rôznym časovým okamihom, je teda vhodná k určeniu kovariancie medzi nameranými hodnotami. **Výberová kovariančná funkcia** medzi funkčnými hodnotami $x_i(t_1)$ a $x_i(t_2)$ v čase t_1 a t_2 je daná vzťahom

$$v(t_1, t_2) = \frac{1}{N}\sum_{i=1}^{n} \{x_i(t_1) - \bar{x}(t_1)\}\{x_i(t_2) - \bar{x}(t_2)\}.$$

Vo funkcionálnom prípade vieme vypočítať aj korelačnú funkciu, ktorá je lepšie interpretovateľnejšia. Nadobúda hodnoty len z intervalu $[-1, 1]$. V prípade, ak máme hodnotu 1 alebo -1, tak máme silnú lineárnu závislosť medzi nameranými hodnotami v čase t_1 a t_2. Ak nadobúda hodnotu 0, žiadna lineárna závislosť medzi nameranými hodnotmai nie je. Keďže v rámci FDA je súbor pozorovaní získaný meraním jednej veličiny v rovnakých jednotkách, korelačná funkcia tu stráca na významnosti. To nebude ale platiť v prípade, kedy budeme chcieť kvantifikovať závislosť a silu lineárneho vzťahu medzi dvomi rôznymi (funkcionálnymi) veličinami, meranými v rôznych jednotkách. **Korelačná funkcia** je daná nasledujúcim

vzťahom

$$r(t_1, t_2) = \frac{v(t_1, t_2)}{\sqrt{s^2(t_1) \cdot s^2(t_2)}} = \frac{v(t_1, t_2)}{s(t_1) \cdot s(t_2)}.$$

Ak máme dvojice pozorovaných funkcií (x_i, y_i), spôsob, akým môžeme vyjadriť ich závislosť, je krížová kovariančná funkcia alebo krížová korelačná funkcia. Majme dvojrozmerný funkcionálny náhodný výber z $L^2[a, b] \times L^2[a, b]$. Lineárny vzťah napozorovaných dvojíc funkcií $(x_1(t), y_1(t)), \ldots, (x_n(t), y_n(t))$, v čase t_1 a t_2, možno určiť pomocou **krížovej kovariančnej funkcie** nasledovne

$$v(t_1, t_2) = \frac{1}{N} \sum_{i=1}^{n} [x_i(t_1) - \bar{x}_i(t_1)][y_i(t_2) - \bar{y}_i(t_2)].$$

Podobne ich lineárny vzťah vieme určiť aj pomocou **krížovej korelačnej funkcie**

$$r(t_1, t_2) = \frac{v(t_1, t_2)}{\sqrt{s_x^2(t_1) \cdot s_y^2(t_2)}} = \frac{v(t_1, t_2)}{s_x(t_1) \cdot s_y(t_2)}.$$

Interpretácia krížovej korelačnej funkcie je v podstate rovnaká, ako pri obyčajnej korelačnej funkcii.

Kapitola 2

Aproximácia funkcií

V tejto časti budeme pracovať s priestorom splajnov a ozrejmíme si pojem B-splajn, ktorý v ďalšej časti budeme používať na reprezentáciu funkcionálnych údajov. Splajn je špeciálna funkcia, ktorá je po častiach polynómom. Za zakladateľa teórie splajnov sa považuje rumunský matematik Isaac Jacob Schoenberg ($\star 21.4.1903 - \dagger 21.2.1990$), ktorý pravdepodobne použil ako prvý slovo *splajn* v spojení s polynomiálnou aproximáciou v roku 1946. V tejto práci sa budeme zaoberať polynomickými splajnami, konkrétne ich B-splajnovou reprezentáciou. Polynomický splajn je možné skonštruovať viacerými spôsobmi. V našom prípade k zostrojeniu splajnu použijeme bázové funkcie, tzv. B-splajny.

Často sa stretávame s dosť komplikovaným teoretickým explicitným vyjadrením funkcie. Nakoľko sú známe jej hodnoty len v niektorých diskrétnych bodoch, je následne zložité používanie danej funkcie. Z tohto dôvodu je vhodnejšie nájsť jej jednoduchšiu verziu a touto verziou nahradiť pôvodný, komplikovanejší tvar funkcie. Je známych niekoľko metód aproximácie danej funkcie. Dôležité sú vlastnosti, ktoré vyžadujeme, aby funkcia spĺňala. Jedna z metód je interpolácia, kedy požadujeme, aby vysvetlená funkcia priamo prechádzala danými diskrétnymi bodmi. Interpolácia polynómom je vhodná len v prípade menšieho počtu vstupných údajov. V prípade rozsiahlejšieho súboru údajov môže nastať situácia, že funkcia bude nevhodne oscilovať. V takomto prípade sa odporúča využiť tzv. splajnovú interpoláciu, teda interpoláciu splajnom. Postup je taký, že lepíme dokopy aspoň dva polynómy, pričom od výsledného produktu vyžadu-

jeme dosiahnutie určitej hladkosti. Spomínaný postup predstavuje čiastkový cieľ tejto práce s funkcionálnymi údajmi.

S funkcionálnymi údajmi sa stretávame často, ukážeme si, ako môžeme tento typ údajov reprezentovať. Typické pozorovania funkcionálných dát sú diskrétne, pričom je v nich zahrnutý šum.

Funkcionálnymi údajmi rozumieme pre danú funkciu n takých dvojíc (f_i, t_i), medzi ktorými je nejaký funkcionálny vzťah. Uvažujeme **model:**

$$f_i = x(t_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

kde pripúšťame aj chyby v modeli, ktoré sú vyjadrené pomocou hodnoty ϵ_i a funkcia $x(t_i)$ vyjadruje jednotlivé známe záznamy. Predpokladáme, že funkcia $x(t)$ je hladká v zmysle existencie spojitých derivácií až do určitého rádu. Požadujeme, aby táto funkcia bola hladká, aby nemohli nastať viaceré skoky za sebou.

V závislosti od predchádzajúcich znalostí o chybe merania môžeme vykonávať dva procesy získavania funkcií z nameraných hodnôt. **Interpolácia** je proces získavania funkcií z nameraných údajov v prípade, že naše pozorovania neobsahujú chybu. **Aproximácia** je proces získavania funkcií v prípade, že naše pozorovania obsahujú chybu, medzi tieto procesy patrí aj **vyhladzovanie**. Vyhladzovanie je proces, pri ktorom získame funkčnú formu hladkejšiu ako skutočné pozorovania. Tento spôsob sa používa napríklad pri prevode diskrétnych údajov na funkcie.

Ak sa zameriavame na interpoláciu alebo vyhladzovanie diskrétnych údajov, tak postupujeme nasledovne. Najskôr si zvolíme priestor funkcií, ktorého dimenzia je známa, následne si zvolíme bázové funkcie (funkcie ktoré tvoria bázu daného priestoru). Potom každú funkciu z tohoto priestoru môžeme vyjadriť ako lineárnu kombináciu bázových funkcií.

Systém bázových funkcií je teda množina známych funkcií ϕ_k, ktoré sú na sebe matematicky nezávislé (napr. v zmysle priestoru $L^2[a, b]$) a ktoré majú vlastnosť, že môžeme ľubovolnú funkciu vyjadriť pomocou lineárnej kombinácie dostatočne veľkého počtu K týchto bázových funkcií. Pokiaľ budeme chcieť napríklad hľadať polynóm stupňa k, tak si môžeme vziať bázu priestoru vyjadrenú pomocou

bázových funkcií v tvare monomiálov,

$$1, t, t^2, t^3, \ldots, t^k.$$

Ďalší, veľmi známy systém bázových funkcií, je Fourierový systém, ktorých báza je periodická. Pre nás najzaujímavejším systémom funkcií sú bázové splajnové systémy, **B-splajny**, ktoré vieme použiť pre neperiodické aj periodické údaje. B-splajny umožňujú zachytiť aj zložité priebehy v rozsiahlých dátových súboroch.

Funkciu $x(t)$ vyjadríme ako lineárnu kombináciu bázových funkcií,

$$x(t) = \sum_{k=1}^{K} c_k \phi_k,$$

kde ϕ_k sú známe bázové funkcie a $(c_1, \ldots, c_K)^T$ je vektor neznámych koeficientov, ktoré určujú jednoznačne našu hľadanú fukciu $x(t)$. Písmenom K označujeme dimenziu priestoru (rozmer), ktorá je definovaná ako počet prvkov jeho ľubovolnej bázy a značíme ju ako dim.

Podrobnejšie si teraz vysvetlíme B-splajnovú reprezentáciu a následne aj vyhladzovanie pomocou týchto splajnov. V tejto kapitole sme čerpali z [9], [8], [1] a [14].

2.1. B-splajnová reprezentácia

Definácia čerpaná z [8].

Definícia 4. Majme konečný interval $[a, b]$ a na ňom sieť $g+2$ bodov (tzv. uzlov)

$$\Delta\lambda : a = \lambda_0 < \lambda_1 < \ldots < \lambda_g < \lambda_{g+1} = b.$$

Polynomickým splajnom stupňa k s defektom d (k a d sú celé nezáporné čísla) na sieti uzlov $\Delta\lambda$ budeme rozumieť funkciu $s_k(t)$ takú, že:

1. $s_k(t)$ je na každom intervale $[\lambda_i, \lambda_{i+1}], i = 0, \ldots, g$, polynómom stupňa najviac k.

2. $s_k(t) \in C^{k-1}[a,b]$, t.j. funkcia $s_k(t)$ má spojité derivácie až do radu $k-1$ na $[a,b]$.

Symbolom $S_k^{\Delta\lambda}[a,b]$ budeme označovať lineárny priestor polynomických splajnov stupňa $k > 0$, definovaných na konečnom intervale $[a,b]$ so sieťou uzlov $\Delta\lambda$.

Uvažujme splajn $s_k(t)$ z Definície 4. Vieme ako vyzerá sieť uzlov, na ktorej je definovaný tento splajn a poznáme jeho stupeň k. Dimenziu priestoru $S_k^{\Delta\lambda}[a,b]$ určíme pomocou vzťahu

$$dim S_k^{\Delta\lambda}[a,b] = (k+1)(g+1) - gk = g + k + 1,$$

kde $(k+1)(g+1)$ predstavuje počet neznámych koeficientov polynómu, t.j. máme $k+1$ koeficientov na $g+1$ intervaloch, ktoré určujú splajn $s_k(t)$. Druhá časť výrazu gk vyjadruje, že v g vnútorných uzloch splajnu má byť splnených k podmienok spojitosti, čím dôjde k eliminácií niektorých neznámych.

2.1.1. Definícia a základné vlastnosti B-splajnov

V tejto práci budeme používať B-splajnovú reprezentáciu, teda polynomický splajn budeme hľadať pomocou tzv. **B-splajnov**, ktorý umožňuje nasledujúci spôsob zostrojenia polynomického splajnu.

Definícia je prevzaná z [1].

Definícia 5. Nech $\Delta\lambda$ je neklesajúca postupnosť uzlov splajnov. Ku každému uzlu λ_i, ktorý má na sieti vpravo $k \in \mathbf{N}$ susedných uzlov takých, že $\lambda_i < \lambda_{i+k}$, definujeme B-splajn $(k+1)$-ého rádu vzťahom

$$B_i^{k+1}(t) = (y_{i+k} - y_i)(t-x)_+^{k+1}[y_i, y_{i+1}, \ldots, y_{i+k}], \quad x \in \mathbf{R}.$$

Stupňom tohoto B-splajnu rozumieme číslo k.

Funkcia $(t-x)_+^k$ je tiež známa ako kladná časť mocniny. Ide o funkciu argumentu t, v závislosti na parametri x. Definujeme ju predpisom

$$(t-x)_+^{k-1} = \begin{cases} (t-x)^{k-1} & \text{pre} \quad t > x \\ 0 & \text{pre} \quad t \leq x \end{cases}$$

Uvažujme sieť uzlov $\Delta\lambda$ z Definície 4 . Našou úlohou je, pre nájdenie polynomických splajnov, zostrojiť toľko lineárne nezávislých B-splajnov, aká je dimenzia príslušného priestoru. Bázu tvorí $g+k+1$ B-splajnov B_i, ktoré ale na sieti uzlov $\Delta\lambda$ nie sme schopní zostrojiť tak, aby tvorili celú bázu daného priestoru funkcií $S_k^{\Delta\lambda}[a,b]$. Z tohto dôvodu musíme rozšíriť sieť uzlov o tzv. pridané uzly. Týmto spôsobom dostaneme novú sieť uzlov, ktorá predstavuje neklesajúcu postupnosť.

Môžeme predpokladať že pridané uzly sú na hranici, teda

$$\lambda_{-k} = \ldots = \lambda_{-1} = \lambda_0 = a, \quad b = \lambda_{g+1} = \lambda_{g+2} = \ldots = \lambda_{g+k+1}.$$

Pridané uzly splňujú podmienku

$$\lambda_{-k} \leq \lambda_{-k+1} \leq \ldots \leq \lambda_0, \quad \lambda_{g+1} \leq \lambda_{g+2} \leq \ldots \leq \lambda_{g+k+1}.$$

Takto vytvorenú sieť uzlov nazývame **rozšírená sieť uzlov** a uzly

$$\lambda_{-k}, \ldots, \lambda_0, \qquad \lambda_{g+1}, \ldots, \lambda_{g+k+1}$$

nazývame **rozširujúce uzly.**

Teraz si uvedieme základné vlastnosti bázových splajnov, ktoré sú prevzaté z [8]:

1. **Nosič:** Nosičom $B_i^{k+1}(t)$ je interval $[\lambda_i, \lambda_{i+k+1}]$ t.j. $B_i^{k+1}(t) = 0$ pre $t \notin [\lambda_i, \lambda_{i+k+1}]$.

2. **Nezápornosť:** Pre B-splajn $B_i^{k+1}(t)$ platí

   $$B_i^{k+1}(t) \geq 0,$$

   pre $t \in [\lambda_i, \lambda_{i+k+1}]$. Rovnosť nastáva vtedy, keď $\lambda_i = \lambda_{i+k+1}$.

3. Pre $t \in [\lambda_j, \lambda_{j+1}]$, kde $j = k-1, \ldots, n+1-k$, platí

   $$\sum_i B_i^{k+1}(t) = 1.$$

   B-splajny tvoria tzv. jednotkový rozklad.

4. B-splajn $B_i^{k+1}(t)$ je na každom intervale $[\lambda_i, \lambda_{i+1}]$ pre $i = 0, 1, \ldots, g-1$ polynómom stupňa najviac k. Pre jednoduché uzly λ_i, t.j. pre uzly, v ktorých je defekt $d = 1$, platí

$$B_i^{k+1}(t) \in C^{k-2}[\lambda_1, \lambda_{g-1}],$$

teda

$$B_i^{k+1}(t) \in S_k^{\Delta\lambda}.$$

V tejto časti textu sme prevažne čerpali z [14]. Báza tohoto priestoru $S_{k-1}^{\Delta\lambda}$ je tvorená postupnosťou bázových splajnov $\{B_i^{k+1}(t)\}_{i=-k}^{g}$ a tento priestor je teda rovný priestoru všetkých funkcií $s_k(t)$, $S_k^{\Delta\lambda}[a, b]$. Preto vieme každý splajn $s_k(t) \in S_k^{\Delta\lambda}[a, b]$ písať v tvare

$$s_k(t) = \sum_{i=-k}^{g} b_i B_i^{k+1}(t),$$

kde b_i sú koeficienty splajnu $s_k(t)$ vzhľadom k báze $\{B_i^{k+1}(t), i = -k, \ldots, g\}$. Každý splajn $s_k(t) \in S_k^{\Delta\lambda}[a, b]$ je teda jednoznačne určený vektorom B-splajnových koeficientov $\mathbf{b} = (b_{-k}, \ldots, b_g)^T$.

V maticovom zápise môžeme písať každý splajn $s_k(t)$ vzťahom

$$s_k(\mathbf{t}) = \mathbf{C}_{k+1}(\mathbf{t})\mathbf{b},$$

kde $\mathbf{C}_{k+1}(\mathbf{t})$ je kolokačná matica a jej definícia je uvedená nasledovne.

Definícia čerpaná z [9].

Definícia 6. Nech je dané $\mathbf{t} = (t_1, \ldots, t_n)^T$ a nech $\{B_i^{k+1}(t)\}_{i=-k}^{g}$ je B-splajnová báza z $S_k^{\Delta\lambda}[a, b]$. Potom

$$\mathbf{C}_{k+1}(\mathbf{t}) = \begin{pmatrix} B_{-k}^{k+1}(t_1) & \ldots & B_g^{k+1}(t_1) \\ \vdots & \ddots & \vdots \\ B_{-k}^{k+1}(t_n) & \ldots & B_g^{k+1}(t_n) \end{pmatrix} \in \mathbf{R}^{n,g+k+1}.$$

sa nazýva **kolokačná matica**.

Na rozšírenej sieti uzlov $\Delta\lambda$ je celkom problematické pracovať s presným znením definície bázových splajnov 5 a to hlavne kvôli výpočtom pomerných diferencií. Z tohoto dôvodu sa pre prácu s nimi využívajú rekurentné vzťahy pre výpočet hodnôt bázových splajnov. Nasledujúca veta je prevzaná z [8].

Veta 1. Pre bázové splajny platia následujúce rekurentné vzťahy

$$B_i^1(t) = \begin{cases} 1 & \text{pre} \quad \lambda_i \leq t \leq \lambda_{i+1} \\ 0 & \text{inak} \end{cases}$$

$$B_i^{k+1}(t) = \frac{t-\lambda_i}{\lambda_{i+k}-\lambda_i} B_i^k(t) + \frac{\lambda_{i+k+1}-t}{\lambda_{i+k+1}-\lambda_{i+1}} B_{i+1}^k(t), \quad i = -k, \ldots, g, \quad k = 1, 2, \ldots$$

Uvažujme, že máme $l \in \{1, \ldots, k-1\}$. Môžeme povedať, že derivácia stupňa l splajnu $s_k(t) \in S_k^{\Delta\lambda}[a,b]$ je opäť splajn $s_{k-l}(t) \in S_{k-l}^{\Delta\lambda}[a,b]$ s rovnakými uzlami.

Využitím vlastností B-splajnov, môžeme deriváciu splajnu písať v maticovom zápise nasledovne

$$s_k^{(l)}(\mathbf{t}) = \mathbf{C}_{k+1-l}(\mathbf{t})\mathbf{b}^{(l)},$$

kde $\mathbf{b}^{(l)} \in \mathbf{R}^{g+k+1-l}$ je daný ako

$$\begin{aligned} \mathbf{b}^{(l)} &= \mathbf{D}_l \mathbf{L}_l \mathbf{b}^{(l-1)} \\ &= \mathbf{D}_l \mathbf{L}_l \cdots \mathbf{D}_1 \mathbf{L}_1 \mathbf{b} \\ &= \mathbf{S}_l \mathbf{b} \end{aligned}$$

kde $\mathbf{b}^{(0)} = \mathbf{b}$. Horná trojuholníková matica $\mathbf{S}_l = \mathbf{D}_l \mathbf{L}_l \cdots \mathbf{D}_1 \mathbf{L}_1 \in \mathbf{R}^{g+k+1-l, g+k+1}$ má úplnú riadkovú hodnosť. Matica $\mathbf{D}_j \in \mathbf{R}^{g+k+1-j, g+k+1-j}$ je diagonálna matica

$$\mathbf{D}_j = (k+1-j) diag(d_{-k+j}, \ldots, d_g)$$

s

$$d_i = \frac{1}{\lambda_{i+k+1-j} - \lambda_i} \quad \forall i = -k+j, \ldots, g$$

a

$$\mathbf{L}_j := \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbf{R}^{g+k+1-j, g+k+2-j}.$$

2.2. Vyhladzujúci splajn

Majme údaje $(t_i, f_i), a \leq t_i \leq b$, kde váhy jednotlivých pozorovaní budeme označovať ako $w_i \geq 0$, $i = 1, \ldots, n, n \geq g + 1$ a máme daný parameter $\alpha > 0$. Naša úloha v tomto prípade znie, že pre $l \in \{1, \ldots, k-1\}$ cheme nájsť splajn $s_k(t) \in S_k^{\Delta\lambda}[a, b]$, ktorý by minimalizoval nasledujúci výraz

$$J_l(s_k) = \int_a^b [s_k^{(l)}(t)]^2 dx + \alpha \sum_{i=1}^{n} w_i [f_i - s_k(t_i)]^2. \tag{1}$$

Takýto splajn nazývame **vyhladzovací splajn**. Funkcionál $J_l(s_k)$ môžeme zapísať v maticovej forme nasledovne

$$J_l(\mathbf{b}) = \mathbf{b}^T \mathbf{N}_{kl} \mathbf{b} + \alpha [\mathbf{f} - \mathbf{C}_{k+1}(\mathbf{t})\mathbf{b}]^T \mathbf{W} [\mathbf{f} - \mathbf{C}_{k+1}(\mathbf{t})\mathbf{b}], \tag{2}$$

kde $\mathbf{t} = (t_1, \ldots, t_n)^T$, $\mathbf{f} = (f_1, \ldots, f_n)^T$, $\mathbf{w} = (w_1, \ldots, w_n)^T$ a $\mathbf{W} = diag(\mathbf{w})$.

Matica $\mathbf{N}_{kl} = \mathbf{S}_l^T \mathbf{M}_{kl} \mathbf{S}_l$ je pozitívne semidefinitná matica, tj. pre každé $\mathbf{z}$ nenulové platí $\mathbf{z}^T \mathbf{N}_{kl} \mathbf{z} \geq 0$, kde

$$\mathbf{M}_{kl} = \begin{pmatrix} \langle B_{-k+l}^{k+1-l}, B_{-k+l}^{k+1-l} \rangle & \ldots & \langle B_g^{k+1-l}, B_{-k+l}^{k+1-l} \rangle \\ \vdots & & \vdots \\ \langle B_{-k+l}^{k+1-l}, B_g^{k+1-l} \rangle & \ldots & \langle B_g^{k+1-l}, B_g^{k+1-l} \rangle \end{pmatrix} \in \mathbf{R}^{g+k+1-l, g+k+1-l}$$

a

$$\langle B_i^{k+1-l}, B_j^{k+1-l} \rangle = \int_a^b B_i^{k+1-l}(x) B_j^{k+1-l}(x) dx$$

definujeme ako skalárny súčin B-splajnov v priestore $L^2([a, b])$. Matica $\mathbf{M}_{kl}$ je pozitívne definitná, s ohľadom k vlastnostiam $B_i^{k+1-l}(t) \geq 0, i = -k + l, \ldots, g$.

Celkovým našim cieľom je nájsť splajn $s_k(t) \in S_k^{\Delta\lambda}$, ktorý bude minimalizovať funkcionál $J_l(s_k)$. Každý splajn s_k je jednoznačne určený vektorom B-splajnových koeficientov $\mathbf{b}$ a môžeme teda funkcionál prepísať do tvaru (2) a stačí nájsť vektor $\mathbf{b}$, ktorý bude jednoznačne určovať riešenie. Inými slovami, riešenie tejto úlohy možno rozumieť tak, že hľadáme minimum funkcie $J_l(\mathbf{b})$. Našou úlohou je nájsť

minimum funkcie, teda chceme nájsť bod podozrivý z extrému, ktorý nájdeme tak, že zderivujeme funkciu $J_l(\mathbf{b})$ podľa $\mathbf{b}$ a položíme rovno 0,

$$\frac{\partial J_l(\mathbf{b})}{\partial \mathbf{b}^T} = 0, \quad l \in \{1, \ldots, k-1\}.$$

Z posledného vzťahu dostaneme systém lineárnych rovníc, ktorý budeme riešiť nasledovne:

$$[\alpha^{-1}\mathbf{N}_{kl} + \mathbf{C}_{k+1}^T(\mathbf{t})\mathbf{W}\mathbf{C}_{k+1}(\mathbf{t})]\mathbf{b} = \mathbf{C}_{k+1}^T(\mathbf{t})\mathbf{W}\mathbf{f}.$$

Keďže funkciu $J_l(\mathbf{b})$ vieme upraviť na rýdzo konvexnú kvadratickú funkciu (viď. [1], str. 55) a jej hessián je vtedy pozitívne definitný, čo odpovedá väčšine situácií v praxi, tak máme zaistenú existenciu jediného minima. Riešením tejto sústavy lineárnych rovníc je $\mathbf{b}^*$ v tvare

$$\mathbf{b}^* = [\alpha^{-1}\mathbf{N}_{kl} + \mathbf{C}_{k+1}^T(\mathbf{t})\mathbf{W}\mathbf{C}_{k+1}(\mathbf{t})]^{-1}\mathbf{C}_{k+1}^T(\mathbf{t})\mathbf{W}\mathbf{f}.$$

Nakoniec vyhladzovací splajn pre dané údaje je daný v tvare

$$s_k^*(t) = \sum_{i=-k}^{g} b_i^* B_i^{k+1}(t),$$

kde vektor $\mathbf{b}^* = (b_{-k}^*, \ldots, b_g^*)^T$.

Kapitola 3

Hustoty

Vo funkcionálnej analýze údajov sa veľmi často stretávame s prípadom, keď pozorovanou funkciou je funkcia hustoty. Napríklad, keď uvažujeme realizácie nejakej spojitej premennej, napríklad výšky dospelého človeka (v cm). Tieto údaje možno vykresliť pomocou histogramu, ktorý môžeme aproximovať pomocou funkcie hustoty rozdelenia pravdepodobnosti. V prípade funkcionálnej analýzy táto hustota bude pre nás predstavovať jedno pozorovanie. Vo všeobecnosti sú funkcie hustoty borelovsky merateľné funkcie, ktoré sú nezáporné a obmedzené tak, aby sa integrovali na jednotku. Pretože sa jedná o nezáporné funkcie, ktorých integrál je rovný jednej (konečný), patria do priestoru L^1, čo je priestor všetkých absolútne integrovateľných funkcií na $I \subseteq \mathbf{R}$. Symbolom I budeme označovať interval $[a, b]$. Bohužiaľ, hustoty nám netvoria štruktúru lineárneho priestoru, pretože napríklad pri súčte dvoch hustôt nemusíme znovu dostať hustotu a tak isto násobením hustoty nemusí znovu vyjsť hustota. Hustoty teda môžeme vnímať ako špeciálny prípad funkcionálnych údajov, s obmedzujúcou podmienkou na hodnotu ich integrálu.

Na hustoty sa samozrejme môžeme pozerať aj konkrétne, na rozdiel od funkcií v štandartnom L^2 priestore hustoty majú dve vlastnosti, invariantnosť na zmenu mierky a relatívnosť mierky. **Invariantnosť na zmenu mierky** znamená, že nielen reprezentácia hustôt s jednotkovým integrálom, ale aj akýkoľvek jej kladný násobok poskytuje rovnakú informáciu o relatívnych príspevkoch mier borelovských množín k miere celého nosiča, typicky intervalu $[a, b]$.

Relatívnosť mierky si ľahšie predstavíme pomocou nasledujúceho príkladu. Relatívny nárast hodnoty premennej (v tomto prípade miery borelovskej množiny) z 0,1 na 0,2 (2 násobok) sa líši od nárastu z 0,5 na 0,6, zatiaľ čo absolútne rozdiely sú v oboch prípadoch rovnaké a to 0,1. Kvôli týmto vlastnostiam hustôt sa zdá, že štandardný priestor L^2 je nevhodný na ich zobrazenie. Napríklad súčet dvoch hustôt podľa geometrickej štruktúry priestoru L^2 vedie k funkcii, ktorá už nie je hustotou rozdelenia pravdepodobností. Tento problém je dobre známy v konečnom dimenzionálnom priestore, kde sú známe osobitné techniky na spracovanie takzvaných kompozičných údajov. [6] Kompozičné údaje opisujú časti jedného celku. Zvyčajne sa prezentujú ako vektory pomerov, percent, koncentrácií alebo frekvencií. Keďže proporcie sú vyjadrené ako skutočné čísla, snažíme sa ich interpretovať alebo dokonca analyzovať ako skutočné viacrozmerné údaje. Tieto techniky vychádzajú hlavne z geometrie založenej na Aitchisonovej geometrii v simplexe, ktorý správne zahŕňa povahu kompozičných údajov. Na zvládnutie tohoto druhu údajov, samozrejme v nekonečnej dimenzií (teda hustôt), bola Aitchisonova geometria zovšeobecnená na takzvané **Bayesové priestory**.

3.1. Bayesov priestor

Motivácia zavedenia teórie Bayesovho priestoru je zovšeobecniť Aitchisonovú geometriu pre konečné kompozičné údaje na nekonečne rozmerné súbory. V skutočnosti môže byť akákoľvek hustota rozdelenia pravdepodobností f považovaná za kompozičný vektor s nekonečne mnoho časťami. Ako sme si už vyššie spomenuli, analýza súborov pravdepodobnostných funkcií hustôt v obvyklom priestore L^2 môže viesť k zavádzajúcim výsledkom. Namiesto toho môžu byť špecifické vlastnosti hustôt zachytené cez Bayesové priestory, ktoré sa pri vyrovnávaní s dátovými obmedzeniami spoliehajú na vhodnú štruktúru Hilbertovho priestoru, v našom prípade na intervale $I = [a, b]$.

Bayesov priestor $B^2(I)$ má štruktúru separabilného Hilbertovho priestoru, ktorý umožňuje konštrukciu izometrického izomorfizmu medzi $B^2(I)$ a $L^2(I)$, kde priestor L^2 je obmedzený na I. Dve kladné funkcie f a g s rovnakým nosičom

považujeme za **ekvivalentné**, ak platí $f = c \cdot g$ pre kladnú konštantu c. Pripomeňme vlastnosť invariantnosti na zmenu mierky funkcií hustoty. Z toho vyplýva, že hustoty v rámci triedy ekvivalencie poskytujú rovnaké relatívne informácie alebo nesú rovnakú relatívnu informáciu o príspevku akejkoľvek miery (či špeciálne pravdepodobnosti) z $\mathbf{R}$ na mieru celého nosiča. Pravdepodobnosť daného javu totiž nemá význam sama o sebe, ale mala by sa porovnávať s pravdepodobnosťou celého priestoru vzorky, krorá je nastavená na 1, a môžeme ju rovnocenne nastaviť na inú pozitívnu konštantu c.

Bayesov priestor $B^2(I)$ pozostáva z hustôt f (ekvivalentných tried hustôt) na nosiči I pre ktorý je logaritmus štvorcovo integrovateľný.

Tak isto aj v tomto preistore $B^2(I)$ dokážeme pracovať s operáciami sčitaním dvoch funkcií (hustôt) a vynásobením hustoty konštantou. Majme dané dve integrovateľné funckie hustoty $f, g \in B^2(I)$ a reálne číslo $c \in \mathbf{R}$. Operáciou $f \oplus g$ označíme perturbáciu a $c \odot f$ mocninnú transformáciu. Definujeme ich pomocou vzťahov

$$(f \oplus g)(t) = \frac{f(x)g(x)}{\int_I f(y)g(y)dy},$$

$$(c \odot f)(t) = \frac{f(x)^c}{\int_I f(y)^c dy}.$$

Výsledné funkcie týchto operácií sú tiež funkcie hustoty, aj keď normovanie na jednotkový integrál nie je z hľadiska invariantnosti na zmenu merítka nutné. Ako môžeme vidieť v [3], je dokázané, že $B^2(I)$ spolu s operáciami $(\oplus, \odot)$ je vektorový priestor. Neutrálne prvky perturbácie a mocninnej transformácie sú $e(x) = \frac{1}{\eta}$, kde $\eta = b - a$ a 1.

V tomto priestore je taktiež možné vyjadriť rozdiel medzi dvomi hustotami $f, g \in B^2(I)$, ktorý budeme značiť ako $f \ominus g$ a je daný vzťahom:

$$(f \ominus g)(t) = (f \oplus [(-1) \odot g])(x), t \in I.$$

Operáciu perturbácie $\oplus$ môžeme interpretovať aj ako bayesovskú aktualizáciu informácií a operáciu $\ominus$ ako nejaké zrušenie informácie. Aby sme $B^2(I)$ mohli

definovať ako Hilbertov priestor, musíme ešte definovať skalárny súčin:

$$\langle f, g\rangle_B = \frac{1}{2\eta}\int_I\int_I \ln\frac{f(x)}{f(y)}\ln\frac{g(x)}{g(y)}dx\,dy, \quad f, g \in B^2(I),$$

kde $\eta = b - a$ je dĺžka intervalu I, ktorý indukuje nasledujúcu normu:

$$||f||_B = \Big[\frac{1}{2\eta}\int_I\int_I \ln^2\frac{f(x)}{f(y)}\Big]^{\frac{1}{2}}.$$

Forma skalárneho súčinu napovedá, že relatívna informácia v hustotách je obsiahnutá v (logaritmických) pomeroch medzi prvkami nosiča I. Ako sme si už spomínali, funkcie hustoty môžeme považovať za funkčné náprotivky ku kompozičným údajom, teda kladným vektorom, ktoré nosia relatívnu informáciu a sú riadené Aitchisonovou geometiou. Aby bolo umožnené ich štatistické spracovanie pomocou štandardných viacrozmerných metód v reálnom priestore, upredňostňovanou stratégiou je ich vyjadrenie v tzv. **clr koeficientoch** (centrované logaritimicke pomery) s ohľadom na generovaný systém, alebo v logaritmicko pomerových súradniciach s ohľadom na ortonormálnu bázu. Tieto súradnice, ktoré tiež nazývame izometrické logaritmicko pomerové súradnice, tak isto ako clr koeficienty, poskytujú izometriu medzi Aitchisonovou geometriou a reálnym euklidovským priestorom. Podobná stratégia sa používa tiež pre hustoty v Bayesovom priestore. Izometrický izomorfizmus medzi $B^2(I)$ a $L^2(I)$ je reprezentovaný pomocou **clr transformácie**, definovanej pre $f \in B^2(I)$ ako

$$clr(f)(x) := f_c(x) = \ln f(x) - \frac{1}{\eta}\int_I \ln f(y)dy. \tag{3}$$

Poznamenajme, že takáto izometria umožňuje počítať operácie a skalárny súčin medzi prvkami v $B^2(I)$ z hľadiska ich náprotivkom v $L^2(I)$ medzi clr-transformáciami. Môžeme ich rátať nasledovne:

- $clr(f \oplus g)(x) = f_c(x) + g_c(x)$,
- $clr(c \odot f)(x) = c \cdot f_c(x)$,

- $\langle f, g\rangle_B = \langle f_c, g_c\rangle_2 = \int_I f_c(x)g_c(x)dx.$

Z konštrukcie clr transformácie plynie nulovosť integrálu clr transformovaných hustôt,

$$\int_I clr(f)(x)dx = \int_I \ln f(x)dx - \int_I \frac{1}{\eta}\int_I \ln f(y)dy\,dx = 0,$$

čo je potrebné vziať do úvahy pri výpočte a analýze clr transformovaných hustôt. Pretože clr priestor je jednoznačne podpriestorom $L^2(I)$, ďalej ho môžeme označovať ako $L_0^2(I)$. Inverznú clr transformáciu vieme získať pomocou vzťahu

$$clr^{-1}[f_c](x) = \frac{\exp(f_c(x))}{\int_I \exp(f_c(y))dy},$$

kde opäť, ako predtým, sa menovateľ používa len na dosiahnutie jednotkového integrálneho obmedzenia výslednej hustoty. Takýmto spôsobom nestratíme relatívnu informáciu, ktorá je nesená pomocou funkcie hustoty.

Kapitola 4

Metóda hlavných komponentov

Jedným z problémov viacrozmerných údajov je grafická interpretácia. Rovnako môže nastať problém pri využití niektorých viacrozmerných metód, kedy je ich interpretovateľnosť limitovaná vysokou dimenzionalitou údajov. Riešením takejto situácie je **metóda hlavných komponentov** (angl. principal component analysis, PCA), ktorá nám umožňuje zhrnúť podstatnú informáciu z údajov (rozptyl) do nových (skrytých, latentných) premenných. Tieto premenné - **hlavné komponenty** - budeme konštruovať tak, aby sme vyčerpali čo najväčšiu časť variability pôvodných údajov. Variabilitu pritom budeme maximalizovať len v triede lineárnych kombinácií pôvodných premenných.

Základnou úlohou PCA vo viacrozmernom prípade je popísať rozptyl množiny korelovaných premenných $\mathbf{x} = (x_1, \ldots, x_p)^T$ pomocou nových nekorelovaných premenných $\mathbf{y} = (y_1, \ldots, y_p)^T$, z ktorých každá je lineárna kombinácia pôvodných premenných. Nové premenné budú mať klesajúcu dôležitosť v tom zmysle, že prvá premenná nám bude vyjadrovať najväčšiu časť variability:

- Prvá premenná vyjadruje najväčšiu časť variability (informácií).
- Druhá premenná vyjadruje druhú najväčšiu časť variability za podmienky, že je nekorelovaná s prvou premennou.
- Tretia premenná vyjadruje tretiu najväčšiu časť variability za podmienky, že je nekorelovaná s oboma predchádzajúcimi premennými.

- A tak ďalej, až dosiahneme cieľového počtu p nových premenných (hlavných komponentov).

Ku grafickej interpretácií využívame často len prvé dve hlavné komponenty. Metóda hlavných komponentov môže slúžiť tiež ako pomocná metóda pri množstve iných metód, ako je napríklad regresná analýza, kde máme jednu vysvetľovanú premennú, ale veľké množstvo vysvetľujúcich premenných a malé množstvo pozorovaní. V tomto zmysle budeme metódu hlavných komponentov využívať aj v tejto práci. Metóda hlavných komponentov je odlišná pre viacrozmerné a pre funkcionálne údaje. V tejto kapitole si na začiatku ukážeme PCA pre viacrozmerné údaje, následne PCA pre funkcionálne údaje a nakoniec si zobrazíme špecálny prípad PCA pre hustoty.

4.1. PCA pre viacrozmerné údaje

V tejto podkapitole sme čerpali z [4]. Predpokladom pre niektoré viacrozmerné metódy je viacrozmerné normálne rozdelenie. PCA za tejto podmienky dosahuje optimálnych výsledkov, nie je to samozrejme nutnosťou. Pre jednoduchosť sa pri výklade pridržíme teoretickej verzie PCA. Predpokladajme, že máme náhodný vektor $\mathbf{x} = (x_1, \ldots, x_p)^T$ z p-rozmerného normálneho rozdelenia s vektorom stredných hodnôt $\boldsymbol{\mu}$ a s variančnou maticou $\boldsymbol{\Sigma}$. Zároveň budeme predpokladať, že $\boldsymbol{\mu} = \mathbf{0}$.

Prvá premenná je lineárna kombinácia $y_1 = \xi_{11}x_1 + \xi_{12}x_2 + \cdots + \xi_{1p}x_p$. Voľba $\boldsymbol{\xi}_1 = (\xi_{11}, \ldots, \xi_{1p})^T$, ktoré by maximalizovalo rozptyl y_1, by bola nekonečno, preto pridáme obmedzujúcu podmienku (najčastejšie $\boldsymbol{\xi}_1^T\boldsymbol{\xi}_1 = 1$, t.j. súčet štvorcov koeficientov je jedna).

Rozptyl y_1 potom spočítame ako $\boldsymbol{\xi}_1^T\boldsymbol{\Sigma}\boldsymbol{\xi}_1$, kde vektor $\boldsymbol{\xi}_1$ hľadáme ako vlastný vektor matice $\boldsymbol{\Sigma}$ odpovedajúci najväčšiemu vlastnému číslu tejto matice (vychádza z výpočtu pomocou Lagrangeových multiplikátorov). Vlastné číslo vypočítame z rovnice $det(I\lambda - \boldsymbol{\Sigma}) = 0$ a vlastný vektor $\boldsymbol{\xi}_1$ je potom taký vektor, pre ktorý platí: $\boldsymbol{\Sigma}\boldsymbol{\xi}_1 = \lambda_1\boldsymbol{\xi}_1$.

Druhá premenná $y_2 = \xi_{21}x_1 + \xi_{22}x_2 + \cdots + \xi_{2p}x_p$ sa vypočíta analogicky s dodatočnou podmienkou $\boldsymbol{\xi}_1^T\boldsymbol{\xi}_2 = 0$ (nekorelovanosť nových premenných).

Potom podobne pre každú ďalšiu premennú platia nasledujúce podmienky:

- $\boldsymbol{\xi}_j^T\boldsymbol{\xi}_j = 1,$
- $\boldsymbol{\xi}_j^T\boldsymbol{\xi}_i = 0, \forall i < j.$

Za uvedenej normujúcej podmienky tiež platí, že rozptyl j-tej hlavnej komponenty je rovný j-temu najväčšiemu vlastnému číslu λ_j. Súčet všetkých λ_i potom vyjadruje celkovú variabilitu v pôvodných údajoch a pomer

$$P_j = \frac{\lambda_j}{st(\boldsymbol{\Sigma})}$$

vyjadruje príspevok j-tej komponenty k celkovej variabilite (st($\boldsymbol{\Sigma}$) značí stopu matice $\boldsymbol{\Sigma}$). Nasledujúci vzťah

$$P^{(m)} = \frac{\sum_{j=1}^{m}\lambda_j}{st(\boldsymbol{\Sigma})}$$

popisuje akú variabilitu vyjadruje prvých m premenných.

Pokiaľ uvažujeme výberovú obdobu PCA, teda predpokladáme na vstupe N p-rozmerných vektorov $\mathbf{x}_1, \ldots, \mathbf{x}_N$, uvažujeme ešte nasledujúce pojmy. Súradnice n objektov v priestore hlavných komponentov nazývame **skóre** a pre i-tý objekt ich určíme ako

$$y_{i1} = \boldsymbol{\xi}_1^T\mathbf{x}_i, \ \ldots\ , y_{ip} = \boldsymbol{\xi}_p^T\mathbf{x}_i.$$

Tieto skóre vieme zapísať do matice $\boldsymbol{\Psi}_{n\times l}$. Vlastné vektory $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p$ príslušné variančnej matici $\boldsymbol{\Sigma}$ nazývame **záťaže**, ktoré určujú smery hlavných komponentov, čo znamená, že nám určujú smery najväčšej variability v údajoch.

Pre výberovú obdobu PCA ešte potrebujeme určiť odhad variančnej matice $\boldsymbol{\Sigma}$. Odhad matice $\boldsymbol{\Sigma}$ určíme ako výberovú variančnú maticu $\mathbf{V}$ danú vzťahom

$$\mathbf{V} = \frac{1}{N}\mathbf{x}^T\mathbf{x}.$$

Čím viac budú pôvodné premenné korelované, tým viac sa budú líšiť hodnoty rozptylu vyjadreného jednotlivými komponentami. Teda pri silnej korelácií bude treba menej premenných k vyjadreniu veľkej časti variability. Väčšinou sa vyžaduje okolo 80% variability v údajoch, toto číslo však výrazne závisí na počte premenných, s ktorými pracujeme.

4.2. PCA pre funkcionálne údaje

V tejto podkapitole sme používali literatúru [7] a [12]. Vo funkcionálnom kontexte dosiahneme rovnaký cieľ pomocou analýzy funkcionálnych hlavných komponentov, ktorú budeme označovať ako FPCA.

Uvažujme náhodnú vzorku funkcionálnych údajov $X_1, \ldots, X_N$ v $L^2(I)$. Opäť označme $\langle x, y\rangle_2 = \int_I x(t)y(t)dt$ skalárny súčin medzi dvoma prvkami x, y v $L^2(I)$ a $||x||_2 = (\int_I |x(t)|^2 dt)^{\frac{1}{2}}$ normu prvku x z priestoru $L^2(I)$. V práci predpokladáme centrované vzorky (t.j. také, od ktorých sme odčítali hodnotu funkcionálneho aritmetického priemeru) z dôvodu ľahšieho zápisu.

Ako sme si uvideli pri PCA viacrozmerných údajov, rovnako v prípade FPCA nás na začiatku zaujíma zdroj najväčšej variability v údajoch. Hľadáme prvok ξ_1 v priestore $L^2(I)$ - tiež nazývaný ako prvá funkcionálna hlavná komponenta (zn. FPC). Túto komponentu dostaneme maximalizáciou výrazu

$$\frac{1}{N}\sum_{i=1}^{N}\langle X_i, \xi\rangle_2^2 \tag{4}$$

cez všetky možné $\xi \in L^2(I)$, za podmienky, že $||\xi||_2 = 1$.

Zostávajúce komponenty FPC $\{\xi_j\}_{j\geq 2}$ by mali vysvetliť čo najviac zo zostávajúcej variability a chceme, aby komponenty boli na seba kolmé. To znamená, že budeme riešiť úlohu (4) s pridanou podmienkou ortogonality $\langle \xi_j, \xi_i\rangle_2 = 0, i < j$. Hodnoty FPC odpovedajú vlastným funkciam kovariančného operátora $V : L^2(I) \to L^2(I)$, definovaného pre $\xi \in L^2(I)$ ako

$$V_\xi = \frac{1}{N}\sum_{i=1}^{N}\langle X_i, \xi\rangle_2 X_i,$$

alebo ekvivalente vieme zapísať aj nasledovne

$$V_\xi = \int_I v(.,t)\xi(t)dt,$$

kde $v : I \times I \to \mathbf{R}$ je výberová kovariančná funkcia

$$v(s,t) = \frac{1}{N}\sum_{i=1}^{N} x_i(s)x_i(t), \quad s,t \in I.$$

Následne možno j-tú FPC ξ_j a k nej príslušné skóre

$$\Psi_{ij} = \langle X_i, \xi_j \rangle_2, \quad i = 1, \ldots, N$$

získať riešením nasledujúcej rovnice

$$V\xi_j = \rho_j \xi_j, \tag{5}$$

kde ρ_j označujeme j-te vlastné číslo, pre ktoré platí $\rho_1 \geq \rho_2 \geq \cdots$. Tak ako aj v PCA viacrozmerných údajov, vieme aj vo funkcionálnom prípade vyjadriť podiel celkovej variability vysvetleným pomocou FPC ξ_j, a to vzťahom $\frac{\rho_j}{\sum_j \rho_j}$. Na riešenie rovnice (5) môžeme použiť niekoľko výpočtových metód.

Jedným zo spôsobou ako redukovať rovnicu (5) na diskrétnu alebo maticovú formu je vyjadriť každú funkciu $X_i, i = 1, \ldots, N$ ako lineárnu kombináciu známych bázových funkcií ϕ_k. Počet použitých K bázových funkcií závisí od mnohých hľadísk: koľko diskrétnych vzorových bodov n bolo v pôvodných údajoch, či bola určitá úroveň vyhladenia uložená pomocou $K < n$, ako efektívne alebo výkonné sú bázové funkcie pri reprodukcií správania pôvodných funkcií atď..

Predpokladajme, že každú funkciu X_i vieme zapísať pomocou bázových funkcií ako

$$X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t). \tag{6}$$

Túto rovnicu (6) vieme zapísať aj kompaktnejšie, definovaním vektorovej funkcie $\mathbf{X} = (X_1, \ldots, X_N)^T$ a vektorovej funkcie $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)^T$. Potom môžeme vyjadriť simultánne rozšírenie všetkých N kriviek ako

$$\mathbf{X} = \mathbf{C}\boldsymbol{\phi},$$

kde $\mathbf{C} = (c_{ik}) \in \mathbf{R}^{N,K}$ značíme ako maticu koeficientov. Pomocou maticového zápisu môžeme zapísať výberovú variančnú funkciu

$$v(s,t) = \frac{1}{N}\boldsymbol{\phi}(s)^T\mathbf{C}^T\mathbf{C}\boldsymbol{\phi}(t), \qquad s,t \in I,$$

kde $\boldsymbol{\phi}(s)^T$ značí transpozíciu vektora funkcií $\boldsymbol{\phi}(s)$.

Definujme si symetrickú maticu $\mathbf{W}$ hodností K so zložkami

$$w_{k_1,k_2} = \int_I \phi_{k_1}\phi_{k_2}\,dt,$$

alebo maticovo $\mathbf{W} = \int_I \boldsymbol{\phi}\boldsymbol{\phi}^T\,dt$. Teraz predpokladajme, že vlastné funkcie ξ z rovnice (5) vieme vyjadriť ako

$$\xi(s) = \sum_{k=1}^{K} b_K\phi_K(s)$$

alebo maticovým zápisom ako

$$\xi(s) = \boldsymbol{\phi}(s)^T\mathbf{b}.$$

Toto využijeme v nasledujúcom zápise

$$\int_I v(s,t)\xi(t)dt = \int_I N^{-1}\boldsymbol{\phi}(s)^T\mathbf{C}^T\mathbf{C}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)^T\mathbf{b}dt = \boldsymbol{\phi}(s)^T N^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b}.$$

Potom rovnicu (5) môžeme vyjadriť ako

$$\boldsymbol{\phi}(s)^T N^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\boldsymbol{\phi}(s)^T\mathbf{b}$$

a upravením dostaneme

$$N^{-1}\mathbf{C}^T\mathbf{C}\mathbf{W}\mathbf{b} = \rho\mathbf{b}.$$

Nesmieme zabudnúť, že platí $||\xi|| = 1$, čo znamená, že $\mathbf{b}^T\mathbf{W}\mathbf{b} = 1$ a podobne, dve funkcie ξ_1 a ξ_2 sú ortogonálne vtedy a len vtedy, ak príslušné vektory koeficientov spĺňajú $\mathbf{b}_1^T\mathbf{W}\mathbf{b}_2 = 0$. Na získanie požadovaných hlavných komponentov si zadefinujeme $\mathbf{u} = \mathbf{W}^{\frac{1}{2}}\mathbf{b}$, ktoré dostaneme vyriešením rovnice

$$\frac{1}{N}\mathbf{W}^{\frac{1}{2}}\mathbf{C}^T\mathbf{C}\mathbf{W}^{\frac{1}{2}}\mathbf{u} = \rho\mathbf{u}$$

a z toho potom dopočítame $\mathbf{b} = \mathbf{W}^{\frac{-1}{2}}\mathbf{u}$ pre každý vlastný vektor.

Na rozdiel od viacrozmernej PCA, kde väčšinou platí $p < n$, t.j. počet premenných je menší ako počet pozorovaní, v FPCA je počet funkčných hodnôt u každeho pozorovania rovný nekonečnu. Na záver ešte poznamenajme, že viacrozmerná PCA často vychádza z korelačnej matice z dôvodu odstránenia vplyvu rozdielnej mierky, pretože v údajoch sa nám často stáva, že jednotlivé vlastnosti sú v rôznych jednotkách. Namiesto toho vo funkcionálnej PCA máme súbor pozorovaní meraním jednej veličiny v rovnakej jednotke, preto vždy vychádzame z kovariančnej funkcie.

4.3. SFPCA

Ako bolo v predchádzajúcich kapitolách naznačené, ak uvažujeme ako funkcie hustoty, tak sa nám niektoré vlastnosti zmenia. Tak isto to bude v prípade metódy hlavných komponentov. Ako funkcionálne kompozície, hustoty rozdelenia pravdepodobnosti sa vyznačujú špeciálnymi vlastnosťami, ako invariantnosť na zmenu mierky a vlastnosti relatívnosti mierky, ktoré sme popísali v kapitole 3. Cieľom tejto sekcie bude odvodiť špeciálnu verziu FPCA pre funkcie hustoty, budeme ju označovať ako SFPCA (z anglíckeho Simplicial functional principal component analysis). Budeme sa držať rovnakej schémy ako pri FPCA, ale zároveň si musíme uvedomiť, že sa budeme nachádzať v Bayesovom priestore $B^2(I)$ a budeme používať špeciálnu geometriu v tomto priestore.

Uvažujme $\widetilde{X}_1, \ldots, \widetilde{X}_N$ vzorku z priestoru $B^2(I)$ a označme si ako $X_1, \ldots, X_N$ príslušné centrované vzorky, pre $i = 1, \ldots, N$, tak, že platí $X_i = \widetilde{X}_i \ominus \bar{X}$, kde $\bar{X}$ značí výberový priemer $\bar{X} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \widetilde{X}_i$. Uvažujme problém, že hľadáme špeciálne funkcionálne hlavné komponenty (zn. SFPC) v $B^2(I)$, tzn. prvky $\{\xi_j\}_{j\geq 1}, \xi_j \in B^2(I)$. Tieto prvky (komponenty) získame maximalizáciou výrazu

$$\frac{1}{N}\sum_{i=1}^{N}\langle X_i, \xi\rangle_B^2 \tag{7}$$

cez všetky možné $\xi \in B^2(I)$ za podmienky, že $||\xi||_B = 1$ a $\langle \xi_j, \xi_k\rangle_B = 0, k < j$,

kde podmienka ortogonality $\langle \xi, \xi_k \rangle_B = 0$, pre $k < j$ platí iba pre $j \geq 2$.

$B^2(I)$ je separabilný Hilbertov priestor (viď. kap. 1.1.), úloha (7) je teda správne sformulovaná. Potom teda riešenie úlohy (7) existuje a je jedinečné. Ako sme si spomínali vyššie, budeme sa chcieť držať schémy z FPCA, teda j-ta SFPC je riešením rovnice

$$V\xi_j = \lambda_j \odot \xi_j,$$

kde (λ_j, ξ_j) odpovedá j-temu vlastnému páru výberového kovariančného operátora $V : B^2(I) \to B^2(I)$, definovaného pre $\xi \in B^2(I)$ ako

$$V\xi = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \langle X_i, \xi \rangle_B \odot X_i.$$

Aby sme mohli používať vzťah (7) v praxi, aplikujeme izometrický izomorfizmus medzi $B^2(I)$ a $L^2(I)$, definovaný prostredníctvom clr-transformácie (3), ktorá nám umožňuje prepísať úlohu (7) ako maximalizáciu úlohy

$$\frac{1}{N} \sum_{i=1}^{N} \langle clr(X_i), clr(\xi) \rangle_2^2$$

s podmienkami $||clr(\xi)||_2 = 1$ a $\langle clr(\xi), clr(\xi_k) \rangle_2 = 0, k < j$ cez všetky možné $\xi \in B^2(I)$.

Pre $j \geq 1$ maximalizačnú úlohu (4) môžeme ekvivalentne prepísať ako hľadanie $\nu \in L_0^2(I)$ ktorý maximalizuje

$$\frac{1}{N} \sum_{i=1}^{N} \langle clr(X_i), \nu \rangle_2^2 \tag{8}$$

s podmienkami $||\nu||_2 = 1, \langle \nu, \nu_k \rangle_2 = 0, k < j$ a $\int_I \nu = 0$, kde ortogonálne obmedzenie má význam len pre $j \geq 2$ a obmedzenie na nulový integrál zodpovedá vlastnostiam clr-transformovaných hustôt.

Teraz si ukážeme, že maximalizáciu výrazu (8) vieme vyriešiť pomocou vlastných funkcií $\{\xi_j\}_{j \geq 1}$ výberového kovariančného operátora $V_{clr} : L_0^2(I) \to L_0^2(I)$ transformovaných vzoriek $clr(X_1), \ldots, clr(X_N)$, pre všetky $\xi \in L_0^2(I)$ ako

$$V_{clr}\xi = \frac{1}{N} \sum_{i=1}^{N} \langle clr(X_i), \xi \rangle_2 clr(X_i).$$

Prvé čo si musíme uvedomiť je, že vlastné funkcie $\{\xi_j\}_{j\geq 1}$ budú riešením úlohy (8), ak by sme ich mali bez podmienky na nulový integrál $\int_I \nu = 0$, pretože v takomto prípade by (8) bolo ekvivalentné s (4) (s podmienkou ortogonality pre $j \geq 2$). Teda aby sme dokázali, že $\nu = \xi_j$ maximalizuje úlohu (8), stačí dokázať, že ξ_j spĺňa tiež obmedzenie $\int_I \xi_j = 0, \forall j \geq 1$. Na zakončenie tejto myšlienky si uvedomme, že vlastnosť nulového integrálu clr-transformovaných vzoriek $clr(X_1), \ldots, clr(X_N)$ implikuje, že V_{clr} pripúšťa nulové vlastné číslo s príslušnou vlastnou funkciou $\xi_0 := \frac{1}{\sqrt{b-a}}$:

$$V_{clr}\xi_0 = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{b-a}}\Big[\int_I clr(X_i)\Big] clr(X_i) = 0.$$

Pretože vlastné funkcie $\{\xi_j\}$ príslušiace nenulovým vlastným číslam $\{\rho_j\}$ sú ortogonálne s vlastnou funkciou ξ_0, ξ_j musia spĺňať podmienku nulovosti integrálu $\int_I \xi_j = 0$ a $\langle \xi_j, \xi_0 \rangle_2 = \frac{1}{\sqrt{b-a}} \int_I \xi_j = 0$.

Každú vlastnú funkciu ξ_j môžeme zapísať jedinečnou lineárnou kombináciou funkcií $clr(X_1), \ldots, clr(X_n)$. Z toho dôvodu podmienka nulovosti integrálu je splnená, pretože platí podľa konštrukcie pre každú funkciu $clr(X_i), i = 1, \ldots, N$. Teda úloha (7) môže byť preformulovaná z hľadiska clr-transformácií na úlohu (8) a SFPC môžeme dostať transformáciou vlastných funkcií $\{\xi_j\}_{j\geq 1}$ príslušných nenulovým vlastným číslam $\{\rho_j\}_{j\geq 1}$ z V_{clr} cez inverznú clr transformáciu,

$$\xi_j = clr^{-1}(\xi_j) =_B \exp(\xi_j), \quad j \geq 1.$$

Na vyrátanie vlastných funkcií ξ_j využijeme známu metódu, ktorá je založená na B-splajnovom bázovom rozšírení (sekcia 2.2.1). Pomocou [9], môžeme uvažovať pre $clr(X_1), \ldots, clr(X_N)$ a $\xi_j, j \geq 1$, B-splajnovú bázu spňujúcu podmienku nulového integrálu, môžeme ich teda zapísať nasledovne:

$$clr(X_i)(.) = \sum_{k=1}^{K} c_{ik}\phi_k(.),$$

$$\xi_j(.) = \sum_{k=1}^{K} b_{jk}\phi_k(.).$$

Hľadanie vlastných funkcií ξ_j prechádza rovnakým spôsobom ako u FPCA na úlohu hľadania vlastných vektorov $\mathbf{b}_j = (b_{jk})$ riešením sústavy

$$\frac{1}{N}\mathbf{C}^T\mathbf{CMb}_j = \rho_i\mathbf{b}_j,$$

s rovnakými argumentami ortogonality, ako boli predtým uvedené, sa podmienka nulovosti integrálu zachová v PCA algoritme a preto nie je nutné ju zavádzať.

Na vypočítanie SFPCA by sme mohli použiť aj alternatívne výpočtové prístupy s ktorými sa avšak v tejto práci nebudeme zaberať, ale môžeme ich nájsť v [2] a [11].

Pri všetkých PCA, či viacrozmernej, funkcionálnej alebo SFPCA, je dôležitým nástrojom na interpretáciu graf skóre pre dvojicu hlavných komponentov, ktorý umožní urobiť si predstavu o údajovej štruktúre (skupiny pozorovaní v údajoch, odľahlé hodnoty).

Na určenie vhodného počtu komponentov sa využíva tzv. **scree diagram**, ktorý zobrazuje podiely vysvetlenej variability jednotlivými komponentami. Vykreslíme krivku, ktorá prejde všetkými bodmi a nájdeme najväčší (najostrejší) zlom, ktorý nám značí poslednú uvažovanú komponentu.

Kapitola 5

Regresná analýza

Regresná analýza je vhodná štatistická metóda pre modelovanie lineárneho vzťahu medzi závislou premennou (vysvetľovaná) a jednou alebo viacerými nezávislými premennými (prediktory). Pokiaľ vysvetľovaná premenná alebo prediktory majú funkcionálny charakter, musíme uvažovať funkcionálnu regresnú analýzu. Mnoho autorov sa zaoberá teoretickými i praktickými aspektami, spojenými s funkcionálnymi lineárnymi modelmi ([5, 12, 15]). Táto práca sa bude predovšetkým zameriavať na funkcionálnu regresiu, oblasť FDA ktorá si zaslúžila najväčšiu pozornosť vo vývoji metodológie.

V tejto kapitole bolo najviac čerpané z [12], [13], [16] a [18].

5.1. Funkcionálny lineárny regresný model pre skalárnu vysvetľovanú premennú s prediktorom z priestoru L^2

V tejto časti si predstavíme regresiu v prípade skalárnej premennej y s nezávislou funkcionálnou premennou $f(t)$ a absolútnym členom.

Uvažujme sadu N párov $(y_1, f_1), \ldots, (y_N, f_N)$, kde $y_1, \ldots, y_N$ je náhodná vzorka reálnej premennej y a $f_1, \ldots, f_N$ je funkcionálná vzorka funkcionálnej premennej f v priestore $L^2(I)$.

Funkcionálny lineárny model pre i-té pozorovanie y_i s prediktorom f_i

vyjadríme vzťahom

$$y_i = \beta_0 + \int_I \beta_1(t) \cdot f_i(t) dt + \epsilon_i = \beta_0 + \langle \beta_1(t), f_i(t) \rangle_2 + \epsilon_i, \quad i = 1, \ldots, N, \quad t \in I, \tag{9}$$

alebo v maticovom zápise, kde $\mathbf{f}(t) = (f_1(t), \ldots, f_N(t))^T$ ako

$$\mathbf{y}_{N\times 1} = \mathbf{1}_{N\times 1}\beta_0 + \int_I \mathbf{f}(t)_{N\times 1} \cdot \beta_1(t) dt + \epsilon_{N\times 1}, \tag{10}$$

kde $\beta_0 \in \mathbf{R}$ a $\beta_1(t) \in B^2(I)$ sú neznáme regresné parametre a ϵ je vektor náhodných chýb modelu, s nulovou strednou hodnotou a konečným rozptylom, takých že sú nezávislé od prediktora.

Odhad regresných parametrov β_0 a β_1 môžeme nájsť minimalizáciou súčtu štvorcových chýb (angl. sum of squared errors)

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \langle f_i, \beta_1 \rangle_2)^2. \tag{11}$$

Rovnicu (11) môžeme riešiť dvomi rôznymi metódami.

Prvá je založená na bázovom rozšírení funkcionálnych premenných $f_1(t), \ldots, f_N(t)$ a regresného parametru $\beta_1(t)$. Druhá je založená na funkcionálnych hlavných komponentoch odvodených z funkcionálnych pozorovaní $f_1(t), \ldots, f_N(t)$.

Ukážeme si najskôr prvú metódu. Predpokladajme, že f_i a β_1 vieme vyjadriť pomocou bázového rozšírenia,

$$f_i(t) = \sum_{k=1}^{K_f} c_{ik}\phi_k(t), i = 1, \ldots, N$$

$$\beta_1(t) = \sum_{k=1}^{K_\beta} b_k \psi_k(t), \tag{12}$$

kde $\{c_{ik}\}$ a $\{b_k\}$ sú bázové koeficienty, a $\boldsymbol{\phi}(t) = \{\phi_1(t), \ldots, \phi_{K_f}(t)\}^T$ a $\boldsymbol{\psi}(t) = \{\psi_1(t), \ldots, \psi_{K_\beta}(t)\}^T$ sú známe bázové systémy, pre ktoré platí $K_f \geq K_\beta$. Rovnicu (12) vieme vyjadriť aj v maticovom tvare ako $\mathbf{f}(t) = \mathbf{C}\boldsymbol{\phi}(t)$ a $\beta_1(t) = \boldsymbol{\psi}(t)^T\mathbf{b}$, kde $\mathbf{C} \in \mathbf{R}^{N,K_f}$ a $\mathbf{b} \in \mathbf{R}^{K_\beta,1}$ je matica a vektor bázových koeficientov a $\boldsymbol{\phi}, \boldsymbol{\psi}$

sú vektory bázových funkcií. Pokiaľ model (10) nahradíme maticovým zápisom (12), zredukujeme model na viacrozmerný

$$\mathbf{y}_{N\times 1} = \mathbf{X}_{N\times(K_\beta+1)}\tilde{\mathbf{b}} + \boldsymbol{\epsilon}_{N\times 1}, \tag{13}$$

kde $\tilde{\mathbf{b}} = (\beta_0, b_1, \ldots, b_{K_\beta})^T$, $\mathbf{X} = [\mathbf{1}_{N\times 1}; \mathbf{C}\mathbf{M}_{\phi,\psi}]$ a

$$\mathbf{M}_{\phi,\psi} = \begin{pmatrix} \langle\phi_1,\psi_1\rangle_2 & \cdots & \langle\phi_1,\psi_{K_\beta}\rangle_2 \\ \vdots & \ddots & \vdots \\ \langle\phi_{K_f},\psi_1\rangle_2 & \cdots & \langle\phi_{K_f},\psi_{K_\beta}\rangle_2 \end{pmatrix} \in \mathbf{R}^{K_f,K_\beta} \tag{14}$$

je matica skalárnych súčinov všetkých párov bázových funkcií $\{\phi_k\}$ a $\{\psi_k\}$. Takýto systém je typicky tvorený B-splajnovými funkciami. V dôsledku toho je odhad $\tilde{\mathbf{b}}$ uvedený ako riešenie nasledujúcej rovnice

$$\mathbf{X}^T\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}^T\mathbf{y},$$

a to ako

$$\hat{\tilde{\mathbf{b}}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Potom odhad hodnôt y je daný ako

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\tilde{\mathbf{b}}}.$$

Pretože je k dispozícií len konečný počet pozorovaní na určenie nekonečne rozmerného parametra β_1, existuje nekonečný počet riešení, ktoré interpretujú rovnako riešenie minimalizačnej úlohy (11). Na vyriešenie tohoto problému je možné prijať stratégiu ponechania dimenzionality K_β z β_1 vzhľadom na N. Toto rozhodne môže redukovať rozptyl v odhadoch a pomáha zlepšiť predikcie budúcich pozorovaní. Napriek tomu sa môže stať, že pri málo rozmernom predpoklade o parametrickom priestore by mohlo dôjsť k strate niektorých dôležitých znakov β_1. Ak by sme uvažovali bohatší bázový systém, pravdepodobne by to malo za následok **overfitting** a nezmyselné odhady s väčším rozptylom, čo by spolu viedlo k zlým predikciam. Z tohoto dôvodu je možné použiť metódu, ktorá pri odhadovaní modelu kombinuje bázu s väčším rozmerom s penalizáciou hrubosti, zavedenou na parametri pri odhadovaní modelu [12].

Na odhadnutie modelu (9) použijeme druhú metódu, ktorá predstavuje alternatívu k prístupu s penalizáciou použitou na regresných parametroch. Odhadneme model pomocou techniky funkcionálnych hlavných komponentov. Táto metóda zahŕňa projektovanie centrovaných funkcionálnych pozorovaní $\tilde{f}_i(t), \ldots, \tilde{f}_N(t)$, ktoré získame ako $\tilde{f}_i(t) = f_i(t) - \bar{f}(t), i = 1, \ldots, N$, kde $\bar{f}(t) = \frac{1}{N}\sum_{i=1}^{N} f_i(t), t \in I$, do menej rozmerného podpriestoru spojeného s prvými K funkcionálnymi hlavnými komponentami (FPC), zn. $\xi_1(t), \ldots, \xi_K(t), t \in I$.

Keďže FPC tvoria ortonormálnu bázu v $L^2(I)$, je jednoduché používať tento systém, ktorý je rovnaký pre $f_i(t), i = 1 \ldots, N$ a $\beta_1(t)$ ako môžeme zvoliť v (12).

Potom môžeme písať

$$\tilde{f}_i(t) = \sum_{k=1}^{K} c_{ik}\xi_k(t), i = 1 \ldots, N$$

$$\beta_1(t) = \sum_{k=1}^{K} b_k\xi_k(t), \tag{18}$$

kde $c_{ik} = \langle \tilde{f}_i, \xi_k \rangle_2$ a $b_k = \langle \beta_1, \xi_k \rangle_2$ sú funkcionálne skóre (projekcie $\tilde{f}_i(t)$ a $\beta_1(t)$ v smere $\xi_k(t)$).

Uvedomme si, že počítanie FPC na centrovaných údajoch nebude mať vplyv na parameter $\beta_1(t)$, ale zmení to intercept β_0 (značený ako $\tilde{\beta}_0$).

Nasledujúci model

$$y_i = \tilde{\beta}_0 + \int_I \beta_1(t)[f_i(t) - \bar{f}(t)]dt + \epsilon_i = \beta_0 + \int_I \beta_1(t) \cdot f_i(t)dt + \epsilon_i, \quad i = 1, \ldots, N, \tag{19}$$

s

$$\beta_0 = \tilde{\beta}_0 - \int_I \beta_1(t) \cdot \bar{f}(t)dt, t \in I,$$

reprezentuje rovnaký model ako (9), keď veličina $\int_I \beta_1(t) \cdot \bar{f}(t)dt$ je zahrnutá v intercepte. Ak využijeme bázové rozšírenie (18), môžeme model (19) prepísať nasledovne

$$\mathbf{y}_{N\times 1} = \mathbf{X}_{N\times(K+1)}\tilde{\mathbf{b}} + \boldsymbol{\epsilon}_{N\times 1}, \tag{20}$$

kde $\tilde{\mathbf{b}} = (\tilde{\beta}_0, b_1, \ldots, b_K)^T$, $\mathbf{X} = [\mathbf{1}_{N\times 1}; \mathbf{C}]$ a $\mathbf{M}_\xi = \mathbf{I}$ následkom vzájomnej ortonormality $\xi_j, j = 1, \ldots, K$. Odhad $\hat{\tilde{\mathbf{b}}}$ je daný ako riešenie nasledujúcej rovnice,

$$\mathbf{X}^T\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}^T\mathbf{y},$$

a odhadované hodnoty y dosiahneme ako

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\tilde{\mathbf{b}}}.$$

Číslo K, počet funkcionálnych hlavných komponentov, ktoré z praktického hľadiska určuje hladkosť β_1, musíme fixovať. Toto môžeme dokázať zahrnutím iba niektorých prvých komponentov, ktoré reprezentujú väčšinu variability v $\tilde{f}_1, \ldots, \tilde{f}_N$, alebo krížovou validáciou. Môžeme vybrať číslo K tak, že minimalizujeme kritérium krížovej validácie, definované ako

$$CV(K) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i^{(-i,K)})^2, \quad K = 1, \ldots, K_{max}, \tag{21}$$

kde $\hat{y}_i^{(-i,K)}$ označuje predikcie i-tej hodnoty premennej y použitím K FPC okrem i-tého pozorovania $(\tilde{f}_i(t), y_i)$ v odhade modelu (19).

Na nájdenie vlastných funkcií ξ_j a príslušných vlastných čísiel λ_j kovariančného operátora V_f v rovnici $V_{\tilde{f}}g = \frac{1}{N}\sum_{i=1}^{N}\langle\tilde{f}_i, g\rangle_2\tilde{f}_i$, ktorá transformuje akýkoľvek prvok $g \in L^2(I)$ na iný prvok v $L^2(I)$, môžeme použiť metódy založené na B-splajnovom bázovom rozšírení funkcionálnych pozorovaní $\tilde{f}(t), i = 1, \ldots, N$ a vlastných funkcií $\xi_j(t), j = 1 \ldots, K$ (viď. sekcia 2.2.1.).

5.2. Funkcionálny lineárny regresný model pre skalárne vysvetľovanné premenné s prediktorom s B^2 priestoru

V tejto časti si rozoberieme problém funkcionálnej regresie s vysvetľujúcou premennou $f(t)$, tvorenú hustotou rozdelenia pravdepodobnosti. V tomto prípade budeme pracovať s metodológiou Bayesových priestorov.

Uvažujme $f_1, \ldots, f_N$ ako vzorku z funkcionálnej premennej f v $B^2(I)$ a $y_1, \ldots, y_N$ je vzorka reálnej premennej y tak, že dostaneme dvojice $(y_i, f_i), i = 1, \ldots, N$. Funkcionálny lineárny model pre i-té pozorovanie y_i v závislosti prediktoru f_i môžeme vyjadriť ako

$$y_i = \beta_0 + \langle \beta_1(t), f_i(t) \rangle_B + \epsilon_i, \quad i = 1, \ldots, N, \quad t \in I, \tag{22}$$

kde $\beta_0 \in \mathbf{R}$ a $\beta_1(t) \in B^2(I)$ sú neznáme regresné parametre a $\boldsymbol{\epsilon}$ je náhodný vektor s identicky rovnako rozdelenými náhodnými chybami s nulovou strednou hodnotou a konečným rozptylom. Ak aplikujeme clr transformáciu (3), model (22) pre $i = 1, \ldots, N$ a $t \in I$ môžeme ekvivalentne prepísať ako

$$y_i = \beta_0 + \langle clr(\beta_1)(t), clr(f_i)(t) \rangle_2 + \epsilon_i = \beta_0 + \int_I clr(\beta_1)(t) \cdot clr(f_i)(t) dt + \epsilon_i, \tag{23}$$

alebo v maticovom zápise

$$\mathbf{y}_{N\times 1} = \mathbf{1}_{N\times 1}\beta_0 + \int_I clr(\mathbf{f})(t)_{N\times 1} \cdot clr(\beta_1)(t) dt + \boldsymbol{\epsilon}_{N\times 1}, \tag{24}$$

a odhady regresných parametrov získame minimalizáciou kritéria SSE,

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \langle \beta_1(t), f_i(t) \rangle_B)^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \langle clr(\beta_1)(t), clr(f_i)(t) \rangle_2)^2.$$

Snažíme sa o minimalizáciu SSE v zmysle L^2 a ako sme mohli vidieť v predchádzajúcej časti, snažíme sa minimalizovať úlohu výpočtovými metódami, založenými na bázovom rozšírení oboch pozorovaných funkcií $clr(f_i), i = 1, \ldots, N$ a regresného parametru $clr(\beta_1)$. Uvedomme si, že v prípade bázového rozšírenia $clr(f_i)$ a $clr(\beta_1)$ v zmysle známych bázových systémov potrebujeme splniť aj obmedzenie na nulovosť integrálu. To je

$$\int_I clr(f_i)(t) dt = \int_I \sum_{k=1}^{K_f} c_{ik}\phi_k(t) dt = 0, i = 1, \ldots, N,$$

$$\int_I clr(\beta_1)(t) dt = \int_I \sum_{k=1}^{K_\beta} b_k \psi_k(t) = 0, \tag{25}$$

s bázovými koeficientami $\{c_{ik}\}$ a $\{b_k\}$, a so známimi bázovými systémami $\{\phi_1(t), \ldots, \phi_{K_f}(t)\}$ a $\{\psi_1(t), \ldots, \psi_{K_\beta}(t)\}$. Pritom je samozrejme potrebné byť opatrný s podmienkami (25) pri odhadovaní modelu (22), pretože toto obmedzenie môže spôsobiť singularitu modelu (22). Podmienka na nulovosť integrálu výsledného splajnu sa totiž premietne do podmienky na odpovedajúce B-splajnové koeficienty [16].

Preto je potrebné predefinovať B-splajnovú bázu tak, aby bola formulovaná priamo v priestore $L_0^2(I)$. Konkrétne, implementáciou vyhladzujúcich splajnov v takejto situácií sa venuje článok [10] a v regresnom kontexte potom hlavne článok [18]. Realizácia príslušného algoritmu, ktorá sa principiálne opiera o poznatky z kapitol 2.2, 4.2 a 5.1, je v komentovanej forme obsiahnutá v Prílohe a na priloženom CD nosiči.

Výsledný odhad funkcionálneho parametru β_1 v priestore $L_0^2(I)$ má nasledujúci tvar

$$clr(\beta_1)(t) = \sum_{j=1}^{K} \hat{b}_j \xi_j(t),$$

a môže byť zapísany v $B^2(I)$ priestore použitím exponenciálu, t.j.

$$\beta_1(t) = \bigoplus_{j=1}^{K} \hat{b}_j \odot \zeta_j(t),$$

kde $\zeta_j = \exp(\xi_j), j = 1, \ldots, N$. Pre väčšinu tried lineárnych regresných modelov, celkovú variabilitu y_i, ktorú môžeme vyjadriť ako celkovú sumu štvorcov $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$, môžeme rozdeliť na dve časti. Na *vysvetleny súčet štvorcov* $ESS = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$ a *reziduálny súčet štvorcov* $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. V takomto prípade vieme identifikovať percento vysvetlenej variability v údajoch, vysvetlenej pomocou vysvetľovaných premenných v modeli a to prostredníctvom **koeficientu determinácie** R^2. Tento koeficient môžeme vypočítať pomocou nasledujúceho vzorca,

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}. \tag{56}$$

Z praktického hľadiska predstavuje koeficient determinácie silný nástroj, ktorý

sa používa na hodnotenie vhodnosti modelu na analyzované údaje. Poznamenajme však, že z uvedených prístupov platí koncept rozdelenia celkovej variability len pre prístup využívajúci penalizácie založenej na SFPCA.

Kapitola 6

Praktická časť

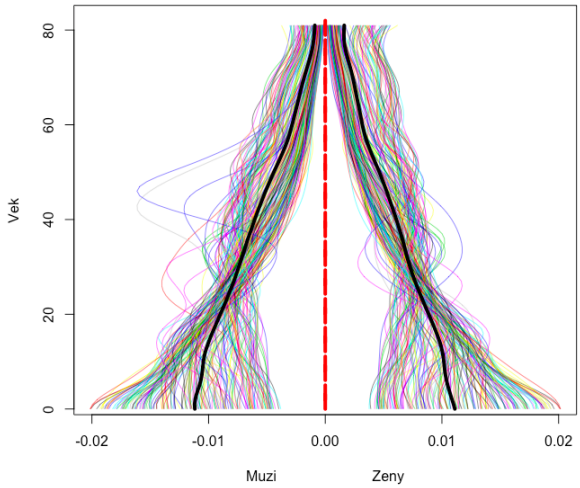
V praktickej časti diplomovej práce si uvedieme prípadovú štúdiu, v ktorej využijeme všetky doposiaľ popísané metódy. Budeme analyzovať závislosť vekovej štruktúry jednotlivých krajín sveta a výskyt rakoviny. Využívame reálne funkcionálne údaje, kde funkcie sú hustoty. Údaje pozostávajú z populačnej pyramídy 223 krajín z roku 2000.

Populačná pyramída je bežný spôsob, ako reprezentovať spoločné rozdelenie pohlavia a veku ľudí žijúcich v danej administratívnej jednotke (napríklad v meste, regióne alebo krajine). Skladá sa z páru funkcií hustôt predstavujúcich vekové rozdelenie mužov a žien vynásobených pomerom mužov a žien. Potom je súčet integrálu oboch častí populačnej pyramídy rovný 1.

Pre každú krajinu máme v databáze tabuľku relatívnej frekvencie cez pohlavie a vek, diskretizovanú v 17 intervaloch po piatich rokoch (posledný interval zodpovedá ľuďom starším ako 80 rokov). Údaje boli extrahované z Medzinárodnej databázy (IDB), ktorú vytvorilo Medzinárodné programové centrum, US Census Bureau. IDB je počítačová databanka, obsahujúca štatistické tabuľky demografických a sociálno-ekonomických údajov za 227 krajín sveta (údaje za štyri krajiny boli neúplné). IDB je k dispozícii na adrese http://www.census.gov/ipc/www/idb/region.php.

Na obrázku 1 môžeme vidieť populačnú pyramídu, ktorá reprezentuje vstupné údaje. Na y-ovej osi je vek a znázornené sú dané husoty, na ľavej strane pre mužov a na pravej pre ženy. Zároveň hrubou čiernou čiarou môžeme vidieť priemernú

hodnotu pre dané pohlavie.



Obr. 1: Populačná pyramída

Pre následné použitie sme vstupné údaje upravili. V ďalších analýzach budeme brať do úvahy len mužskú populáciu a do úvahy sme vzali len 182 krajín. V nasledujúcej tabuľke sú vstupné údaje pre populačnú pyramídu:

	Men.0-4	Men.5-9	Men.10-14	Men.15-19	Men.20-24	Men.25-29	Men.30-34	Men.35-39	Men.40-44	Men.45-49	Men.50-54	Men.55-59
1	0.016414047	0.013851561	0.012632067	0.010873153	0.009416685	0.008220106	0.006765727	0.005737437	0.004772340	0.003909601	0.003137978	0.002474548
2	0.009363974	0.009671520	0.010792529	0.010564254	0.009645341	0.007221198	0.005753093	0.005311268	0.006094039	0.005516637	0.004338135	0.003716669
3	0.011028981	0.011447483	0.011679177	0.012345564	0.010928714	0.009023977	0.007607009	0.006434295	0.005335842	0.004189672	0.003248844	0.002203034
4	0.012424295	0.013641393	0.013082343	0.009762986	0.008257629	0.006359189	0.005354647	0.005025623	0.005130445	0.004766480	0.004067668	0.003639646
5	0.005388653	0.005347719	0.005216145	0.005207374	0.005850621	0.007297926	0.009540517	0.010900107	0.010394281	0.008581495	0.007523062	0.006011432
6	0.017306206	0.014166523	0.012297470	0.010270692	0.008816529	0.007694806	0.006613884	0.005993257	0.004708524	0.003715545	0.002940763	0.002300577
7	0.007649044	0.008388237	0.009272055	0.009014945	0.008211474	0.009014945	0.009577374	0.009255986	0.008115057	0.006443837	0.004644062	0.003390648
8	0.009966196	0.010028466	0.008525086	0.007905349	0.007718539	0.008498399	0.009936544	0.010861701	0.009556992	0.006920887	0.003525679	0.001657573
9	0.009099259	0.008872695	0.008950378	0.008613545	0.008881851	0.007905076	0.006857442	0.006302445	0.006054873	0.005746702	0.005139194	0.004289932
10	0.005328250	0.006524371	0.010645029	0.010899376	0.008526353	0.006347199	0.006393564	0.007144893	0.008622927	0.006877874	0.004947421	0.002585088
11	0.006669411	0.007430332	0.007577973	0.007163442	0.006252041	0.005996508	0.007367868	0.008299144	0.008239520	0.007430332	0.005968115	0.004991411
12	0.006640977	0.007090678	0.007203259	0.007141029	0.006796481	0.007407865	0.007729279	0.007669351	0.007618416	0.006934038	0.006713245	0.005416592
13	0.005098576	0.005742914	0.005956820	0.005972904	0.006091779	0.006485222	0.008335421	0.009138243	0.008323108	0.006880770	0.006318586	0.005515323
14	0.008133593	0.009547789	0.011102114	0.010362766	0.008772331	0.007962791	0.008362740	0.007111293	0.007036997	0.005051127	0.003507266	0.001659140
15	0.009471299	0.010025655	0.009760789	0.009268324	0.008682024	0.008117020	0.008104376	0.007696429	0.006424006	0.005258062	0.004169980	0.003250269
16	0.009869637	0.010067383	0.009624968	0.008561892	0.007402228	0.008251714	0.009643859	0.010855016	0.012024735	0.011076833	0.006660299	0.003287035
17	0.011560626	0.010503320	0.012528670	0.014198509	0.010473901	0.008646259	0.007103357	0.005903585	0.005169161	0.004442099	0.003557250	0.002776019
18	0.006707712	0.007046821	0.007855911	0.007883387	0.007550062	0.008072825	0.008820456	0.008781412	0.008193574	0.007191430	0.005190035	0.003684650
19	0.004780297	0.005496285	0.007421438	0.008365071	0.007680510	0.007547181	0.006963690	0.007310596	0.008207553	0.007029716	0.005859251	0.003445330
20	0.005612815	0.005900885	0.006233550	0.005977657	0.006259050	0.006347851	0.007212625	0.007929480	0.007835540	0.007268082	0.006754797	0.005780238
21	0.015108042	0.014217545	0.013044156	0.011374948	0.009452507	0.007517139	0.006873790	0.006099643	0.004532337	0.003510280	0.002647158	0.001923962
22	0.018557507	0.015674319	0.013388306	0.011069822	0.009272757	0.007795687	0.006147865	0.004502547	0.003315887	0.002381098	0.001895037	0.001529696
23	0.006006879	0.006535335	0.006400876	0.006366479	0.005772358	0.006044403	0.006963727	0.009121326	0.009349593	0.008214509	0.006873046	0.005728580
24	0.015312562	0.013664563	0.012268978	0.010666630	0.008734509	0.007873646	0.006940104	0.005901319	0.005036348	0.004267359	0.003427219	0.002741890
25	0.012971351	0.013082421	0.012467724	0.011522683	0.009555538	0.007863345	0.006061952	0.005186206	0.004674337	0.003976396	0.003277153	0.002576016
26	0.006447300	0.005567568	0.008335864	0.007594867	0.007462337	0.007329252	0.008079280	0.009488678	0.009120197	0.008010467	0.006275723	0.004240503
27	0.013516068	0.013694546	0.013008285	0.012427101	0.010807852	0.008537159	0.006037586	0.004302201	0.003438845	0.002825358	0.002200936	0.001767184
28	0.009178847	0.009651096	0.009734525	0.009975188	0.010009468	0.008954386	0.008122831	0.007236905	0.006121347	0.005092836	0.004120561	0.003170433
29	0.010262241	0.010616191	0.009921402	0.009135988	0.009249982	0.009393043	0.008978678	0.008878934	0.009163347	0.007281318	0.004656624	0.002797394
30	0.004064589	0.004736439	0.006234680	0.007065689	0.007657239	0.007669967	0.006972845	0.006219617	0.006659042	0.006908447	0.007132554	0.006083158

Obr. 2: Ukážka dát

Uvažujeme nasledovné krajiny:

Afganistan	Albánsko	Alžírsko
Americká Samoa	Angola	Argentína
Arménsko	Austrália	Rakúsko
Azerbajdžan	Bahamy	Bahrajn
Bangladéš	Barbados	Bielorusko
Belgicko	Belize	Benin
Bhután	Bolívia	Bosna a Hercegovina
Botswana	Brazília	Brunej
Bulharsko	Burkina Faso	Burundy
Kambodža	Kamerun	Kanada
Kapverdy	Stredoafrická republika	Čad
Chile	Čína	Kolumbia
Komory	Kongo Brazzaville	Kongo
Kostarika	Pobrežie Slonoviny	Chorvátsko
Kuba	Cyprus	Česko
Dánsko	Džibutsko	Dominikánska republika
Ekvádor	Egypt	Salvádor
Rovníková Guinea	Eritrea	Estónsko
Etiópia	Fidži	Fínsko
Francúzsko	Francúzska Guayana	Francúzska Polynézia
Gabon	Gambia	Gaza Strip
Gruzínsko	Nemecko	Ghana
Grécko	Guadelup	Guam
Guatemala	Guinea	Guinea-Bissau
Guayana	Haiti	Honduras
Maďarsko	Island	India
Indonézia	Irán	Irak
Írsko	Izrael	Taliansko
Jamajka	Japonsko	Jordánsko
Kazachstan	Keňa	Kuvajt
Kirgizsko	Laos	Litva
Libanon	Lesotho	Libéria
Libya	Lotyšsko	Luxembursko
Macedónsko	Madagaskar	Malawi
Malajzia	Maldivy	Mali
Malta	Martinik	Mauritánia
Maurícius	Mexiko	Moldavsko
Mongolsko	Čierna Hora	Maroko
Mozambik	Namíbia	Nepál
Holandsko	Nová Kaledónia	Nový Zéland
Nikaragua	Niger	Nigéria

Severná Kórea	Nórsko	Omán
Pakistan	Panama	Papua Nova Guinea
Paraguaj	Peru	Filipíny
Poľsko	Portugalsko	Portoriko
Katar	Reunion	Rumunsko
Rusko	Rwanda	Svätá Lucia
Samoa	Sao Tome and Principe	Saudská Arábia
Senegal	Srbsko	Sierra Leone
Singapur	Slovensko	Slovinsko
Solmon Ostrovy	Somálsko	Južná Afrika
Južná Kórea	Španielsko	Sri Lanka
Sudán	Surinam	Švédsko
Švajčiarsko	Sýria	Tadžikistan
Tanzánia	Thajsko	Togo
Trinidad a Tobago	Tunisko	Turecko
Turkménsko	Uganda	Ukrajina
Spojené arabské emiráty	Veľká Británia	Spojené štáty americké
Uruguaj	Uzbekistan	Vanuatu
Venezuela	Vietnam	Jemen
Zambia	Zimbabwe	

Hlavným cieľom tejto časti bude funkcionálna regresia. Na začiatku si uvedieme na jednom ilustratívnom príklade postup, ktorý budeme aplikovať. Chceme zistíť závislosť medzi vekovou štruktúrou daného štátu a výskytom rakoviny prostaty. Z tohto dôvodu sme údaje obmedzili len na mužov. Premennú výskyt rakoviny sme uvažovali len pre 182 krajín. Údaje pre jednotlivé závislé premenné boli extrahované zo stránky Cancer Today, ktorá spolupracuje s asociáciou National Agency for Research on Cancer, ktorá robí výskumy pod World Health Organization, možno nájsť na adrese https://gco.iarc.fr/today/online-analysis-table. Stránka Cancer Today poskytuje súbor údajov na skúmanie úmrtnosti a prevalencie 36 špecifických typov rakoviny v 185 krajinách alebo teritóriách sveta v roku 2018, podľa pohlavia a vekovej skupiny, ako súčasť projektu GLOBOCAN.

Prvý krok, ktorý spravíme, bude použitie diskrétnej verzie clr transformácie na dané funkcie hustôt. Následne s touto transformáciou budeme pracovať v ďalšej analýze. Ukážku clr transformácie môžeme vidieť na nasledujúcom obrázku.

[0,5)	[5,10)	[10,15)	[15,20)	[20,25)	[25,30)	[30,35)	[35,40)	[40,45)
1.537602417	1.367862858	1.275703505	1.12576165	0.98194804	0.84604803	0.651334645	-0.20666966	0.302301675
0.758698077	0.791013788	0.900682440	0.87930434	0.78830330	0.49884917	0.271565927	0.19165891	0.329139384
1.072695840	1.109939276	1.129976908	1.18546620	0.37041585	0.87205454	0.701239455	0.53381167	0.346616094
1.103665316	1.197120259	1.155274944	0.86260978	0.69514899	0.43391234	0.261976268	0.19856091	0.219203886
1.683456840	0.790126140	1.341785225	1.16168608	1.00901993	0.87293723	0.721562751	0.62302668	0.381766158
0.623609833	0.598395435	0.607112617	0.56875282	0.59942683	0.48292194	0.340751337	0.25635450	-0.476866927
0.155117350	0.357639046	0.847187519	0.87080004	0.62525622	0.33010811	0.337386355	0.44849233	0.636519082
0.239866538	0.305388410	0.321141009	0.31246433	0.26301242	-0.34399746	0.391623032	0.38383944	0.377175921
-0.009967116	0.109038346	0.145608381	0.14830484	0.16801176	0.23059765	0.481585624	0.57353973	0.480107342
0.693684168	0.853991008	0.311669759	0.93590059	0.76928396	0.67246097	0.721467523	0.55936548	0.548862915
0.791600128	0.848481368	0.821707298	0.76993663	0.70458874	0.63729715	0.635738223	0.58409052	0.403375971
0.861930788	0.881768506	0.836828269	0.71978891	0.57424875	0.68288865	0.838789053	0.95709499	1.059433492
1.093690177	0.997776561	1.174104781	1.29922212	0.99497171	0.80321190	-0.086494528	0.42164496	0.288795556
0.382980575	0.432299259	0.540988908	0.54448030	0.50127844	0.56822615	0.656796235	0.65235988	0.583072854
0.027119424	0.166689155	0.466989584	-0.10646561	0.50130270	0.48379086	0.403326252	0.45194155	0.567671573

Obrázok tabuľky údajov po clr transformácií

V ďalšom kroku analýzy budeme clr transformované proporcie vyhladzovať použitím systému vyhladzovacích B-splajnov s nosičom I, ktoré budú spĺňať podmienku nulovosti integrálu. Pre tento prípad sme zmenili pre vek interval na $[0, 90]$, z ktorého pri vyhladzovaní budeme vychádzať. Jednotlivé intervaly, na ktorých budeme vyhladzovať budú nasledovné:

$$[0,5), [5,10), [10,15), [15,20), [20,25), [25,30), [30,35), [35,40), [40,45), [45,50),$$

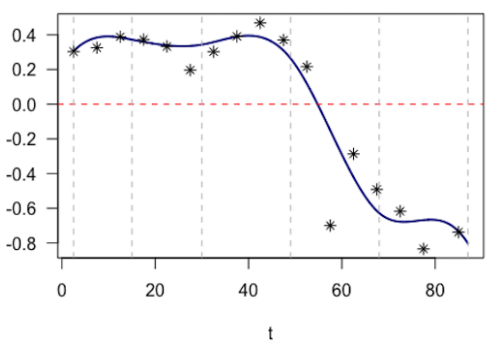
$$[50,55), [55,60), [60,65), [65,70), [70,75), [75,80), [80,90).$$

Vyhladzovanie budeme uskutočňovať pomocou kubického vyhladzovacieho splajnu ($k = 3$) na nosiči $I = [0, 90]$ a máme zvolených 6 neekvidistantných uzlov $\{\lambda_0 = 2{,}5, \lambda_1 = 15, \lambda_2 = 30, \lambda_3 = 49, \lambda_4 = 68, \lambda_5 = 87\}$. Krajné uzly sme zvolili $[2{,}5; 87]$ a nie $[0, 90]$ pretože sme chceli zmierniť odstrelenie splajnov na koncoch intervalu.

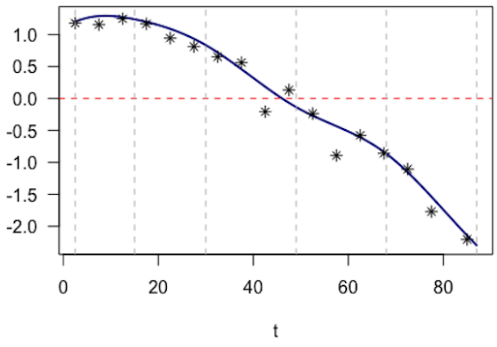
Výsledný splajn nájdeme minimalizáciou funkcionálu

$$J_l(s_k) = (1-\alpha)\int_a^b [s_k^{(2)}(x)]^2 dx + \alpha \sum_{i=1}^{n} w_i [f_i - s_k(x_i)]^2, \quad \alpha \in (0,1).$$

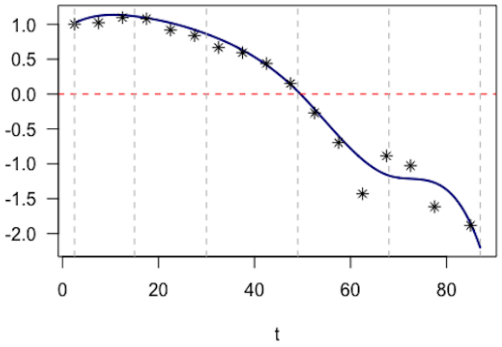
Vektor váh $\mathbf{w}_i$ sme nastavili ako vektor jednotiek a vyhladzovací parameter α na hodnotu 0,5. Na nasledujúcich niekoľkých obrázkoch si ukážeme ako sa nám podarilo vyhladiť jednotlivé pozorovanie pre daný štát. Nakoľko máme 182 krajín, tak si uvedieme len grafy pre niekoľko krajín, kde na osi x máme vek a na osi y hodnoty clr hustôt.



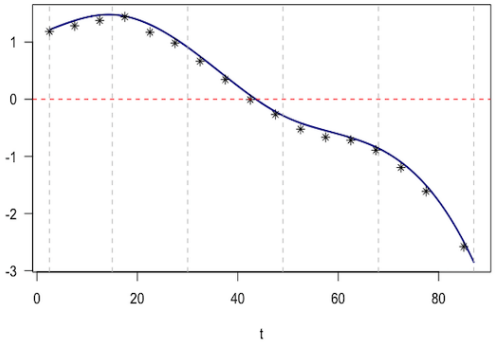
(a) Spojené štáty americké



(b) Uzbekistan



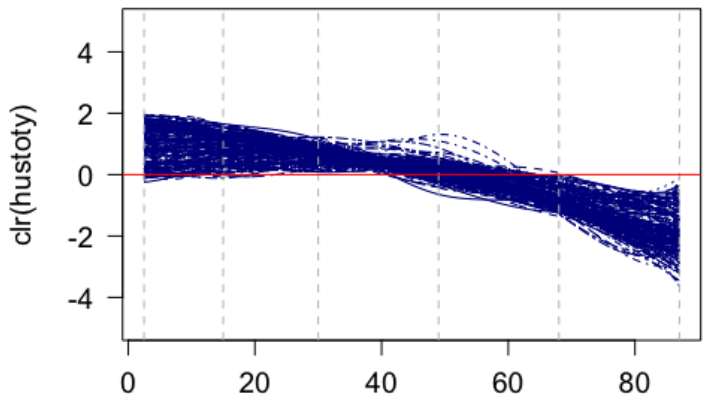
(c) Vietnam



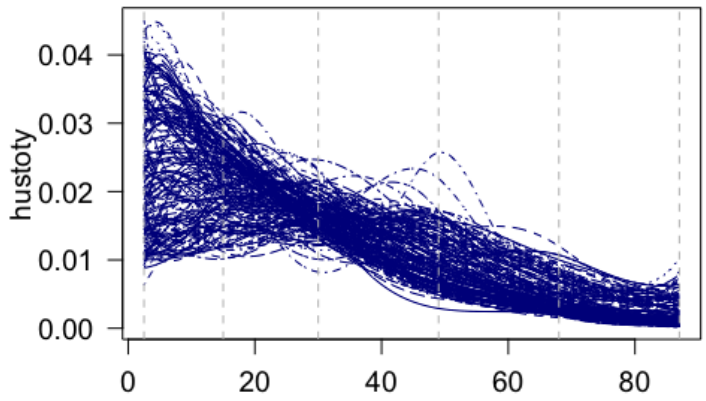
(d) Zimbabwe

Obr. 4: Vyhladzovanie clr transformovaných údajov pomocou B-splajnov.

Na obrázku 5 môžeme vidieť vyhladené hustoty pre všetky krajiny s príslušným vyhladzovacím splajnom v $L_0^2(I)$ priestore a na obrázku 6 vyhladené clr transformované hustoty v B^2 priestore. Vertikálne prerušované čiary vyznačujú pozície jednotlivých uzlov.



Obr. 5: Vyhladené funkcie hustôt pomocou vyhladzujúceho splajnu v $L_0^2(I)$ priestore.



Obr. 6: Výsledné vyhladené clr transformované funkcie hustôt v B^2 priestore.

Ako bolo spomenuté na začiatku praktickej časti, v práci budeme uvažovať reálnu závislú kompozičnú premennú a to výskyt rakoviny prostaty v danej krajine. Hodnota tejto premennej presne vyjadruje hrubú mieru (počet nových prípadov v roku 2018, ktoré sa vyskytnú v konkrétnej populácii za rok), vyjadrenú ako počet prípadov na 100000 obyvateľov populácie. Pre každú krajinu máme hod-

notu výskytu rakoviny prostaty, na ktorú použijeme logaritmickú transformáciu a budeme túto premennú značiť ako $y_1, \ldots, y_{182}$.

Teraz, pre modelovanie zavislosti početnosti vyskytu rakoviny prostaty y_i na vekovej štruktúry populácie $f_i(x)$, uvažujeme regresný model v nasledujúcom tvare

$$y_i = \beta_0 + \langle \beta_1(x), f_i(x) \rangle_B + \epsilon_i, \quad i = 1, \ldots, 182, \quad x \in I,$$

kde $\beta_0 \in \mathbf{R}$ a $\beta_1 \in B^2(I)$ sú neznáme regresné parametre, ktoré budeme chcieť odhadnúť a ϵ_i sú chyby modelu. Keďže dané hustoty sme upravili pomocou clr transformácií hustôt, tak model môžeme upraviť nasledovne:

$$y_i = \beta_0 + \int_I clr(\beta_1)(x) \cdot clr(f_i)(x) dx + \epsilon_i.$$

Ďalej sme využili prístup vyjadrenia kompozičných splajnov pomocou B-splajnovej bázy, kde ďalej budeme využívať SFPC. O správnom výbere počtu komponentov sa môžeme rozhodnúť vďaka koeficientu determinácie. Rôzne počty komponentov a príslušné koeficienty determinácie sú uvedené v nasledujúcej tabuľke:

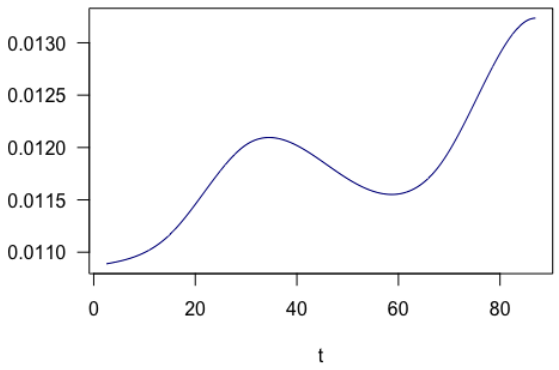
Počet komponentov	Hodnota R^2
$K = 3$	0,6280411
$K = 4$	0,6299921
$K = 5$	0,6303666
$K = 6$	0,6335434

Tabuľka 6.1: Hodnota koeficientu determinácie pre rôzny počet komponentov

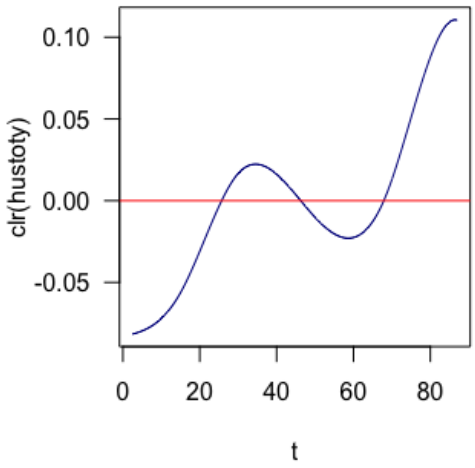
Je zrejmé, že pridávaním počtu komponentov sa nám hodnota koeficientu determinácie výrazne nezvýšila, čo znamená, že pridaním komponentov sa nám prediktívna schopnosť regresného modelu oveľa nezvýši. Z hodnoty koeficientu determinácie vyplýva najvhodnejšie $K = 3$, vyšší počet komponentov by nám mohol viesť k nežiadúcemu zahrnutiu šumu do odhadu regresného koeficientu. Ďalší spôsob na stanovenie počtu SFPC je pomocou krížovej validácie (leave-one-out cross validation). Číslo K chceme vybrať tak, že minimalizujeme kritérium krížovej validácie $CV(K)$. Krížovú validácu robíme tak, že údaje rozdelíme do disjunktných

podmnožin rovnakej veľkosti. Každá podmnožina je použitá práve raz pre testovanie modelu vzniknutého zo zostávajúcich údajov. Kritérium $CV(K)$ potom vyberáme to najmenšie. Overili sme, že minimum dostaneme pre $K = 3$. Tento počet aj budeme uvažovať, naviac prvé tri komponenty spolu zachytia skoro 94% vysvetlenej variability v údajoch.

Pretože výsledné odhady regresných parametrov sú vo forme clr transformovyných funkcií hustôt, sú následne vrátené naspäť do originálneho priestoru B^2. Z tohoto dôvodu ich aj v tomto priestore môžeme interpretovať nasledovne: vyššie hodnoty hustôt, ktoré odpovedajú jednotlivým rozdeleniam veku v danom štáte indikujú väčší vplyv rozdelenia veku na výskyt rakoviny prostaty a naopak. Budeme však uvažovať aj clr hustotu regresného parametra, pretože v tomto prípade môžeme povedať, že jeho kladné hodnoty prispievajú k rastu hodnôt výskytu rakoviny. Naopak záporné funkčné hodnoty majú opačný efekt. Vďaka tomu môžeme povedať, približne od ktorého roku začína príspevok k nárastu výskytu rakoviny (oproti jeho priemernej hodnoty, reprezentovanej absolútnym členom), samozrejme s uvážením vplyvu vekovej štruktúry populácie. Na obrázku 7 môžeme vidieť odhad regresného parametru β_1 v priestore B^2.



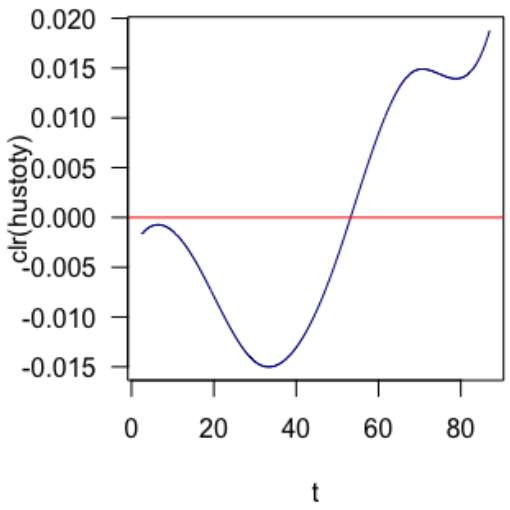
Obr. 7: Odhad regresného parametra β_1.



Obr. 8: Odhad regresného parametra β_1.

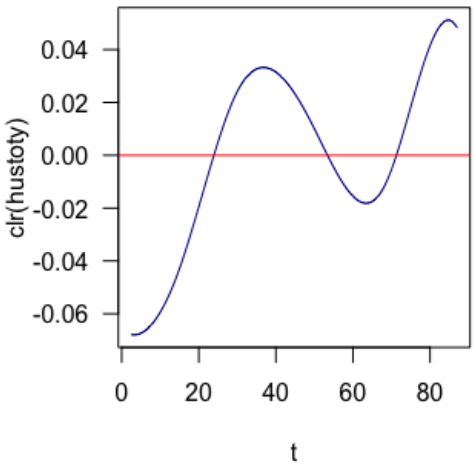
Z tohoto obrázku môžeme zhodnotiť, tak ako sa dá predpokladať, že vzhľadom k hustote populácie sa nám výskyt rakoviny prostaty takýmto spôsobom zvyšuje, ale s tým, že berieme do úvahy vekovú štruktúru obyvateľstva. Avšak je ešte potrebné si uvedomiť, že pri odhade y-ovej hodnoty vlastne berieme skalárny súčin dvoch clr hustôt, takže integrujeme potom tieto clr hustoty. Zároveň môžeme zhodnotiť, že výskyt tohoto typu rakoviny rozhodne prevláda u staršej populácie mužov. Čo je však veľmi zaujímavé, že k väčšiemu výskytu tohoto druhu rakoviny dochádza medzi 35. a 45. rokom - samozrejme relatívne vzhľadom k výskytu rakoviny u mladších vekových ročníkov. Tieto výsledky sú pravdepodobne ovplyvnené tým, že parameter bol skúmaný na vzorke 182 krajín. Nie všetky skúmané krajiny majú rovnaké podmienky pre život, ani rovnakú úroveň zdravotnej starostlivosti. Preto v nasledujúcej analýze sa budeme zameriavať zvlášť na krajiny Európy a krajiny z Afriky. Chceme zistiť, či nárast, ktorý sme evidovali z výsledkov dochádza medzi 35. a 45. rokom, nie je spôsobený krajinami z Afriky, resp. krajinami s menej rozvinutou zdravotníckou starostlivosťou.

Prvú analýzu prevedieme na európskych krajinách. Postupujeme rovnako ako pri rakovine prostaty, preto postup už nebudeme bližšie komentovať. Znova použijeme tri komponenty, ktoré vysvetľujú väčšiu časť variability pôvodných údajov. Na obrázku 9 môžeme vidieť odhad regresného parametra β_1 v priestore L^2.



Obr. 9: Odhad regresného parametra β_1.

Z obrázku možno posúdiť, že v europskych krajinách hodnoty parametra systematicky rastú približne od 35. roku a kladných hodnôt dosiahnu až od 55. roku, čo znamená že mladšia generácia mužov touto rakovinou zrejme nie je zásadnejšie zasiahnutá. Nasledujúcou analýzou sa pozrieme teda, či by tento nárast mohol byť spôsobený krajinami v Afrike, resp. všeobecnejšie krajinami s nižšiou úrovňou zdravotnej starostlivosti. Na obrázku 10 môžeme vidieť odhad regresného parametra β_1 pre africké krajiny.



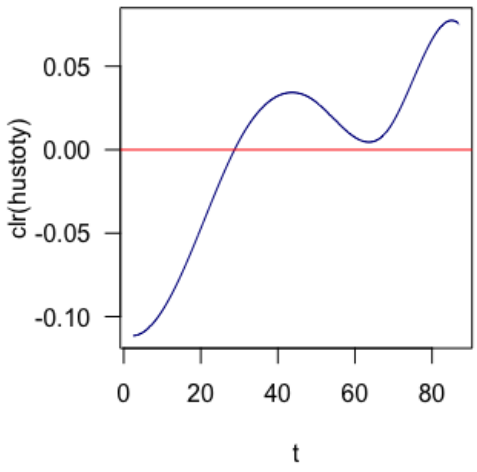
Obr. 10: Odhad regresného parametra β_1.

Obrázok 10 potvrdzuje našu teóriu, keďže môžeme na ňom vidieť efekt dominantnejšieho výskytu rakoviny (vzhľadom k vekovej štruktúre) od 20. roku

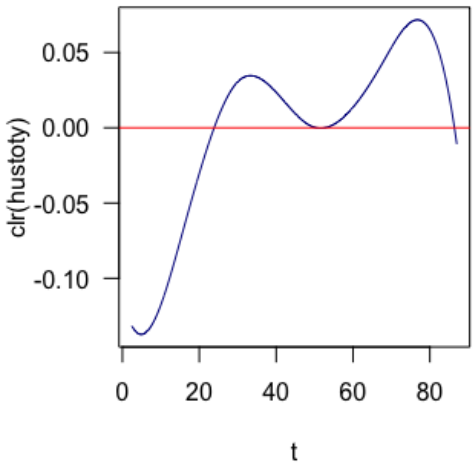
až po 50. rok. Tento výsledok môže byť spojený s faktom, že pohlavné dospievanie v Afrike je rýchlejšie ako v iných krajinách. Zároveň nám to pomáha vysvetliť nárast krivky na obrázku 8. v období od 35. roku po 45. rok.

V ďalšej analýze sa budeme zaoberať ešte rôznymi druhmi rakoviny, kde opäť budeme sledovať spomínanú závislosť. Budeme sledovať rakovinu pľúc, pečene a nakoniec hrubého čreva. Údaje pre tieto premenné boli čerpané z rovnakého zdroja ako rakovina prostaty. Navyše, v niektorých prípadoch si tieto výsledky porovnáme medzi pohlaviami a budeme uvažovať 3 aj 4 komponenty, aby sme mohli zaznamenať rozdiel.

Ďalším typom rakoviny, ktorou sa budeme zaoberať, bude rakovina pľúc u mužskej populácie. Na začiatku analýzy budeme používať prvé tri hlavné komponenty, ktoré vysvetľujú spolu 93,08% variability. Neskôr sa pozrieme ako sa zmenia výsledky, keď použijeme prvé štyri hlavné komponenty, ktoré zachytávajú spolu 96,08% vysvetlenej variability. Na nasledujúcich obrázkoch môžeme vidieť odhad regresného parametra β_1, v prípade štyroch a v prípade troch komponentov.



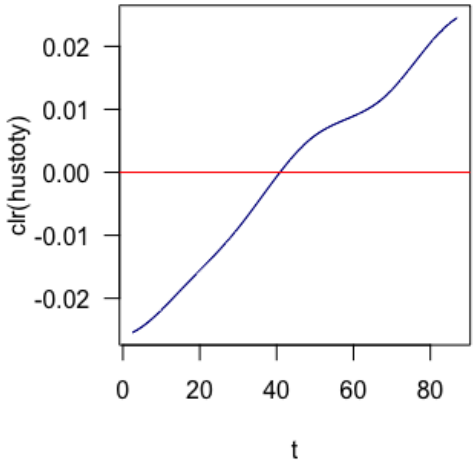
Obr. 11: Odhad regresného parametra β_1 pre tri hlavné komponenty.



Obr. 12: Odhad regresného parametra β_1 pre štyri hlavné komponenty.

Koeficienty determinácie sa v týchto dvoch prípadoch veľmi nemenia. V prípade troch hlavných komponentov je $R^2 = 71{,}14\%$ a v prípade štyroch komponentov je $R^2 = 71{,}92\%$. V tomto prípade sa javí lepšia varianta s tromi hlavnými komponentami, keďže aj graf pre odhad regresného parametra je lepšie interpretovateľnejší, v prípade štyroch hlavných komponentov sme zjavne do odhadu regresného parametra zahrnuli aj nežiadúci šum. V tomto prípade môžeme vidieť, ako aj v predchádzajúcom prípade, nárast krivky v období medzi 30-tym a 45-tym rokom. Tento nárast je zrejme spôsobený, tak ako aj pri rakovine prostaty, nevyspelými krajinami.

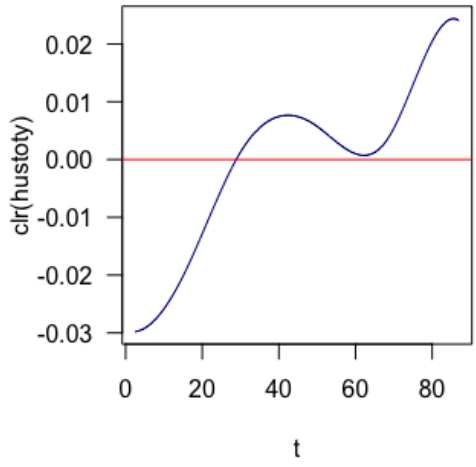
Ďalším typom rakoviny, ktorú zanalyzujeme, je rakovina pečene. V tomto prípade výsledky porovnáme aj medzi pohlaviami. Na začiatku sa pozrieme na skupinu mužov. Rozhodli sme sa analyzovať výsledky len pre tri hlavné komponenty kvôli akceptovatelnej hodnote koeficientu determinácie a súčasne dostatočnej hodnote vysvetlenej variability (93,08%). Na obrázku 13. teda uvažujeme odhad regresného parametra β_1 pre tri hlavné komponenty.



Obr. 13: Odhad regresného parametra β_1, na y-ovej osi sú hodnoty hustôt a x-ovej osi vek.

Môžeme konštatovať, že funkčné hodnoty od 40-teho roku prispievajú k rastu hodnôt výskytu rakoviny a to pomerne monotónne aj vzhľadom k vekovej štruktúre populácie. Koeficient determinácie je v tomto prípade rovný 24,36%.

Následne sa pozrieme aké sú výsledky rakoviny pečene pre skupinu žien. Na obrázku 14 môžeme vidieť graf odhadu regresného parametru β_1 pre tri hlavné komponenty.

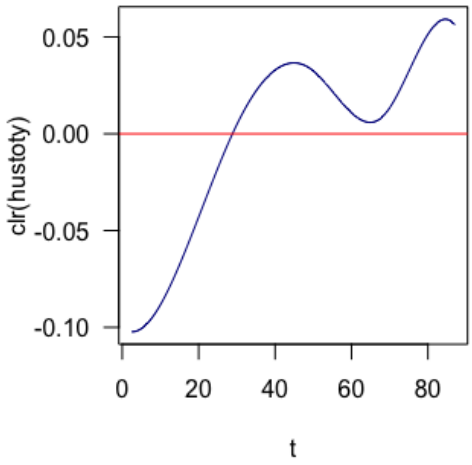


Obr. 14: Odhad regresného parametra β_1, na y-ovej osi sú hodnoty hustôt a x-ovej osi vek.

V porovnaní s výsledkom u mužov možno zhodnotiť, že závislosť rakoviny pečene na vekovej štruktúre je viac markantná u žien ako u mužov. V odhade β_1 pre ženy možno pozorovať aj pokles, u mužov pokles nenájdeme.

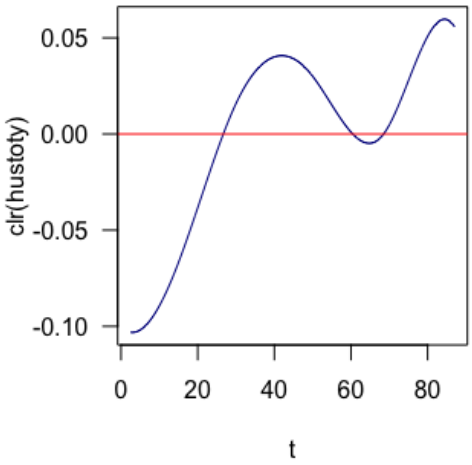
Ako posledný príklad si uvedieme rakovinu hrubého čreva. Na začiatku bu-

deme uvažovať skupinu mužov a pri tomto type rakoviny budeme uvažovať len prvé tri hlavné komponenty, keďže pri zmene na štyri komponenty zmena koeficientu determinácie nebola veľmi výrazná.



Obr. 15: Odhad regresného parametra β_1, na y-ovej osi sú hodnoty hustôt a x-ovej osi vek.

Na obrázku 15 môžeme vidieť krivku regresného parametru v preistore L^2 pre tri hlavné komponenty. Koeficient deremínácie v tomto prípade vyšiel 81,97%. Na obrázku 16 si zobrazíme ten istý druh rakoviny pre skupinu žien.



Obr. 16: Odhad regresného parametra β_1, na y-ovej osi sú hodnoty hustôt a x-ovej osi vek.

V porovnaní s mužskou populáciou možno vidieť rozdiel, avšak nie tak veľmi evidentný ako pri rakovine pečene. V tomto prípade koeficient determinácie vyšiel 78,02%.

Na záver môžeme zhodnotiť, že rakovina je rozhodne závislá od vekovej štruktúry populácie. Pri skúmaní rôznych druhov rakovín nám vždy vyšiel veľmi podobný výsledok, reflektujúci okrem všeobecne vyššieho nárastu výskytu rakoviny s vekom tiež vekovú štruktúru populácie ako takej. Všetky výsledky v tejto práci boli konzultované v Komplexnom onkologickom centre v Nemocnici Komárno. Aj podľa toho je zrejmé, že pre hlbšie závery z uskutočnenej analýzy by bolo treba skúmať aj ďalšie súvislosti výskytu jednotlivých druhov rakovín.

Záver

V diplomovej práci sme sa zaoberali regresnou analýzou s hustotou ako vysvetľujúcou premennou. Ukázalo sa, že aby sa táto regresia dala nejako relevantne vyriešiť, je potrebné využiť metodiku Bayesových priestorov. Dôležitú úlohu pri tejto analýze tiež hrala vhodná voľba aproximácie funkcií pomocou splajnov. To bolo následne demonštrované na reálnom príklade. Výsledky nám ukázali, že závislosť medzi vekovou štruktúrou populácie a niektorými typmi rakoviny naozaj existuje.

Pri štatistickom spracovaní funkcionálnych údajov som bola veľmi prekvapená rôznou metodikou a geometrickými vlastnosťami takýchto údajov, ktoré neboli vždy ľahko pochopiteľné. Avšak toto štatistické odvetvie mi príde veľmi zaujímavé a hlavne veľmi prínosné pre moje štúdium, keďže som počas tejto práce zistila, v akých rôznych odvetviach sa dá funkcionálna analýza využiť. Najťažšia časť z celej práce sa veľmi ťažko určuje, keďže s problematikou funkcionálnej analýzy som sa stretla prvýkrát. Problematika funkcionálnej analýzy bola pre mňa zo začiatku zložitá a mala som problém správne pochopiť niektoré tvrdenia. Na tomto mieste by som chcela ešte raz poďakovať svojmu vedúcemu, ktorý mi pomáhal previesť touto novou problematikou a pomohol mi pri správnom pochopení jednotlivých častí.

Za prínos považujem asi všetko čo sa tohoto odvetvia týka keďže si myslím že v budúcnosti využijem skoro všetko z tejto problematiky. Verím, že moja práca bude prínosná aj pre ostatných záujemcov o danú problematiku. Keďže k funkcionálnej analýze existuje veľmi veľa literatúry v angličtine a málo v českom alebo slovenskom jazyku, verím že veľa záujemcov ocení aj takýto rýchly

prehľad cez štatistické spracovanie hustôt a pomôže im zorientovať sa v danej problematike.

Literatúra

[1] Bělašková, Z., *Vyhlazující splajny v R*, 86 s. Diplomová práca. Univerzita Palackého, Katedra matematickej analýzy a aplikácií matematiky. Vedúci práce Jitka Machalová, 2015.

[2] Delicado, P., *Dimensionality reduction when data are density functions*, Comput. Statist. Data Anal. 55, 401–420, 2011.

[3] Egozcue, J. J., Díaz-Barrero, J. L., Pawlowsky-Glahn, V., *Hilbert space of probability density functions based on atchison geometry*, Acta Mathematica Sinica, 22(4), 1175-1182, 2006.

[4] Everitt , B., Hothorn , T., *An introduction to Applied Multivariate Analysis with R*, Springer,Heidelberg, 2011.

[5] Faraway, J., *Regression analysis for a functional response*, Technometrics 3, 254–261. `http://dx.doi.org/10.2307/1271130`, 1997.

[6] Filzmoser, P., Hron, K., Templ, M, *Applied compositional data analysis*, Springer, Cham, 2018.

[7] Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P., *Simplicial principal component analysis for density functions in Bayes spaces*, MOX-report 25/2014, Politecnico di Milano, 2014.

[8] Kobza, J., *Splajny*, vyd. 1, Olomouc, Vydavateľstvo Univerzity Palackého v Olomouci, 1993.

[9] Machalová, J., Hron, K., Monti. G.S., *Preprocessing of centred logratio transformed density functions using smoothing splines*, Journal of Applied Statistics, 2015.

[10] Machalová, J., Talska, R., Hron, K., Gaba, A., *Compositional splines for representation of density functions*, arXiv:1905.06858.

[11] Menafoglio, A., Guadagnini, A., Secchi, P., *A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers*, Stoch. Environ. Res. Risk Assess. 28 (7), 2014.

[12] Ramsay, J., Silverman, B., *Functional Data Analysis*, second ed. Springer, New York, 2005.

[13] Ramsay, J., O., Hooker, G., Graves, S., *Functional Data Analysis with R and MATLAB*, Springer, New York, 2009.

[14] Smělá, A., *Interpolace splajny*, 96 s. Bakalárska práca. Univerzita Palackého, Katedra matematické analýzy a aplikacií matematiky. Vedúci práce Jitka Machalová, 2011.

[15] Shena, Q., Xub, H., *Diagnostics for linear models with functional responses*, Technometrics 1, 26–33, 2007.

[16] Talská, R., Menafoglio, A., Machalová, J., Hron, K., Fišerová, E., *Compositional regression with functional response*, Comput. Stat. Data Anal. 123, C (July 2018), 66–85. `https://doi.org/10.1016/j.csda.2018.01.018`, 2018.

[17] Talská, R., *Bayesovy prostory*, 76 s. Diplomová práca. Univerzita Palackého, Katedra matematickej analýzy a aplikácií matematiky. Vedúcí práce Karel Hron, 2015.

[18] Talská, R., Hron, K., Matys-Grygar, T, *Compositional scalar-on-function regression with application to particle size distributions*, rukopis článku, 2020.

[19] *The R Project for Statistical Computing [online]*, dostupné z: `http://www.r-project.org`.

Prílohy

Ukážka kódu - odhad parametru v regresii s hustotou ako vysvetľujúcou premennou

```
# Použité knižnice:
library(robCompositions)
library(compositions)
library(fda)
library(scatterplot3d)
# Načítanie populačnej pyramídy:
data <- read.csv2("data.csv",header=FALSE)
# Numerická integrácia:
trapzc <- function(passo,y)
{
int<-passo*(0.5*y[1]+sum(y[2:(length(y)-1)]) +0.5*y[length(y)])
return (int)
}
# Spätná transformácia z L^2 do B^2:
clr2density <- function(z, z_step, clr)
{
if(is.fd(clr))
return(exp(eval.fd(z,clr))/trapzc(z_step,exp(eval.fd(z,clr))))
if(!is.fd(clr))
return(exp(clr)/trapzc(z_step,exp(clr)))
}
# Názvy krajín:
states = read.csv2("data.csv",header=FALSE)[,1]
# Vek v intervale [0,90]:
age_parnition = c(seq(0,80,by=5),90) #čísla od 0 po 90 po 5
age_cut = c() #zostavenie vekových intervalov
for(i in 1:(length(age_parnition)-1)) age_cut[i] = paste("[",
age_parnition[i],",",age_parnition[i+1], ")",sep=)
n_class = length(age_parnition)-1
```

```
t_mid = c()
for (i in 1:n_class){t_mid[i]=(age_parnition[i+1]+
age_parnition[i])/2}
t_mid #stredy intervalov
# DĺŽky podintervalov
dlzky = diff(age_parnition)
# Rozdelenie populačnej pyramídy podľa pohlavia
muzi = as.matrix(data[,c(18:2)])
# Konštantný súčet
muzi_norm = constSum(muzi)
# Hustoty
dens_men = muzi_norm/dlzky
# Clr transformácia pre hustoty:
l.dens_men = cenLR(dens_men)$x.clr
# Nastavenia pre vyhladzovanie
knots=c(2.5,15,seq(30,87,l=4)) #uzly
w = rep(1,17) #váhy
k = 4
der = 2
alfa = 0.5 #vyhladzovací parameter
ch = 1
# Vyhladenie splajnami pre jednotlivé štáty
z_coef = NULL
J = c()
for (i in 1:length(states)){
spline = SmoothingSpline0(knots=knots, t=t_mid, f=l.dens_men[i,],
w=w, k=k, der=der, alfa=alfa, ch=ch)
abline(v=knots,col="gray",lty=2)
z_coef = cbind(z_coef,spline[[2]]) #z-koeficienty
 J[i] = spline[[1]]
}
sum(J) #hodnota funkcionálu ktorú chceme minimalizovať
# ZB-splajnová báza
Z = ZsplineBasis(knots=knots,k)
Zco= Z$C0
# Clr hustoty
data.l=Zco%*%(z_coef)
t.fine = seq(knots[1],max(knots),l=1000) #sekvencia bodov
od prvého po posledný uzol
t.step=diff(t.fine[1:2]) #krok pri sekvencii bodov
# Graf clr hustôt
N=182
matplot(t.fine,data.l,
```

```
lty=1:N, type="l",las=1,cex.lab=1.2,cex.axis=1.2,
ylim=c(-5,5),ylab="clr(hustoty)",xlab=, col="darkblue")
# Vyjadrenie kompozičných splajnov pomocou B-splajnovej báze:
bazeB = create.bspline.basis(range(knots),nbasis = dim(z_coef)[1]+1,
norder=k,breaks=knots)
b_coef =Z$D%%Z$K%%(z_coef) #b-koeficienty
data.fd = fd(b_coef, bazeB) #definujeme si funkcionálny data objekt
# Využijeme SFPCA a skóre na zredukovanie veľkého počtu údajov
ncom=3 #nastavenie počtu komponentov
pcafd = pca.fd(data.fd,ncom, centerfns = T) #funkcionálna PCA
score=pcafd$scores #skóre
mean=pcafd$meanfd #stredné hodnoty
expVar = pcafd$values[1:5]/sum(pcafd$values)*100
#vysvetlená variabilita 5 komponentami
# REGRESIA
# Načítanie vysvetľovanej premennej výskyt rakoviny prostaty:
druhapremenna = read.csv2("druhaprem.csv",header=FALSE,
dec = ",",sep = )
druhapremenna1 = druhapremenna[,-1]
druhapremenna_names=as.character(read.csv2("druhaprem.csv",
header=FALSE,dec = ",",sep = )[,1])
# Logaritmická transformácia
l.response=log(druhapremenna1)
# Matica plánu vytvorená zo skórov SFPCA
X = as.matrix(cbind(rep(1,n),score))
# Metóda najmenších štvorcov na odhad regresných parametrov
MNC = solve(t(X)%%X)%%t(X)%%as.vector(l.response)
# Odhad regresného parametru β_0
BETA0=MNC[1,]
# Odhad regresného parametru β_1
BETA1=MNC[-1,]
# Odhady β_1(t)
odhad.l = 0
odhad.b = c()
for (j in 1:ncom)
{
odhad.l = odhad.l + BETA1[j]*eval.fd(t.fine, pcafd$harmonics[j,])
}
odhad.b = clr2density(t.fine, t.step, odhad.l)
# Graf pre odhady β_1 v L^2: y osa clr hustoty
matplot(t.fine, (odhad.l),type="l",col="darkblue",lty=1,las=1,
ylab="clr(hustoty)",xlab="t")
abline(h=0,col="red")
```

```
# Graf pre odhady β_1 v B^2: y osa hustoty
matplot(t.fine, (odhad.b),type="l",col="darkblue",lty=1,las=1,
xlab="t")
# Koeficient determinácie na vyjadrenie lineárneho vzťahu:
haty=X%%MNC #hat matica
rezidua=l.response-haty #reziduá
SS_tot=l.response-mean(l.response) #(y_i- str.hodnota)
SS_reg=haty-mean(l.response) #(odhady hodnot y_i- str. hodnota)
SS_TOT=sum(SS_tot^2) #celková suma štvorcov
SS_rez=sum(rezidua^2) #reziduálny súčet štvorcov
SS_REG=sum(SS_reg^2) #vysvetlený súčet štvorcov
R2=1-(SS_rez/SS_TOT) #koeficient determinácie
```

Ukážka druhej premennej - typy rakoviny muži

Krajiny	Prostata	Pľúca	Pečeň	Hrubé črevo
Afganistan	1.60	4	2.1	2.6
Albania	44.50	64.7	17.9	15.2
Algeria	12.20	15.4	1.5	13.7
American.Samoa	38.20	33.3	13.7	23.5
Angola	13.40	1.6	2.5	2.5
Argentina	53.00	32.5	6.2	39
Armenia	46.10	82.2	19.1	26.6
Australia	148.10	58.7	14	78.1
Austria	130.40	68.5	17.5	58.9
Azerbaijan	5.80	24.8	3.9	13.5
Bahamas	102.30	12.8	5.1	30.7
Bahrain	3.90	5.8	2	7.3
Bangladesh	2.70	11	2.8	3.8
Barbados	220.40	22.6	5.1	86.1
Belarus	76.60	82.5	7.2	63.7
Belgium	132.60	109.4	11.2	92.7
Belize	36.30	10	7.4	5.8
Benin	22.90	0.54	4.1	4.8
Bhutan	0.69	5.5	6.5	5.3
Bolivia	29.00	10	5.4	5.5
Bosnia.and.Herzegovina	54.50	111.9	17.1	60.8
Botswana	6.20	3.5	3.5	3
Brazil	82.00	18.5	7.1	23.9
Brunei	15.20	25.1	14.3	37.6
Bulgaria	124.40	98.2	11	79.6
Burkina.Faso	7.00	1.7	8	2.5
Burundi	13.70	1.1	4.9	3.5
Cambodia	2.40	12.8	21.3	8
Cameroon	17.90	1.4	5.9	3.6
Canada	116.80	68.1	14.2	71.1
Cape.Verde	33.00	9.4	8.7	3.3
Central.African.Republic	14.10	1.5	4.5	3.3
Chad	7.50	0.81	3.6	2.8
Chile	72.90	24	9.7	33.3
China	13.60	70.8	40	41.5
Colombia	52.30	13.4	4.9	18.1
Comoros	12.60	0.24	4.3	1.9
Congo..Brazzaville.	18.70	1.6	5.4	3
Congo..Kinshasa.	13.60	1.5	5.8	3.9
Costa.Rica	73.90	12.5	10	22.6

Cote.d.Ivoire	19.70	1.6	5.6	2.6
Croatia	116.80	105.7	20.4	97.2
Cuba	91.70	72.2	8.7	31.5
Cyprus	126.20	66.4	8.6	56.8
Czech.Republic	176.50	80.5	13	86.9
Denmark	163.00	84.7	14.9	103.6
Djibouti	4.50	2.1	2.3	4.1
Dominican.Republic	58.10	14.2	7.4	13.5
Ecuador	39.40	7.2	5.7	10.7
Egypt	6.20	8.5	36.8	5.3
El.Salvador	47.60	7.1	6.9	8.9
Equatorial.Guinea	13.30	2.7	4.7	3.3
Eritrea	3.60	1.6	1.9	3.7
Estonia	203.80	95.1	10	67.9
Ethiopia	3.20	2	1.4	4.1
Fiji	42.60	7.3	11.5	9.3
Finland	170.40	59.7	12.7	66.8
France	202.50	100.4	26.1	80.5
French.Guiana	73.20	22.1	12.4	22.1
French.Polynesia	83.20	64.6	15.8	18.6
Gabon	18.10	6.9	3.3	4.4
Gambia.The.Republic	3.10	2.7	22.9	0.84
Gaza.Strip.and.West.Bank	7.50	12.1	1.2	10.2
Georgia	33.50	55.8	12.6	17.7
Germany	154.50	97.9	15.3	76.5
Ghana	14.50	0.99	13.7	4.5
Greece	117.80	143.4	20.4	75.8
Guadeloupe	384.10	26	10.6	42.8
Guam	53.60	68	29.8	25
Guatemala	25.20	2.6	10.1	4.3
Guinea	13.80	2.1	14.8	1.2
Guinea.Bissau	8.30	1.3	8.9	3.2
Guyana	29.60	3.5	2.5	3.8
Haiti	36.40	4	6.4	7.4
Honduras	16.10	4.6	6.4	6.1
Hungary	119.50	140.3	16.7	132.6
Iceland	104.40	47.8	7.1	54.3
India	3.70	6.9	2.7	5.2
Indonesia	8.50	16.7	10.6	14.2
Iran	14.60	11.4	4.8	13.7
Iraq	2.80	7.9	1.4	3.8
Ireland	208.80	68.6	11.5	73.5
Israel	69.70	36.9	4.8	31.8

Italy	151.60	93.1	30.2	93.1
Jamaica	90.80	27.3	3.8	26.4
Japan	113.80	127.9	37.8	133.1
Jordan	7.90	19	2	9.2
Kazakhstan	10.60	39.6	7.3	15.6
Kenya	11.30	1.5	3	4.5
Kuwait	9.20	4.6	3.7	8.1
Kyrgyzstan	4.30	17.6	8.8	5.7
Laos	2.00	18.1	20.5	9.4
Latvia	155.70	98	7.1	82.9
Lebanon	49.20	37.1	4.4	25.4
Lesotho	12.60	3.2	3.7	2.5
Liberia	15.30	1.8	10.1	1.8
Libya	9.70	18.2	3	9.5
Lithuania	117.20	95.6	12.1	67.8
Luxembourg	134.10	72.1	16.2	57.3
Macedonia	67.70	87.2	11.6	49.6
Madagascar	13.40	0.7	4.5	2.8
Malawi	5.50	0.87	1.8	1.8
Malaysia	10.90	20.7	8.8	20.2
Maldives	5.90	12.7	8.7	11.9
Mali	5.60	1.9	4.6	4
Malta	135.00	65.9	6.9	82.9
Martinique	329.60	30.3	14.3	63.4
Mauritiania	8.80	1.9	9.7	3.1
Mauritius	28.20	25	5.7	24.9
Mexico	38.50	7	5.3	12
Moldova	43.00	68.3	27.8	64.1
Mongolia	2.20	23.8	84.2	4.3
Montenegro	63.40	101.5	10.6	37
Morocco	22.30	32.3	1.4	12.3
Mozambique	11.10	0.86	4.5	1.9
Namibia	16.40	2.9	2.5	4
Nepal	0.87	11.8	1.2	3.6
Netherlands	149.40	76.5	7.3	100.1
New.Caledonia	120.70	78.8	17.7	41.9
New.Zealand	163.70	47.4	16.5	77.1
Nicaraqua	34.40	6	10.4	8.1
Niger	1.50	0.33	5.2	2.7
Nigeria	13.20	0.78	3.1	3.6
Korea.North	3.50	54.7	30.6	26.1
Norway	202.60	62.2	9.2	93.6
Oman	4.50	3	2.8	8.2

Pakistan	4.40	7.7	2.6	3.5
Panama	65.70	11.6	5.5	20.2
Papua.New.Guinea	15.10	8.7	9.9	10.3
Paraguay	38.00	17	3.3	13.3
Peru	46.70	9.6	6.8	13.9
Philippines	13.60	22.6	12.8	16.4
Poland	83.70	97.8	8.4	76.2
Portugal	135.70	82.1	20.2	125.3
Puerto.Rico	159.70	26.2	14.3	55.7
Qatar	3.60	3	1.8	5
Reunion	95.90	51.7	13.6	43.5
Romania	63.60	90.4	23.5	68.5
Russia	59.90	72.3	8.9	49.5
Rwanda	11.50	1.8	7.9	5.8
Saint.Lucia	91.00	18.2	3.4	18.2
Samoa	38.20	33.3	13.7	23.5
Sao.Tome.and.Principe	12.50	9.6	6.7	2.9
Saudi.Arabia	3.20	3.6	3.5	12.5
Senegal	12.00	1.6	8.4	2.8
Serbia	74.30	125.4	10.7	88.2
Sierra.Leone	10.20	1.5	6.5	2.7
Singapore	117.80	75.4	34.9	69.7
Slovakia	89.20	91.4	12.6	102.8
Slovenia	170.10	96.7	20.9	125.8
Solomon.Islands	13.60	5.7	10.1	3.8
Somalia	5.10	1.7	1.7	3.9
South.Africa	44.20	19.6	5.5	12.5
Korea.South	64.80	75.2	47.6	102.1
Spain	139.40	89.8	21.9	99.9
Sri.Lanka	5.30	10.6	4.7	7.3
Sudan	4.50	1.6	2.8	3.6
Suriname	62.10	25.6	8.8	17.9
Sweden	211.60	39.4	12.8	67.4
Switzerland	160.10	57.4	15.7	61.3
Syria	12.30	16.8	2.2	9.4
Tajikistan	1.20	4.6	3.6	3.6
Tanzania	14.90	0.38	4	3.4
Thailand	19.20	45.9	48.3	26.7
Togo	7.80	1.4	5.2	3.8
Trinidad.and.Tobago	89.60	28.7	5.3	28.3
Tunisia	14.20	30	2.8	14.7
Turkey	42.90	72.8	6.9	28.6
Turkmenistan	3.30	11.4	6.1	4.7

Uganda	9.50	1.4	5.3	3
Ukraine	54.60	66.7	4.9	55
United.Arab.Emirates	3.70	2.1	0.9	6.4
United.Kingdom	171.60	82.2	14.7	80.8
United.States	131.50	75	16.9	50
Uruguay	90.10	69.1	6.8	68.7
Uzbekistan	2.80	10.3	4.9	5
Vanuatu	15.40	8.4	15.4	2.8
Venezuela	46.20	18.3	4	13.8
Vietnam	8.30	35	41	15.9
Yemen	0.81	1.8	2.8	4.4
Zambia	14.10	1.3	1.2	1.9
Zimbabwe	15.80	2.2	3.9	4.6

Ukážka druhej premennej - typy rakoviny ženy

Krajiny	Hrubé črevo	Pečeň
Afganistan	1.8	1.4
Albania	12.2	11.7
Algeria	12.6	1.2
American.Samoa	13.6	2.1
Angola	2.3	1.3
Argentina	31.4	4.3
Armenia	40.1	13.6
Australia	65.5	5.7
Austria	42.4	8.3
Azerbaijan	9	3.9
Bahamas	27	2
Bahrain	8.4	1.7
Bangladesh	2.9	0.94
Barbados	67.6	4.7
Belarus	56.9	4.7
Belgium	70.1	6.3
Belize	6.8	2.6
Benin	2.8	1.1
Bhutan	1.8	1.8
Bolivia	5.2	6.7
Bosnia.and.Herzegovina	43.3	14.4
Botswana	1.5	1.9
Brazil	25.2	4.8
Brunei	26.6	2.8
Bulgaria	52	5
Burkina.Faso	3.4	5.1
Burundi	4	2.3
Cambodia	8.1	10.3
Cameroon	3.4	1.9
Canada	62.2	6.9
Cape.Verde	7.6	7.2
Central.African.Republic	3	1.9
Chad	2.6	1.6
Chile	31.7	7.7
China	31.5	14.5
Colombia	18.9	4.3
Comoros	2.4	2.7
Congo..Brazzaville.	2.6	2.3
Congo..Kinshasa.	4.6	2.8
Costa.Rica	22.9	7.2

Cote.d.Ivoire	3.3	3.4
Croatia	66.6	10.6
Cuba	42.3	5.9
Cyprus	29.1	4
Czech.Republic	61.1	7.1
Denmark	90.5	6.4
Djibouti	4.3	1.7
Dominican.Republic	13.9	5.8
Ecuador	13.3	5.9
Egypt	5.6	14.1
El.Salvador	12.2	9
Equatorial.Guinea	2.7	1.7
Eritrea	3.6	1.6
Estonia	75.8	6.1
Ethiopia	4.7	1.6
Fiji	9.6	5.1
Finland	57.5	7.2
France	64	6.8
French.Guiana	15.2	2.1
French.Polynesia	14.2	7.8
Gabon	3.5	2
Gambia.The.Republic	0.18	6.8
Gaza.Strip.and.West.Bank	9.6	0.64
Georgia	15.4	7.1
Germany	64.7	6.4
Ghana	3.9	5
Greece	55.9	9.2
Guadeloupe	39	5
Guam	20.8	7.3
Guatemala	5	10.7
Guinea	1	9.4
Guinea.Bissau	3.2	4.3
Guyana	5.4	2.6
Haiti	9.7	5.5
Honduras	6.1	2.1
Hungary	92.4	6.2
Iceland	45.2	1.8
India	3.1	1.4
Indonesia	8.2	3.2
Iran	10.4	3.7
Iraq	3.2	1.3
Ireland	50.2	3.6
Israel	27.8	3.2

Italy	73.7	11.8
Jamaica	39.3	3.5
Japan	100.6	18.6
Jordan	13.2	1.8
Kazakhstan	17.5	4.9
Kenya	4.6	2.3
Kuwait	8.8	1.8
Kyrgyzstan	5.8	6.1
Laos	9.3	9.3
Latvia	78.1	8.5
Lebanon	22.7	3.1
Lesotho	1.7	2.3
Liberia	2.5	6.9
Libya	11.4	2.9
Lithuania	60.2	5.5
Luxembourg	52.1	7.5
Macedonia	45.9	6.1
Madagascar	3.5	2.4
Malawi	1.9	1.3
Malaysia	18	3.1
Maldives	5.7	2.6
Mali	5.6	1.7
Malta	56.7	3.3
Martinique	49	6.7
Mauritiania	4	4.4
Mauritius	20.6	4.2
Mexico	10.8	5.8
Moldova	44.2	14.9
Mongolia	5	59.7
Montenegro	29.8	6
Morocco	10.5	1
Mozambique	1.8	3.4
Namibia	3.7	1.3
Nepal	6	0.7
Netherlands	74.7	4.4
New.Caledonia	30.2	9.4
New.Zealand	65.8	5.4
Nicaraqua	12	7.6
Niger	2.5	1.7
Nigeria	3.2	2.1
Korea.North	25.2	14.4
Norway	88.9	3.2
Oman	6.5	1.7

Pakistan	3	1.7
Panama	17.3	6.5
Papua.New.Guinea	5.9	7.4
Paraguay	12.6	1.9
Peru	14.5	7.4
Philippines	13	5.3
Poland	53.2	5.2
Portugal	76.9	7.4
Puerto.Rico	44.4	5.3
Qatar	6.2	0.74
Reunion	35.8	5.9
Romania	45.3	12.1
Russia	49.7	5.7
Rwanda	7.6	4
Saint.Lucia	21.8	1.1
Samoa	13.6	2.1
Sao.Tome.and.Principe	1.9	0.95
Saudi.Arabia	8.1	1.6
Senegal	3.7	4.9
Serbia	53	7
Sierra.Leone	2.9	4.1
Singapore	75.4	12.9
Slovakia	67.8	6.3
Slovenia	65.5	7.3
Solomon.Islands	4.2	2.9
Somalia	4	1.6
South.Africa	11.7	3.2
Korea.South	63.4	16.9
Spain	61	7
Sri.Lanka	6.5	2.7
Sudan	3.1	1.7
Suriname	15.2	3.2
Sweden	61.2	6.7
Switzerland	48.3	6.5
Syria	11.3	2
Tajikistan	2.2	2.4
Tanzania	5.3	1.3
Thailand	24	19.7
Togo	3.7	2.6
Trinidad.and.Tobago	27.3	4.3
Tunisia	13.7	3.3
Turkey	20.4	3.8
Turkmenistan	4.8	3.6

Uganda	3.1	2.9
Ukraine	46.9	3.3
United.Arab.Emirates	8.4	1.3
United.Kingdom	63.3	8.3
United.States	45	6.5
Uruguay	62.5	2.5
Uzbekistan	5.5	4.1
Vanuatu	4.3	4.3
Venezuela	13.9	3.4
Vietnam	14.6	11.8
Yemen	4	1.5
Zambia	2.2	0.97
Zimbabwe	4.7	3