UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Bayesovy prostory

Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Bc. Renáta Talská**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: Prezenční
Rok odevzdání: 2015

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Renáta Talská

Název práce: Bayesovy prostory

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2015

Abstrakt: Pravděpodobnostní hustoty představují speciální případ funkcionálních dat mající relativní charakter, známý z kompozičních dat v mnohorozměrné statistice. Cílem této diplomové práce je popsat speciální Hilbertovy prostory, zvané Bayesovy, které umožňují zachytit geometrické vlastnosti hustot. Následně je metodika těchto prostorů využita při statistickém zpracovaní hustot pomocí funkcionální metody hlavních komponent. Diplomová práce též zahrnuje aplikaci na reálná data.

Klíčová slova: Hilbertův prostor, funkcionální analýza dat, bázové splajny, Aitchisonova geometrie, Bayesovy prostory, clr transformace, metoda hlavních komponent, funkcionální metoda hlavních komponent

Počet stran: 80

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Renáta Talská

Title: Bayes spaces

Type of thesis: Master's

Department:
Department of Mathematical Analysis and Application of Mathematics

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

The year of presentation: 2015

Abstract: Propability densities can be considered as a special case of functional data carring relative information, known from compositional data in multivariate statistics. The goal of this thesis is to descibe Hilbert spaces, called Bayes spaces, which enable to capture special geometry of densities. Futhermore, methodology of Bayes spaces will be used for statistical analysis of densities in case of functional principal component analysis. The thesis involves also aplication to real-world data.

Key words: Hilbert space, functional data analysis, basis spline, Aitchoson geometry, Bayes spaces, clr transformation, principal component method, functional principal component analysis

Number of pages: 80

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana doc. RNDr. Karel Hron, Ph.D. s použitím uvedené literatury.

V Olomouci dne 23. března 2015

Obsah

Poděkování

Ráda bych poděkovala především svému vedoucímu diplomové práce panu doc. RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci, cenné rady a čas, který mi věnoval při konzultacích. Za totéž bych ráda poděkovala paní RNDr. Jitce Machalové, Ph.D., která mi byla nápomocná v oblasti numerické matematiky. Nemohu opomenout ani svoji rodinu a blízké, kterým patří velký dík za plnou podporu během mého studia.

Úvod

V mé diplomové práci se budu zabývat problematikou Bayesových prostorů, Hilbertových prostorů, které slouží pro popis speciálního případu funkcionálních dat, a to pravděpodobnostních hustot. Bayesovy prostory představují jiný pohled na hustoty, jakožto na funkcionální data nesoucí pouze relativní informaci. Respektování relativního charakteru hustot je klíčovým bodem pro práci s tímto typem dat, který následně umožňuje jejich statistické zpracování.

Práce je dělena do šesti kapitol, které jsou dále členěny na podkapitoly. V první kapitole si přiblížíme základní poznatky z oblasti funkcionální analýzy dat, budou objasněny pojmy vedoucí k zavedení Hilbertových prostorů, které následně využijeme v kapitole třetí při zavedení výběrového prostoru a geometrické reprezentace hustot. Druhá kapitola pojednává o funkcionálních datech jako takových, budou zde vysvětleny souhrnné statistické charakteristiky pro případ funkcionálních dat, které slouží k jejich základnímu popisu. Ve druhé části druhé kapitoly se budeme věnovat problematice reprezentace funkcionálních dat pomocí široce používaného přístupu interpolace bázovými splajny. V reálném životě totiž není možné sledováním určitých jevů získat spojité funkce, ty musíme vytvořit z často velmi obsáhlého souboru diskrétních pozorování majících funkcionální charakter. Následuje kapitola třetí, jejíž součástí bude i motivace, která vedla k zavedení Bayesových prostorů. Bude zde představena specifická geometrie hustot zobecněním Aitchisonovy geometrie, která respektuje relativní charakter mnohorozměrných dat, známých pod názvem kompoziční data. Následně si ukážeme, že množina hustot společně se zavedenou geometrií reprezentací má strukturu Hilbertova prostoru. Ve čtvrté kapitole se budeme zabývat clr transformací, která nám umožní použití standardních metod pro statistické zpracování hustot. Funkcionální metoda hlavních komponent, ve které bude zohledněna metodika Bayesových prostorů, bude obsahem kapitoly páté. Její součástí bude i simulační studie, kde si fungování této statistické metody představíme na souborech hustot pocházejících z exponenciálních rodin, které jsou v současnosti hojně využívány v praxi.

Teoretická část bude na vhodných místech doplněna ilustrativními příklady.

V závěru práce se pak pokusím uplatnit získané teoretické poznatky při statistickém zpracování souboru reálných dat, tvořeném pravděpodobnostními hustotami, za pomoci volně dostupného statistického softwaru R.

1. Základní poznatky z funkcionální analýzy

V úvodní kapitole se seznámíme se základními pojmy z oblasti funkcionální analýzy, které později využijeme při popisu speciálních lineárních prostorů zvaných Bayesovy prostory. Při tvorbě kapitoly byly informace čerpány zejména z [10], [11].

Připomeňme si, že množina X tvoří reálný lineární prostor, jestliže jsou na X definovány operace sčítání a násobení reálným číslem splňující dvě podmínky:

1. Vzhledem ke sčítání je X komutativní grupa.

2. Vzhledem k násobení reálným číslem platí následující vlastnosti. Pro všechna $\mathbf{x}, \mathbf{y} \in X$ a $\alpha, \beta \in \mathbf{R}$ máme:

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

1.1. Normovaný lineární prostor

Normovaným lineárním prostorem rozumíme každý lineární prostor X, který je vybavený **normou**, což je funkcionál $\|\cdot\| : X \to \mathbf{R}$, splňující pro každé $\mathbf{x}, \mathbf{y} \in X$ a $\alpha \in \mathbf{R}$ podmínky:

- $\|\mathbf{x}\| \geq 0$, přičemž $\|\mathbf{x}\| = 0$ právě když $\mathbf{x} = \mathbf{0}$,
- $\|\lambda \cdot \mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|$,
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

Poznámka 1.1 Funkcionálem, jednoduše řečeno, rozumíme zobrazení, které přiřazuje funkci číslo. Odtud vzešel i název této matematické disciplíny jakožto funkcionální analýza.

Definice 1.1 (Izomorfní a izometrické zobrazení) *Nechť X a Y jsou reálné lineární prostory. X a Y nazveme izomorfní, jestliže existuje prosté lineární zobrazení $T : X \to Y$, které zachovává operace sčítání a násobení reálným číslem. Pokud navíc zachovává i normu prvků, pak T nazýváme izometricky izomorfním zobrazením X na Y.*

Později se seznámíme s Hilbertovými prostory, kde T bude zachovávat i skalární součin.

Normované lineární prostory, které jsou navíc *úplné*, nazýváme **Bachanovy prostory**. Přičemž úplností rozumíme, že každá cauchyovská posloupnost prvků z tohoto prostoru konverguje vzhledem k dané normě k prvku z tohoto prostoru.

1.2. Lineární prostor se skalárním součinem a Hilbertův prostor

Lineární prostor X, na němž je definovaný skalární součin, nazveme *lineární prostor se skalárním součinem*, označovaný též jako pre-Hilbertův prostor. Přitom skalárním součinem rozumíme funkcionál $\langle \cdot, \cdot \rangle : X \times X \to \mathbf{R}$ splňující pro všechna $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$ a $\alpha \in \mathbf{R}$ podmínky:

- $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, přičemž $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ právě když $\mathbf{x} = \mathbf{0}$,
- $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$,
- $\langle \lambda \cdot \mathbf{x}, \mathbf{y} \rangle = \lambda \cdot \langle \mathbf{x}, \mathbf{y} \rangle$,
- $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$.

Norma definovaná pomocí skalárního součinu, $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, je pak normou na tomto prostoru, nazývána též jako norma indukovaná skalárním součinem.

Následně je-li norma v Bachanově prostoru indukovaná skalárním součinem, pak tento prostor nazveme prostorem **Hilbertovým**.

Poznámka 1.2 Zavedení Hilbertova prostoru umožňuje měřit úhly a vzdálenosti mezi prvky z lineárních prostorů nekonečné dimenze, jako například v prostorech nekonečných posloupností či funkcí.

1.2.1. Příklady Hilbertova prostoru

- Prostory $\mathbf{R}^n$

  Prostory všech n-tic, na nichž je definován skalární součin daným vztahem

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

  tvoří Hilbertův prostor konečné dimenze.

- Prostor l^2

  Prostor všech nekonečných posloupností reálných čísel $\mathbf{x} = \{x_i\}_{i=1}^{\infty}$, jejichž součet druhých mocnin absolutních hodnot je konečný, tj.

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

  spolu se skalárním součinem

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

  tvoří Hilbertovův prostor nekonečné dimenze.

- Prostor $L^2[a, b]$

  Uvažujme prostor Lebesgueovsky měřitelných funkcí z $[a, b] \to \mathbf{R}$, které jsou integrovatelné s kvadrátem na $[a, b]$, tj. integrál

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

  existuje a je konečný. Pak tento prostor funkcí, na němž je definovaný skalární součin jako

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

tvoří Hilbertův prostor nekonečné dimenze. Norma indukovaná tímto skalárním součinem je dána vztahem

$$\|f\|_2 = \sqrt{\langle f, f\rangle_2} = \sqrt{\int_a^b f(t)\, f(t)\, dt} = \sqrt{\int_a^b f(t)^2\, dt}$$

a vzdálenost d dvou funkcí f a g z $L^2[a,b]$ určíme pomocí normy následovně

$$d_2(f,g) = \|f-g\|_2 = \sqrt{\langle f-g, f-g\rangle_2} = \sqrt{\int_a^b (f(t)-g(t))^2\, dt}.$$

Skalární součin nám umožňuje na prostorech se skalárním součinem, tedy i Hilbertově prostoru, určit kolmost (ortogonalitu) prvků. Přičemž platí, že každý prvek je kolmý na nulový prvek v daném prostoru. Uvažujme množinu nenulových prvků $\{\mathbf{x}_i, i \in I\}$ z Hilbertova prostoru, kde I značí indexovou množinu. Pak tuto množinu nazveme **ortogonální**, platí-li pro

$$\forall i \neq j : \langle \mathbf{x}_i, \mathbf{x}_j\rangle = 0, \quad i,j \in I.$$

Je-li navíc $\|\mathbf{x}_i\| = 1$ pro každé $i \in I$, pak tuto množinu nazveme **ortonormální**. V případě, že I je spočetná, hovoříme o $\{\mathbf{x}_i, i \in I\}$ jako o posloupnosti.

Poznámka 1.3 Každou ortogonální množinu lze převést na množinu ortonormální podělením každého prvku příslušnou normou.

1.2.2. Ortonormální báze v Hilbertově prostoru

Množina $\{\mathbf{e}_i, i \in I\}$ prvků z $\mathcal{H}$ tvoří ortonormální bázi prostoru $\mathcal{H}$, jestliže

1. všechny prvky báze jsou navzájem kolmé a mají jednotkovou délku, tzn. jsou **ortonormální**,

2. lineární obal báze je hustý v $\mathcal{H}$ (uzávěr lineárního obalu je roven $\mathcal{H}$).

Z druhé podmínky vyplývá, že libovolný prvek lze zapsat jako součet nekonečné řady a díky ortogonalitě bude toto vyjádření jediné. Druhou podmínku lze ekvivaletně vyjádřit jako požadavek na **úplnost** množiny $\{\mathbf{e}_i, i \in I\}$, kterou přibližuje následující definice.

Definice 1.2 (O úplnosti) *Nechť $\mathcal{H}$ je Hilbertův prostor. O množině $\{\mathbf{x}_i, i \in I\}$ prvků z $\mathcal{H}$ řekneme, že je **úplná**, jestliže jediný prvek $\mathbf{z} \in \mathcal{H}$, který je kolmý ke všem prvkům uvažované množiny, je nulový prvek. (Posloupnost $\{\mathbf{x}_i, i \in I\}$ je **úplná**, jestliže z podmínky $\langle \mathbf{z}, \mathbf{x}_i \rangle = 0$ pro všechna $i \in I$ a nějaké $\mathbf{z} \in \mathcal{H}$ plyne, že $\mathbf{z} = 0$).*

Věta 1.1 *V každém Hilbertově prostoru existuje ortonormální báze.*

Často se vyžaduje, aby byl Hilbertův prostor *separabilní*. Separabilitu nám objasní následující definice.

Definice 1.3 ***Separabilním*** *Hilbertovým prostorem rozumíme Hilbertův prostor, na němž existuje nejvýše spočetná množina prvků $\mathbf{e}_i$ taková, že libovolný prvek $\mathbf{x}$ lze vyjádřit jako lineární kombinaci $\mathbf{x} = \sum_{i \in I} \alpha_i \mathbf{e}_i$.*

Věta 1.2 *Nechť $\mathcal{H}_1$ a $\mathcal{H}_2$ jsou separabilní Hilbertovy prostory nekonečné dimenze. Potom jsou izometricky izomorfní.*

Důležitým důsledkem věty 1.2 je, že každý separabilní Hilbertův prostor je izometricky izomorfní s l^2. Tento poznatek následně využijeme ve třetí kapitole.

1.2.3. Fourierovy řady v Hilbertově prostoru

Nyní si uvedeme základní pojmy z teorie Fourierových řad v Hilbertově prostoru $\mathcal{H}$ a důležitá tvrzení, která později využijeme.

Definice 1.4 *Uvažujme ortonormální posloupnost nenulových prvků $\{\mathbf{e}_n\}_{n=1}^{\infty} \in \mathcal{H}$. Nechť $\mathbf{x} = \sum_{n=1}^{\infty} a_n \mathbf{e}_n$, kde $a_n \in \mathbf{R}$. Potom čísla*

$$a_n = \langle \mathbf{x}, \mathbf{e}_n \rangle , \quad n \in \mathcal{N}$$

*nazveme **Fourierovými koeficienty** prvku $\mathbf{x}$ vzhledem k ortonormální posloupnosti $\{\mathbf{e}_n\}_{n=1}^{\infty}$.*

Definice 1.5 *Uvažujme ortonormální posloupnost nenulových prvků* $\{\mathbf{e}_n\}_{n=1}^{\infty} \in \mathcal{H}$, $\mathbf{x} \in \mathcal{H}$ *a Fourierovy koeficienty* a_n *z definice* 1.4. *Potom řadu*

$$\sum_{n=1}^{\infty} a_n \mathbf{e}_n = \sum_{n=1}^{\infty} \langle \mathbf{x}, \mathbf{e}_n \rangle \, \mathbf{e}_n$$

nazveme (abstraktní) ***Fourierovou řadou*** *prvku* $\mathbf{x}$ *v prostoru* $\mathcal{H}$ *vzhledem k ortonormální posloupnosti* $\{\mathbf{e}_n\}_{n=1}^{\infty}$.

Věta 1.3 *Nechť* $\{\mathbf{e}_n\}_{n=1}^{\infty}$ *je ortonormální posloupnost prvků z* $\mathcal{H}$, $\mathbf{x} \in \mathcal{H}$ *a nechť* a_n *jsou Fourierovy koeficienty z definice* 1.4. *Potom*

- *Fourierova řada* $\sum_{n=1}^{\infty} a_n \mathbf{e}_n$ *vždy konverguje k nějakému prvku* $\mathbf{z} \in \mathcal{H}$,
- *vždy platí Besselova nerovnost:* $\sum_{n=1}^{\infty} |a_n|^2 \leq \|\mathbf{x}\|^2$,
- *Parsevalova rovnost* $\sum_{n=1}^{\infty} |a_n|^2 = \|\mathbf{x}\|^2$ *platí právě tehdy, když* $\mathbf{x} = \sum_{n=1}^{\infty} a_n \mathbf{e}_n$.

Poznámka 1.4 Parsevalovu rovnost lze psát v souladu s definicí 1.4 ve tvaru

$$\sum_{n=1}^{\infty} |\langle \mathbf{x}, \mathbf{e}_n \rangle|^2 = \|\mathbf{x}\|^2,$$

která nám takto připomíná známou Pythagorovu větu.

Věta 1.4 *Nechť* $\{\mathbf{e}_n\}_{n=1}^{\infty}$ *je ortonormální posloupnost prvků z* $\mathcal{H}$. *Pak následující tvrzení jsou ekvivalentní:*

- $\{\mathbf{e}_n\}_{n=1}^{\infty}$ *tvoří ortonormální bázi na* $\mathcal{H}$,
- *pro všechna* $\mathbf{x} \in \mathcal{H}$ *platí Parsevalova rovnost vzhledem k* $\{\mathbf{e}_n\}_{n=1}^{\infty}$,
- $\mathbf{x} = \sum_{n=1}^{\infty} \langle \mathbf{x}, \mathbf{e}_n \rangle \, \mathbf{e}_n$ *pro každé* $\mathbf{x} \in \mathcal{H}$.

Poslední tvrzení říká, že každý prvek $\mathbf{x} \in \mathcal{H}$ lze rozvést ve Fourierovu řadu vzhledem k $\{\mathbf{e}_n\}_{n=1}^{\infty}$.

2. Data jako funkce

Časový průběh meteorologických dat jako je teplota, tlak, úhrn srážek, dále vývoj tělesné výšky nebo váhy člověka, případně vývoj cen cenných papírů na burze cenných papírů v závislosti na čase, jsou jen některé z mnoha případů, kdy je zcela přirozené taková vstupní data brát jako funkciónalní. Funkcionální analýza dat, krátce FDA (functional data analysis), je obor matematické statistiky, který zpracovává informace o datech majících funkcionální charakter [12]. Její cíle se shodující s cíly ostatních statistických metod:

- snaží se reprezentovat data způsobem, který napomůže jejich další analýze,
- usiluje o takové zobrazení dat, které nám pomůže odhalit jejich důležité vlastnosti. Například zobrazením druhých derivací napozorovaných funkcí, které popisují výšku jednotlivých dívek v různých letech, můžeme odhalit, ve kterých letech došlo u dívek k nejrychlejšímu růstu nebo naopak, kdy se rychlost růstu začala zpomalovat. Čímž dostaneme dodatečnou informaci o datech, kterou bychom neměli k dispozici, pokud bychom data analyzovali pomocí mnohorozměrých statistických metod,
- zkoumá zdroje variability mezi daty apod.

S nástupem výkonnější počítačové techniky, která je potřebná k časově náročným výpočtům, zaznamenala FDA zejména v posledních letech velkého rozmachu. Ve této kapitole si nejprve přiblížíme základní pojmy z popisné statistiky pro případ funkcionálních dat, a to statistické charakteristiky míry polohy a variability. V druhé části pak přejdeme k problematice reprezentace funkcionálních dat a popíšeme si proces jejich získávání, neboť v praxi nemáme k dispozici soubor pozorování, který je tvořen přímo funkcemi. Namísto toho obsahuje diskrétní pozorování, která jsou měřena v různých časových okamžicích.

2.1. Statistické charakteristiky funkcionálních dat

Míry polohy pro funkcionální data [12] nám dávají informaci o tom, kde se koncentrují hodnoty měřené veličiny (teploty, úhrnu srážek apod.) v jednotlivých časových okamžicích přes celý soubor pozorování. Funkcionální aritmetický průběr je nejčastěji počítanou statistickou charakteristikou polohy, která je založena na myšlence stejného příspěvku jednotlivých statistických jednotek na společné průměrné hodnotě v čase t. Měli bychom mít na paměti, že jeho hodnoty mohou být silně ovlivněny odlehlými pozorováními, tj. naměřenými hodnotami veličiny, které jsou oproti zbývajícím naměřeným hodnotám v čase t příliš velké či malé.

Mějme funcionální náhodný výběr $X_1, \ldots, X_n$ v $L^2[a,b]$. Výběrovým **aritmetickým průměrem** pak rozumíme funkci proměnné $t \in [a,b]$ z $L^2[a,b]$ danou pro realizaci tohoto výběru vztahem

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

Geometrický průměr [7] patří mezi neméně používanou míru polohy. Jeví se výhodný zejména v situacích, kdy má věcný smysl počítat součin mezi hodnotami měřené veličiny v čase t. Je dobré si uvědomit, jaký je jeho vztah k aritmetickému průměru, neboť platí, že logaritmus z geometrického průměru je roven aritmetickému průměru zlogaritmovaných hodnot veličiny naměřené na jednotlivých statistických jednotkách.

Pro n napozorovaných funkcí $x_1(t), \ldots, x_n(t)$ určíme **geometrický průměr** ze vztahu

$$\bar{x}(t) = \sqrt[n]{x_1(t) \cdot \ldots \cdot x_n(t)} = \sqrt[n]{\prod_{i=1}^{n} x_i(t)}.$$

K určení míry koncentrace (rozptýlení) hodnot kolem odhadnuté míry polohy nám slouží charakteristiky variablity, zejména rozptyl a směrodatná odchylka. Máme-li k dispozici funkcionální náhodný výběr v $L^2[a,b]$ o rozsahu n, definujeme

výběrový **rozptyl** jako funkci $s^2(t)$ s hodnotami

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

a výběrovou **směrodatnou odchylku** $s(t) = \sqrt{s^2(t)}$, tj.

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

Při statistickém zpracování dat je důležitá znalost míry závislosti mezi naměřenými hodnotami. Zde půjde o to určit, jak spolu souvisí hodnoty jednotlivých pozorování $x_i(t_1)$ a $x_i(t_2)$ v časových okamžicích t_1 a t_2. K určení kovariance mezi funkčními hodnotami $x_i(t_1)$ a $x_i(t_2)$ v čase t_1 a t_2 slouží **kovarianční funkce** $\sigma(t_1, t_2)$. Odhadneme ji pomocí výběrové kovarianční funkce $v(t_1, t_2)$ jako

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

Pro případ, kdy jsou časové okamžiky t_1 a t_2 totožné, přechází kovarianční funkce ve výběrový rozptyl, což můžeme jasněji vidět následným dosazením do vztahu

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

$$\frac{1}{n-1} \sum_{i=1}^{n} [x_i(t_1) - \bar{x}(t_1)]^2 = s^2(t_1).$$

Kovarianční funkce se zobrazuje nejčastěji v podobě 3D grafů nebo pomocí rovinných grafů, v němž různé úrovně kovarianční funkce mohou být rozlišeny pomocí kontur či barev. Předpokládejme, že máme k dispozici data měřená v časovém intervalu $[0, T]$. Pak hodnoty v takovém grafu odpovídají hodnotám kovarianční funkce, která je spočtená pro všechny možné kombinace času $[t_1, t_2]$ z kartezského součinu $[0, T] \times [0, T]$, kdy na úhlopříčce (diagonále) dostáváme hodnoty varianční funkce, které odpovídají hodnotám kovarianční funkce v čase

t_i a t_j pro $i = j$. Je zřejmé, že pro určení celkové informace o kovarianční struktuře dat bychom si vystačili pouze s jednou polovinou grafu (nad/pod diagonálou), platí totiž $v(t_1, t_2) = v(t_2, t_1)$.

Pro snadnější interpretaci míry závislosti se častěji používá **korelační funkce**, která je očištěna od vlivu různého měřítka. Nabývá hodnot z intervalu $[-1, 1]$, kdy krajní hodnoty vypovídají o lineární závislosti mezi naměřenými hodnotami v čase t_1 a t_2. Jelikož v rámci FDA je soubor pozorování získán měřením jedné veličiny ve stejných jednotkách, korelační funkce zde ztrácí na významnosti. To nebude ale platit v případě, kdy budeme chtít kvantifikovat závislost a sílu lineárního vztahu mezi dvěma různými veličinami měřených v různých jednotkách. Diskutovanou charakteristiku získáme podělením výběrové kovarianční funkce příslušnými rozptyly ve tvaru

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

Výše uvedené vztahy můžeme vnímat jako jistou funkcionální obdobu varianční a korelační matice v mnohorozměrné statistické analýze dat.

Korelační funkce se zobrazuje stejným způsobem jako kovarianční funkce. Proto i nyní předpokládejme, že máme k dispozici data měřené v časovém intervalu $[0, T]$. Hodnoty v rovinném grafu následně odpovídají hodnotám korelační funkce, která je spočtená pro všechny možné kombinace času $[t_1, t_2]$ z $[0, T] \times [0, T]$. Na úhlopříčce (diagonále) dostáváme korelace rovny 1, které odpovídají korelacím mezi stejnými hodnotami, tj. hodnotami měřenými v časovém okamžiku $t = t_1 = t_2$. Směry kolmé na úhlopříčku (obsahující jednotkové korelace), pak vypovídají o tom, jak rychle se korelace mění, když dochází k postupnému navyšování rozdílu mezi časovými okamžiky. Zřejmě opět celou informaci vyčteme z poloviny grafu (nad/pod diagonálou), neboť $r(t_1, t_2) = r(t_2, t_1)$.

Nakonec se zmíníme ještě o křížové kovarianční a křížové korelační funkci. Jak už bylo řečeno dříve, zejména druhou charakteristiku lze využít v případě, kdy chceme matematicky vyjádřit závislost mezi naměřenými hodnotami dvou veličin.

Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je dvourozměrný funkcionální náhodný výběr z $L^2[a,b] \times L^2[a,b]$. Těsnost lineárního vztahu hodnot n napozorovaných dvojic funkcí $(x_1(t), y_1(t)), \ldots, (x_n(t), y_n(t))$ v čase t_1 a t_2, lze kvantifikovat jak pomocí **křížové kovariační funkce**,

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

tak i pomocí **křížové korelační funkce**

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

2.2. Reprezentace funkcionálních dat

Sledováním jevů v reálném životě nelze získat funkcionální pozorování v podobě spojitých funkcí, neboť bychom museli mít k dispozici nespočetně mnoho pozorování. Funkcionálními daty rozumíme pro danou funkci N takových dvojic (f_i, t_i), mezi nimiž je nějaký funkcionální vztah. Uvažovanou situaci lze popsat pomocí následujícího modelu

$$f_i = x(t_i) + \varepsilon_i, \quad i = 1, \ldots, N,$$

kde navíc uvažujeme i chybu v modelu a předpokládáme, že funkce $x(t)$ je hladká ve smyslu existence spojitých derivací až do určitého řádu. V případě velkých skoků v po sobě jdoucích naměřených hodnotách f_i bychom totiž tato data vnímali spíše jako mnohorozměrná než jako funkcionální.

Proces získávání funkcí z naměřených dat nazveme *interpolací* v případě, že naše pozorování neobsahují chybu. Jinak mluvíme o *aproximaci* a převod diskrétních dat na funkce může zahrnovat i vyhlazování, o kterém se můžete více dočíst v [12]. Pro stanovení těchto funkcí použijeme metodu lineárních kombinací bázových funkcí $\{\varphi_1, \ldots, \varphi_K\}$, o kterých předpokládáme, že jsou vzájemně lineárně nezávislé a pomocí nichž lze volbou dostatečně velkého K danou spojitou funkci aproximovat s předem stanovenou přesností. Funkci $x(t)$ rozvineme

pomocí těchto systémů jako

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

kde φ_k jsou známé bázové funkce a $(c_1, \ldots, c_K)$ je vektor neznámých koeficientů.

K nejpoužívanějším systémům bázových funkcí patří systémy Fouerierovy, které se typicky používají pro periodická data, a pro nás zajímavější bázové splajnové systémy, B-splajny, používané zejména pro neperiodická data, kterými se nyní budeme zabývat podrobněji [8], [9].

Přístup bázových splajnů je velmi oblíbeným nástrojem pro reprezentaci dat zejména proto, že umožňuje zachytit i složité průběhy v rozsáhlých datových souborech. Podrobně si nyní projdeme problematikou interpolace pomocí B-splajnů a v závěru kapitoly se pouze okrajově zmíníme o vyhlazovacích splajnech, které jsou jakýmsi kompromisem mezi interpolací pomocí B-splajnů a aproximací dat metodou nejmenších čtverců využívající B-splajny.

2.2.1. B-splajnová reprezentace

Interpolační splajn je funkce, která je po částech polynomem nižšího stupně, jejíž jednotlivé části na sebe navazují dostatečně hladce. Označme symbolem $\Delta\lambda$ rostoucí posloupnost $g + 2$ uzlů, tj. $a = \lambda_0 < \lambda_1 < \ldots < \lambda_g < b = \lambda_{g+1}$, a dále pomocí $s_k(t)$ polynomický splajn definovaný na konečném intervalu $[a, b]$ s vlastnostmi

- $s_k(t)$ je na každém intervalu $[\lambda_i, \lambda_{i+1}]$, $i = 0, 1, \ldots, g-1$, polynomem stupně nejvýše k,
- $s_k(t) \in C^{k-1}[\lambda_i, \lambda_{i+k+1}]$, tzn. $s_k(t)$ má v uzlech λ_i spojité derivace až do řádu $k - 1$.

Lineární prostor splajnů stupně $k > 0$ definovaných na $[a, b]$ na síti uzlů $\Delta\lambda$, budeme značit pomocí $\mathcal{S}_k^{\Delta\lambda}[a, b]$. Jeho dimenze je rovna

$$\dim \mathcal{S}_k^{\Delta\lambda}[a, b] = k + g + 1.$$

Obecně totiž máme $g+1$ podintervalů, odtud $(k+1)(g+1)$ neznámých koeficientů. Protože ale jednotlivé části interpolačního splajnu včetně $(k-1)$ jeho derivací na sebe musí v sousedních uzlech navazovat, máme k podmínek spojitosti ve všech vnitřních uzlech, tzn. kg podmínek spojitosti (g je počet vnitřních uzlů), které musíme odečíst. Celkově tedy máme

$$(k+1)(g+1) - kg = g + k + 1$$

parametrů.

Každý splajn $s_k(t)$ lze tedy jednoznačně vyjádřit jako lineární kombinaci nějakých $g+k+1$ bázových funkcí. V této práci budeme pracovat s B-splajnovou bází. Na dané síti uzlů $\Delta\lambda$ lze ale zkonstruovat pouze $g+1-k$ lineárně nezávislých B-splajnů. Abychom získali všechny B-splajny, je potřeba rozšířit danou síť uzlů $\Delta\lambda$ o $2k$ uzlů, které splňují podmínku

$$\lambda_{-k} \leq \lambda_{-k+1} \leq \ldots \leq \lambda_0, \qquad \lambda_{q+1} \leq \lambda_{q+2} \leq \ldots \leq \lambda_{q+k+1}.$$

My budeme nadéle pracovat s takovou rozšířenou sítí uzlů, kdy se přidané uzly budou rovnat λ_0 a λ_{g+1}, tj.

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

Potom každý splajn $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ lze jednoznačně vyjádřit jako

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

kde vektor $\mathbf{b} = (b_{-k}, \ldots, b_g)'$ nazýváme vektorem B-splajnových koeficientů splajnu $s_k(t)$. Funkce $B_i^{k+1}(t)$ nazýváme B-splajny stupně k na síti uzlů $\lambda_i, \ldots, \lambda_{i+k+1}$ s vlastnostmi

1. nezápornost

$$B_i^{k+1}(t) \geq 0, \quad \text{pro všechna} \quad t \in \mathbf{R},$$

2. nosičem $B_i^{k+1}(t)$ je interval $[\lambda_i, \lambda_{i+k+1}]$, tj.

$$B_i^{k+1}(t) = 0 \quad \text{pro} \quad t \notin [\lambda_i, \lambda_{i+k+1}],$$

3. $B_i^{k+1}(t)$ je na každém intervalu $[\lambda_i, \lambda_{i+1}], i = 0, 1, \ldots, g-1$ polynomem stupně nejvýše k.

Důsledkem druhé vlastnosti je tvrzení, že na $[\lambda_i, \lambda_{i+1}]$ je celkem $k+1$ nenulových B-splajnů $B_i^{k+1}(t)$.

Příklad 1 Určete bázi prostoru kubických splajnů $\mathcal{S}_3^{\Delta\lambda}[0,7]$ na ekvidistantní síti uzlů $\Delta t = \{0,1,2,3,4,5,6,7\}$.

Řešení. Protože na této síti uzlů nelze sestrojit všech $g+k+1 = 6+3+1 = 10$ B-splajnů, je potřeba danou síť rozšířit o 6 uzlů. Výslednou bázi na rozšířené síti uzlů $\{0,0,0,0,1,2,3,4,5,6,7,7,7,7\}$ můžeme vidět na obrázku 1. ∘

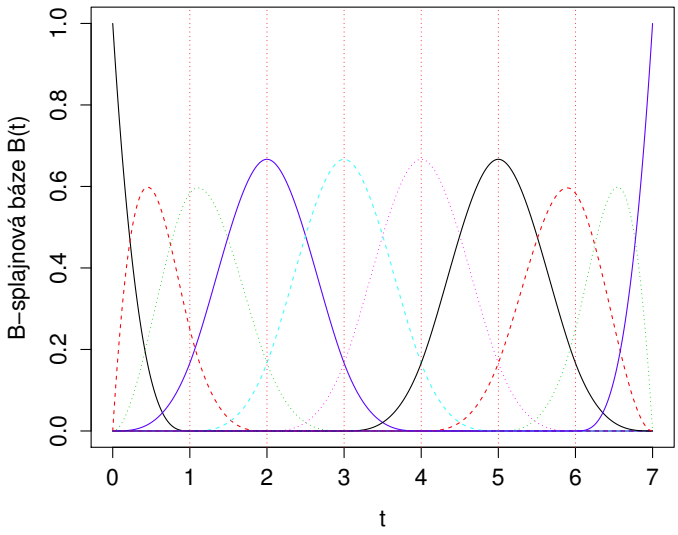
Všimněte si, že všechny tři výše zmíněné vlastnosti zde opravdu platí. Každá ze čtyř bázových funkcí uprostřed je kladná pouze přes čtyři sousední podintervaly a díky ekvidistantní síti uzlů mají také stejný tvar. Naopak tři bázové funkce vlevo a tři bázové funkce vpravo se následkem násobnosti příslušných uzlů svými tvary liší. Nicméně stále platí, že žádná z nich není kladná přes více než čtyři sousední podintervaly. Protože kubické splajny mají $k-1 = 2$ spojitých derivací, každá z bázových funkcí vytváří hladký přechod v místě, kdy dosáhnou nulových hodnot.

2.2.2. Interpolace splajny

Nechť je dána ekvidistantní síť bodů interpolace t_j a funkční hodnoty v těchto bodech $f(t_j)$ pro $j = 0, 1, \ldots, g+1$. Úkolem je najít interpolační splajn $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ ve tvaru $s_k(t) = \sum_{i=-k}^{g} b_i B_i^{k+1}(t)$ splňující v bodech t_j podmínky interpolace

$$s_k(t_j) = \sum_{i=-k}^{g} b_i B_i^{k+1}(t_j) = f(t_j) \quad \text{pro všechna} \quad j = 0, 1, \ldots, g+1, \tag{2}$$

kde b_i jsou neznámé koeficienty. Abychom mohli danou soustavu lineárních rovnic zapsat v maticovém tvaru, je potřeba nadefinovat důležitý pojem kolokační matice, o které pojednává následující definice.



Obrázek 1: Deset bázových funkcí, které tvoří lineární prostor všech kubických splajnů definovaných na $[0, 7]$. Svislé čárkované přímky odpovídají šesti vnitřním uzlům.

Definice 2.1 *Nechť je dán vektor bodů interpolace* $\mathbf{t} = (t_1, \ldots, t_n)'$ *a nechť* $\left\{B_i^{k+1}(t)\right\}_{i=-k}^{g}$ *je B-splajnová báze* $\mathcal{S}_k^{\Delta\lambda}[a, b]$. *Pak*

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

nazveme kolokační maticí.

Pomocí kolokační matice z předchozí definice lze soustavu (2) psát ve tvaru

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

kde $\mathbf{s}_k(\mathbf{t}) = (s_k(t_1), \ldots, s_k(t_n))'$. Jednoznačnost řešení této soustavy a tedy i jednoznačnost hledaného interpolačního splajnu záleží na regularitě matice této soustavy, tj. na kolokační matici $\mathbf{C}_{k+1}(\mathbf{t})$.

Uvažujeme-li $g+2$ podmínek interpolace, z celkové počtu parametrů $g+k+1$ nám po jejich odečtení zbývá volných $k-1$. Pro kubické splajny (tedy $k = 3$) máme právě dva volné parametry, které lze volit několika způsoby. V této práci

použijeme takové podmínky, které povedou na tzv. *přirozené kubické splajny*. V dodatečných podmínkách budeme proto požadovat, aby druhá derivace interpolačního kubického splajnu v krajních bodech intervalu $[a, b]$ byla nulová. Abychom byli schopni tyto podmínky použít, je potřeba si nejdříve uvést pár vět k derivacím splajnů.

Pro $l \in \{1, \ldots, k-1\}$ je l-tá derivace splajnu $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$ splajnem stupně $s_{k-1} \in \mathcal{S}_{k-1}^{\Delta\lambda}[a, b]$ na stejné síti uzlů $\Delta\lambda$. Podle [9] lze l-tou derivaci splajnu maticově zapsat ve tvaru

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

kde $\mathbf{b}^{(l)} \in \mathbf{R}^{g+k+1-l}$ určíme pomocí

$$\mathbf{b}^{(l)} = \mathbf{D}_l\mathbf{L}_l\mathbf{b}^{(l-1)} = \mathbf{D}_l\mathbf{L}_l \ldots \mathbf{D}_1\mathbf{L}_1\mathbf{b} = \mathbf{S}_l\mathbf{b}$$

s $\mathbf{b} = \mathbf{b}^{(0)}$. $\mathbf{S}_l = \mathbf{D}_l\mathbf{L}_l \ldots \mathbf{D}_1\mathbf{L}_1 \in \mathbf{R}^{g+k+1-l,g+k+1}$ je horní trojúhelníková matice s plnou řádkovou hodností, $\mathbf{D}_j \in \mathbf{R}^{g+k+1-j,g+k+1}$ je diagonální matice

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

kde prvky d_i pro všechna $i = -k+j, \ldots, g$ získáme dosazením do

$$d_i = \frac{1}{\lambda_{i+k+1-j} - \lambda_i}$$

a dvojdiagonální matice $\mathbf{L}_j$ je ve tvaru

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

Příklad 2 Vykreslete graf přirozeného kubického interpolačního splajnu pro předepsané body interpolace $\mathbf{t} = \{0, 1, 2, 3, 4, 5, 6, 7\}$ a funkční hodnoty $\mathbf{f} = \{2, 3, 3.5, 2.5, 1, 2.5, 2, 3\}$.

Řešení. Jelikož nemáme zadanou síť uzlů splajnu, vyvstává otázka, jak ji určit. V této práci ji budeme volit vždy shodnou s body interpolace. Odpovídající rozšířená síť uzlů pak odpovídá uzlům z příkladu 1. Hledaný splajn musí splňovat podmínky interpolace, tj.

$$s_3(t_j) = \sum_{i=-3}^{6} b_i B_i^4(t_j) = f(t_j) \quad \text{pro} \quad j = 0, \dots 7,$$

a dodatečné podmínky ve tvaru $s_3^{(2)}(a) = s_3^{(2)}(b) = 0$ tj.

$$s_3^{(2)}(0) = s_3^{(2)}(7) = 0.$$

Provedeme následující kroky v softwaru R [6], [13] s využitím knihovny `fda`:

1. Vygenerujeme rozšířenou síť uzlů $\Delta\lambda$.

2. Vytvoříme B-splajnovou bázi `baze3` kubických splajnů pomocí příkazu `create.bspline.basis` mající tři vstupy `rangeval`, `nbasis` a `norder`, tj. definiční obor, počet a řád B-splajnů.

3. Spočteme kolokační matici $\mathbf{C}_4(\mathbf{t})$. Příkaz `eval.basis`, vstupy body interpolace a sestrojená báze z druhého kroku.

4. Okrajové podmínky: $s_4^{(2)}(a) = \mathbf{C}_2(a)\mathbf{S}_2\mathbf{b} = 0$ a $s_4^{(2)}(b) = \mathbf{C}_2(b)\mathbf{S}_2\mathbf{b} = 0$, kde $\mathbf{C}_2(a)$ a $\mathbf{C}_2(b)$ značí příslušný řádek kolokační matice.

   - Spočteme matice $\mathbf{D}_2, \mathbf{L}_2, \mathbf{D}_1, \mathbf{L}_1$, čímž získáme matici $\mathbf{S}_2$.
   - Vytvoříme B-splajnovou bázi `baze1` lineárních splajnů pomocí příkazu z kroku 2.
   - Spočteme kolokační matice $\mathbf{C}_2$ v krajních bodech intervalu $[a, b]$, tj. $\mathbf{C}_2(a)$ a $\mathbf{C}_2(b)$ s využitím `baze1`.

5. Doplníme kolokační matici z kroku 3 o řádky $\mathbf{C}_2(a)\mathbf{S}_2$ a $\mathbf{C}_2(b)\mathbf{S}_2$ na regulární matici $\mathbf{C}$ a vektor funkčních hodnot doplníme dvěma nulami na $\mathbf{f}_d$.

6. Pomocí příkazu `solve` získáme neznámé koeficienty $\mathbf{b}$ řešením soustavy lineárních rovnic $\mathbf{Cb} = \mathbf{f}_d$.

7. Výsledný přirozený kubický splajn získáme pomocí příkazu `fd` se dvěma vstupy (`b` a `baze3`) a vykreslíme ho pomocí příkazu `plot`.

Konkrétně v našem příkladě máme:
Načteme knihovnu `fda` pomocí příkazu `library(fda)`, do proměnné `t` si uložíme body interpolace a do proměnné `f` funkční hodnoty v bodech interpolace:

```
t=c(0,1,2,3,4,5,6,7)
f=c(2,3,3.5,2.5,1,2.5,2,3)
```

Dále provedeme kroky 1-7 :

1. ```
   > lambda # rozšířená síť uzlů splajnu
   [1] 0 0 0 0 1 2 3 4 5 6 7 7 7 7
   ```

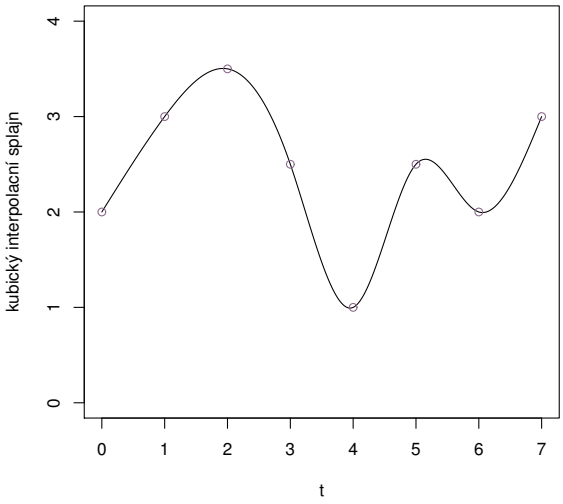
2. ```
   nbasis = 10
   norder = 4
   baze3 = create.bspline.basis(c(0,7),nbasis,norder)
   ```

3. ```
   C4 = eval.basis(t, baze3)
   ```

4. - Určíme matice $\mathbf{D}_2$, $\mathbf{D}_1$ a $\mathbf{L}_2$ $\mathbf{L}_1$ a spočteme matici $\mathbf{S}_2$:
     ```
     S2 = D2%*%L2%*%D1%*%L1
     ```
   - ```
     nbasis1 = 8
     norder1 = 2
     baze1 = create.bspline.basis(c(0,7),nbasis1,norder1)
     ```
   - ```
     C2a = eval.basis(t[1],baze1)
     C2b = eval.basis(t[length(t)],baze1)
     ```

5. ```
   C = rbind(C4,C2a%*%S2,C2b%*%S2)
   fd = (2.0,3.0,3.50,2.50,1.0,2.50,2.0,3.0,0.0,0.0)
   ```

6. ```
   b = solve(C,fd) # hledané neznámé koef. b
   > t(round(b,2))
   [1,] 2 2.35 3.05 3.78 2.82 -0.06 3.42 1.4 2.47 3
   ```
   ```



Obrázek 2: Výsledný přirozený kubický splajn z příkladu 2.

7.
```
splajn = fd(b,baze3)
plot(splajn,ylab="kubický interpolační splajn",xlab="t",
ylim=c(0,4))
points(t,f, col="plum4")
```

Výsledný přirozený kubický splajn můžeme vidět na obrázku 2.

2.2.3. Vyhlazující splajny

Vyhlazování splajny [9] je metodou aproximace dat, která je kompromisem mezi interpolací splajny a aproximací dat ve smyslu metody nejmenších čtverců. Nechť je dána rostoucí posloupnost bodů interpolace $\{t_1, \ldots, t_n\}$ z intervalu $[a, b]$ a nechť f_i jsou funkční hodnoty v těchto bodech. Dále nechť w_i jsou nezáporné váhy a α je daný kladný parametr. Index i uvažujeme vždy tak, že $i = 1, 2, \ldots, n, n \geq g + 1$. Naším úkolem je sestrojit **vyhlazující splajn** $s_k(t) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$, který bude na uvažovaném intervalu $[a, b]$ minimalizovat funkci

$$J_l(s_k) = \int_a^b \left[s_k^{(l)}(t)\right]^2 dt + \alpha \sum_{i=1}^{n} w_i \left[f_i - s_k(t_i)\right]^2 \tag{3}$$

pro $l \in \{1, \ldots, k-1\}$.

V článku [8] je detailně popsán jak postup převedení úlohy (3) na úlohu hledání B-splajnových koeficientů minimalizací funkce $J_l(\mathbf{b})$, tak jednoznačnost řešení této úlohy. V případě, že je funkce $J_l(\mathbf{b})$ ryze konvexní, hledaný vektor B-koeficientů je dán jednoznačně. Označme ho jako $\mathbf{b}^* = (b^*_{-k}, \ldots, b^*_g)'$. Potom splajn

$$s^*_k(t) = \sum_{i=-k}^{g} b^*_i B^{k+1}_i(t) \tag{4}$$

nazveme nejlepší aproximací funkce $J_l(s_k)$ ve smyslu metody vyhlazovacích splajnů.

Vyhlazovací splajny představují vhodnou formu aproximace zejména v situacích, kdy jsou naše pozorovaná data f_i zatížena chybami, a proto nemá smysl vyžadovat jejich přesnou interpolaci.

3. Hustoty jako funkce

Při statistické analýze funkcionálních dat se často můžeme setkat s tím, že naše pozorování budou tvořena hustotami. Například tak může být náš datový soubor dat tvořen hustotami, kdy každá z nich popisuje rozdělení příjmů domácností v určitých regionech dané země. Otázkou je, zda můžeme na takový soubor dat použít standardní statistické metody. Protože se jedná o nezáporné funkce, jejichž integrál je konečný (roven jedné), patří do prostoru L^1 (prostor všech absolutně integrovatelných funkcí na $I \subseteq \mathbf{R}$). Bohužel se ukazuje, že hustoty společně se standardním sčítáním funkcí a násobením funkce reálnou konstantou nemají strukturu lineárního prostoru (součet dvou hustot a násobek hustoty nemusí být nezbytně hustotou) [3], [4], [5].

Hustoty můžeme vnímat jako speciální případ funkcionálních dat s omezující podmínkou na hodnotu jejich integrálu. S podobným případem se můžeme setkat v rámci mnohorozměrných dat nesoucích pouze relativní informaci (s konstantním součtem), vyjádřených nejčastěji v podobě proporcí či procent [1]. Relativní informací rozumíme, že jediná relevantní informace je obsažena v podílech mezi složkami mnohorozměrného pozorování. Kompoziční data, jak se tento typ dat nazývá, vykazují specifické vlastnosti. Jsou *invariantní na změnu měřítka*, tzn. nezáleží na jednotkách, ve kterých jsou kompozice vyjádřeny, protože relativní informace se tím nezmění. Další důležitou vlastnostní je vlastnost *relativního měřítka*: zatímco absolutní změna při nárustu hodnoty proměnné z 0.15 na 0.3 je stejná jako při nárustu z 0.45 na 0.6, relativní nárust je v prvním případě dvakrát větší, ve druhém je větší jen o jednu třetinu. Jelikož euklidovská geometrie nerespektuje uvedené přirozené vlastnosti kompozičních dat, byla vytvořena geometrie nová, Aitchisonova. Výběrovým prostorem pro tento typ dat je simplex, Aitchisonova geometrie má vlastnosti konečně rozměrného Hilbertovova prostoru.

Kompoziční data reprezentujeme pomocí vektorů majících konečný počet kladných složek s libovolným, ale pevně zvoleným součtem (rovným např. jedné v případě proporcionálních dat). Každou hustotu lze pak vnímat jako kompoziční vektor mající nekonečně mnoho složek a jako takové převzaly od kompozič-

ních dat jejich přirozené vlastnosti týkající se měřítka. Tyto vlastnosti jsou zcela ignorovány v případě, že jsou hustoty analyzovány jako běžná funkcionální data, a proto použití standardních statistických metod může v takovém případě vést k zavádějícím, někdy až nesmyslným závěrům. Myšlenka speciálních lineárních prostorů, zvaných Bayesovy, popisujících geometrické vlastnosti hustot, přišla spolu s myšlenkou zobecnění Aitchisonovy geometrie pro hustoty definováním Hilbertova prostoru na množině hustot. Z teoretických důvodů se zde omezujeme jen na hustoty, které jsou spojité na omezeném intervalu I. Tato podmínka pro nás ale není nikterak omezující, neboť se v praxi v důsledku realizuje každé měření hodnot statistického znaku (i s teoreticky neomezeným oborem hodnot) na omezeném intervalu. V této kapitole si ukážeme, že prostor spojitých hustot definovaných na konečném intervalu I společně s operací pertubace (odpovídající sčítaní na tomto prostoru) a mocninnou trasformací (ekvivaletní násobení reálnou konstantou) tvoří lineární prostor. Následně zobecněním Aitchisonova skalárního součinu získáme strukturu pre-Hilbertova prostoru. A nakonec uvidíme, že hustoty, pro něž je součet druhých mocnin odpovídajících Fourierových koeficientů konečný, budou tvořit úplný Hilbertův prostor.

3.1. Bayesův prostor hustot

Hustoty popisují relativní příspěvky borelovských množin na omezeném intervalu $I = [a, b]$ na celkové pravděpodobnosti realizace náhodné proměnné na I, proto lze o nich hovořit jako o funkcionálních datech nesoucích pouze relativní informaci. Podobně jako diskrétní kompozice jsou hustoty taktéž invariatní na změnu meřítka, neboť zobecněním omezující podmínky na hodnotu integrálu na $\int_a^b f(t) = c, c \in \mathbf{R}$, kde f je hustota definovaná na $I = [a, b]$, nedojde ke změně relativní informace (nezmění se relativní příspěvky uvažovaných podmnožin I na celkové pravděpodobnosti realizace náhodné proměnné na $[a, b]$). Proto omezující podmínku na jednotkovou hodnotu integrálu můžeme vnímat jako takovou reprezentaci hustot, která nám umožňuje snadnější interpretaci. Stejně tak pro hustoty platí vlastnost relativního měřítka: přestože absolutní změna při nárustu prav-

děpodobnosti realizace v daných borelovských množinách z 0.15 na 0.3 je stejná jako z 0.45 na 0.6, relativní změna je rozdílná. V prvním případě je dvakrát tak větší, ve druhém dojde ke zvětšení pouze o jednu třetinu.

Z důvodů uvedených výše omezíme naši pozornost na spojité hustoty definované na $[a,b]$. Hustota je pak reálná funkce daná zobrazením $f : [a,b] \to \mathbf{R}$, splňující dvě podmínky:

1. $f(t) > 0$,
2. $\int_a^b f(t) = 1$.

Přitom druhou podmínku můžeme ekvivalentně psát jako $\|f\|_1 = \int_a^b |f|\, dt = 1, f \in L^1[a,b]$ a chápeme ji vlastně pouze jako vhodně zvolenou reprezentaci dané hustoty. Z uvedené definice hustot vyplývá, že musí být na intervalu $[a,b]$ omezené, a pomocí $\mathcal{B}_b^2[a,b]$ označíme množinu takových hustot definovaných na intervalu $[a,b]$.

3.1.1. Geometrie spojitých kompozic

Nejdříve si nadefinujeme základní operace pro práci s hustotami. Jak už bylo nastíněno, operace pertubace a mocninná transformace získáme zobecněním Aitchisonovovy geometrie na simplexu pro spojité kompozice (hustoty).

Definice 3.1 *Nechť jsou dány hustoty $f, g \in \mathcal{B}_b^2[a,b]$. Pak pertubaci definujeme jako operaci $\oplus : \mathcal{B}_b^2[a,b] \times \mathcal{B}_b^2[a,b] \to \mathcal{B}_b^2[a,b]$ danou vztahem*

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_a^b f(s)g(s)ds} = \mathcal{C}(fg), \quad t \in (a,b),$$

kde $\mathcal{C}$ je operace uzávěru, prostřednictvím které lze měnit hodnotu integrálu spojitých kompozic (bez ztráty relativní informace obsažené v hustotách).

Neutrálním prvkem operace pertubace je hustota mající rovnoměrné rozdělení na intervalu $[a,b]$,

$$e(t) = \frac{1}{b-a},$$

neboť

$$(f \oplus e)(t) = \frac{f(t) \cdot \frac{1}{b-a}}{\int_a^b f(s) \cdot \frac{1}{b-a} ds} = \frac{f(t) \cdot \frac{1}{b-a}}{\frac{1}{b-a} \int_a^b f(s) ds} = \frac{f(t)}{1} = f(t).$$

Poznámka 3.1 Všimněte si, že pertubace hustot je ekvivalentní použití Bayesovy věty pro hustoty. Odtud pak vznikl i název prostoru hustot jakožto **Bayesova prostoru**.

Definice 3.2 *Nechť je dána hustota $f \in \mathcal{B}_b^2[a,b]$ a reálná konstanta α. Pak mocninnou transformaci definujeme jako operaci $\odot : \mathbf{R} \times \mathcal{B}_b^2[a,b] \to \mathcal{B}_b^2[a,b]$ danou vztahem*

$$(\alpha \odot f)(t) = \frac{f^\alpha(t)}{\int_a^b f^\alpha(s) ds} = \mathcal{C}(f^\alpha), \quad t \in [a,b].$$

Neutrálním prvkem mocninné transformace je 1, protože platí

$$(1 \odot f)(t) = \frac{f^1(t)}{\int_a^b f^1(s) ds} = f(t).$$

Následně zavedeme operaci $\ominus$, analogii odčítání funkcí v $L^2[a,b]$. Pro $f, g \in \mathcal{B}_b^2[a,b]$ a operaci $\ominus$ platí

$$(f \ominus g)(t) = (f \oplus [(-1) \odot g])(t), \quad t \in [a,b].$$

Věta 3.1 *Množina spojitých hustot definovaných na konečném intervalu společně s operacemi pertubace a mocninná transformace tvoří lineární prostor.*

Důkaz: Musíme dokázat, že množina spojitých hustot definovaných na intervalu $[a,b]$ tvoří společně s operací pertubace komutativní grupu a mocninná transformace má stejné vlastnosti jako vnější součin v euklidovském prostoru.

1. $(\mathcal{B}_b^2[a,b], \oplus)$ tvoří komutativní grupu, tzn. pro $f, g, h \in \mathcal{B}_b^2[a,b]$ platí:

   - $(f \oplus g)(t) \in \mathcal{B}_b^2[a,b]$, tj. $(f \oplus g)(t)$ je opět spojitá hustota na $[a,b]$, neboť funkce $f(t)g(t)$ je spojitá;

- *komutativita*: $f \oplus g = g \oplus f$;
- *asociativita*: $(f \oplus g) \oplus h = f \oplus (g \oplus h)$;
- existuje jediný *neutrální prvek* $e(t) = \frac{1}{b-a}$ takový, že platí $f \oplus e = e \oplus f = f$;
- ke každé funkci f existuje jediný *inverzní prvek* f^{-1} tak, že $f \oplus f^{-1} = f^{-1} \oplus f = e$, totiž

$$\left(f \oplus f^{-1}\right)(t) = \frac{f(t)f^{-1}(t)}{\int_a^b f(s)f^{-1}(s)ds} = \frac{1}{\int_a^b 1ds} = \frac{1}{b-a}$$
$$= \left(f^{-1} \oplus f\right)(t) = e(t).$$

2. Vlastnosti operace mocninné transformace. Pro $f, g \in \mathcal{B}_2^b[a,b]$ a reálné konstanty α a β platí:
   - $(\alpha \odot f)(t) \in \mathcal{B}_2^b[a,b]$ pro $\alpha \in \mathbf{R}$. Toto tvrzení platí, stačí si uvědomit, že funkce f^α je opět spojitá;
   - *asociativita*: $\alpha \odot (\beta \odot f) = (\alpha \cdot \beta) \odot f$;
   - *distributivní vlastnost* 1: $\alpha \odot (f \oplus g) = (\alpha \odot f) \oplus (\alpha \odot g)$;
   - *distributivní vlastnost* 2: $(\alpha + \beta) \odot f = (\alpha \odot f) \oplus (\beta \odot f)$;
   - existuje *neutrální prvek*: $1 \odot f = f$.

Protože uvedené vlastnosti jsou zřejmé z definic 3.1 a 3.2, trojice $(\mathcal{B}_b^2[a,b], \oplus, \odot)$ tvoří lineární prostor. □

Abychom dostali strukturu pre-Hilbertova prostoru, musíme na $(\mathcal{B}_b^2[a,b], \oplus, \odot)$ nadefinovat skalární součin. Následně pak získáme normu a vzdálenost, čímž se zmíněný prostor stane též normovaným a metrickým.

Definice 3.3 *Nechť $f, g \in \mathcal{B}_b^2[a,b]$. Skalární součin je definovaný jako funkcionál $\langle \cdot, \cdot \rangle_B : \mathcal{B}_b^2[a,b] \times \mathcal{B}_b^2[a,b] \to \mathbf{R}$ daný vztahem*

$$\langle f, g \rangle_B = \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f(t)}{f(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds,$$

kde η značí délku intervalu $[a,b]$, tj. $\eta = b - a$.

Věta 3.2 *Bayesův skalární součin splňuje standardní vlastnosti skalárního součinu. Pro* $f, g, h \in \mathcal{B}_b^2[a,b]$ *a* $\alpha \in \mathbf{R}$ *platí:*

1. *komutativita:* $\langle f, g\rangle_B = \langle g, f\rangle_B$,
2. *pozitivita:* $\langle f, f\rangle_B \geq 0$, *přičemž* $\langle f, f\rangle_B = 0$ *pro* $f = e$,
3. *distributivita:* $\langle f \oplus h, g\rangle_B = \langle f, g\rangle_B + \langle h, g\rangle_B$,
4. $\langle \alpha \odot f, g\rangle_B = \alpha \cdot \langle f, g\rangle_B$.

Důkaz:

1. *komutativita*: zřejmě platí $\langle f, g\rangle_B = \langle g, f\rangle_B$.
2. *pozitivita*: pomocí definice 3.3 si postupně rozepíšeme skalární součiny hustot $\langle f, f\rangle_B$ a $\langle e, e\rangle_B$. V prvním případě pro $f \neq e$ dostáváme

$$\langle f, f\rangle_B = \frac{1}{2\eta}\int_a^b\int_a^b \ln\frac{f(t)}{f(s)}\cdot\ln\frac{f(t)}{f(s)}dt\,ds = \frac{1}{2\eta}\int_a^b\int_a^b\left(\ln\frac{f(t)}{f(s)}\right)^2 dt\,ds > 0;$$

   protože $\ln\frac{f(t)}{f(s)} \neq 0$ pro $t, s \in [a,b]$ vyjma $t = s$, máme určitý integrál z kladné funkce (vyjma množiny míry nula), výsledkem je tedy kladné číslo. Pro $f = e$ obdržíme

$$\langle e, e\rangle_B = \frac{1}{2\eta}\int_a^b\int_a^b\left(\ln\frac{\frac{1}{b-a}}{\frac{1}{b-a}}\right)^2 dt\,ds = \frac{1}{2\eta}\int_a^b\int_a^b (\ln 1)^2\,dt\,ds = 0,$$

   neboť určitý integrál z nulové funkce je roven nule.

3. *distributivita*: skalární součin mezi pertubací hustot f, h a hustotou g rozepíšeme následovně,

$$\langle f \oplus h, g\rangle_B = \left\langle \frac{f(t)h(t)}{\int_a^b f(\xi)g(\xi)d\xi}, g\right\rangle_B = \frac{1}{2\eta}\int_a^b\int_a^b \ln\frac{\frac{f(t)h(t)}{\int_a^b f(\xi)g(\xi)d\xi}}{\frac{f(s)h(s)}{\int_a^b f(\xi)g(\xi)d\xi}}\cdot\ln\frac{g(t)}{g(s)}dt\,ds;$$

pokrácením a následným využitím pradidla pro počítání s logaritmy, $\ln ab = \ln a + \ln b,\ a, b > 0$, obdržíme

$$= \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f(t)h(t)}{f(s)h(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds.$$

$$\frac{1}{2\eta} \int_a^b \int_a^b \left(\ln \frac{f(t)}{f(s)} + \ln \frac{h(t)}{h(s)} \right) \cdot \ln \frac{g(t)}{g(s)} dt\, ds.$$

Tím se dostáváme k součtu dvou skalárních součinů, což jsme chtěli dokázat:

$$\frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f(t)}{f(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds + \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{h(t)}{h(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds =$$

$$= \langle f, g \rangle_B + \langle h, g \rangle_B .$$

4. Obdobnými úpravami dokážeme i poslední tvrzení,

$$\langle \alpha \odot f, g \rangle_B = \left\langle \frac{f^\alpha(t)}{\int_a^b f^\alpha(\xi) d\xi}, g \right\rangle_B =$$

$$\frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{\frac{f^\alpha(t)}{\int_a^b f^\alpha(\xi) d\xi}}{\frac{f^\alpha(s)}{\int_a^b f^\alpha(\xi) d\xi}} \cdot \ln \frac{g(t)}{g(s)} dt\, ds = \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f^\alpha(t)}{f^\alpha(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds,$$

nakonec využijeme toho, že platí $\ln a^c = c \ln a,\ a > 0, c \in \mathbf{R}$, a možnosti vytknutí konstanty před integrál,

$$\frac{1}{2\eta} \int_a^b \int_a^b \ln \left(\frac{f(t)}{f(s)} \right)^\alpha \cdot \ln \frac{g(t)}{g(s)} dt\, ds =$$

$$\alpha \cdot \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f(t)}{f(s)} \cdot \ln \frac{g(t)}{g(s)} dt\, ds = \alpha \cdot \langle f, g \rangle_B .$$

□

Pro další účely si rozepíšeme vztah pro skalární součin z definice 3.3. S využitím pravidla, že logaritmus podílu je roven rozdílu logaritmů, dostáváme

$$\langle f, g \rangle_B = \frac{1}{2\eta} \int_a^b \int_a^b \left[\ln f(t) - \ln f(s) \right] \left[\ln g(t) - \ln g(s) \right] dt\, ds =$$

$$\frac{1}{2\eta}\int_a^b\int_a^b \left[\ln f(t)\ln g(t) - \ln f(t)\ln g(s) - \ln f(s)\ln g(t) + \ln f(s)\ln g(s)\right] dt\, ds.$$

Jelikož platí následující rovnosti

$$\int_a^b\int_a^b \ln f(t)\ln g(t)\, dt\, ds = \int_a^b\int_a^b \ln f(s)\ln g(s)\, dt\, ds$$

a

$$\int_a^b\int_a^b \ln f(t)\ln g(s)\, dt\, ds = \int_a^b\int_a^b \ln f(s)\ln g(t)\, dt\, ds,$$

celkově obdržíme

$$\langle f,g\rangle_B = \frac{1}{2\eta}\left[2\int_a^b\int_a^b \ln f(t)\ln g(t)\, dt\, ds - 2\int_a^b\int_a^b \ln f(t)\ln g(s)\, dt\, ds\right].$$

Nakonec ještě spočteme integrál přes proměnnou s u prvního výrazu,

$$\langle f,g\rangle_B = \frac{1}{\eta}\int_a^b \left[\ln f(t)\ln g(t)s\right]_a^b\, dt - \frac{1}{\eta}\int_a^b \ln f(t)\, dt \int_a^b \ln f(s)\, ds =$$

$$\frac{1}{\eta}\int_a^b \ln f(t)\ln g(t)\,(b-a)\, dt - \frac{1}{\eta}\int_a^b \ln f(t)\, dt \int_a^b \ln f(s)\, ds,$$

a protože $\eta = b-a$, dostáváme konečný tvar úpravy

$$\langle f,g\rangle_B = \int_a^b \ln f(t)\ln g(t)\, dt - \frac{1}{\eta}\int_a^b \ln f(t)\, dt \int_a^b \ln f(s)\, ds. \tag{5}$$

Věta 3.3 $(\mathcal{B}_b^2[a,b], \oplus, \odot)$ *spolu s Bayesovým skalárním součinem tvoří pre-Hilbertův prostor.*

Nakonec si Bayesovův skalární součin dáme do souvislosti se skalárním součinem v L^2. Pro každou $f \in \mathcal{B}^2[a,b]$ lze najít takový násobek $\alpha \in \mathbf{R}$, že $\int_a^b \ln(\alpha f(t)) = 0$. Výběrem reprezentanta, zvolením defakto hodnoty integrálu uvažované hustoty, nedojde ke změně ve výsledku skalárních součinů. Z toho důvodu Bayesův skalární součin odpovídá skalárnímu součinu v L^2 prostoru logaritmů spojitých hustot definovaných na omezeném intervalu $[a,b]$, násobených příslušnou konstantou α.

Definice 3.4 *Nechť $f \in \mathcal{B}^2[a,b]$. Normu f definujeme jako funkcionál $\|\cdot\|_B : \mathcal{B}_b^2[a,b] \to \mathbf{R}$ daný vztahem*

$$\|f\|_B = \sqrt{\langle f, f\rangle_B} = \sqrt{\frac{1}{2\eta}\int_a^b \int_a^b \ln^2 \frac{f(t)}{f(s)} dt\, ds},$$

kde η značí délku intervalu $[a,b]$, tj. $\eta = b - a$.

Vztah pro výpočet normy lze upravit pomocí (5) jako

$$\|f\|_B = \sqrt{\int_a^b \ln^2 f(t)\, dt - \frac{1}{\eta}\left(\int_a^b \ln f(s)\, ds\right)^2}.$$

Definice 3.5 *Nechť $f, g \in \mathcal{B}_b^2[a,b]$. Vzdálenost mezi f a g definujeme jako funkcionál $d_B(\cdot,\cdot) : \mathcal{B}_b^2[a,b] \times \mathcal{B}_b^2[a,b] \to \mathbf{R}$ daný vztahem*

$$d_B(f,g) = \sqrt{\frac{1}{2\eta}\int_a^b \int_a^b \left(\ln \frac{f(t)}{f(s)} - \ln \frac{g(t)}{g(s)}\right)^2 dt\, ds}.$$

Díky normě jsme získali normovaný pre-Hilbertův protor. Aby byl tento prostor úplný, je potřeba určit limitu cauchyovské posloupnosti prvků v $\mathcal{B}_b^2[a,b]$. Za tímto účelem určíme úplnou ortonormální množinu v $\mathcal{B}_b^2[a,b]$, bázi v $\mathcal{B}_b^2[a,b]$. Vzhledem k této bázi pak následně najdeme Fourierovy koeficienty s vlastností konečného součtu druhých mocnin z jejich absolutních hodnot příslušných k prvkům z $\mathcal{B}_b^2[a,b]$. Prvky z $\mathcal{B}_b^2[a,b]$ pak lze vyjádřit jako součet nekonečné Fourierovy řady, čímž se z uvažované báze stane báze (úplného) Hilbertova prostoru. Dokažme proto nyní existenci úplné ortonormální množiny v $\mathcal{B}_b^2[a,b]$.

Věta 3.4 *Nechť množina spojitých funkcí $\{\varphi_j\}_{j\geq 0}$, kde $\varphi_0(t) = \frac{1}{\sqrt{b-a}}, t \in [a,b]$, tvoří ortonormální bázi prostoru $L^2[a,b]$. Potom množina funkcí $\{\psi_j\}_{j\geq 1}$, kde $\psi_j = \mathcal{C}[\exp(\varphi_j)]$, tvoří ortornormální bázi v $\mathcal{B}_b^2[a,b]$.*

Důkaz: Dokážeme, že libovolné dvě hustoty ψ_j a ψ_k jsou ortogonální pro všechna $j \neq k$. Tzn. skalární součin hustot ψ_j a ψ_k pro $j \neq k$ se musí rovnat nule. Skalární součin si rozepíšeme pomocí vztahu (5),

$$\langle \psi_j, \psi_k \rangle_B = \int_a^b \ln \left\{ \mathcal{C} \left[\exp \left(\varphi_j(t) \right) \right] \right\} \cdot \ln \left\{ \mathcal{C} \left[\exp \left(\varphi_k(t) \right) \right] \right\} \, dt$$

$$-\frac{1}{\eta} \int_a^b \ln \left\{ \mathcal{C} \left[\exp \left(\varphi_j(t) \right) \right] \right\} \, dt \cdot \int_a^b \ln \left\{ \mathcal{C} \left[\exp \left(\varphi_k(s) \right) \right] \right\} \, ds,$$

a provedeme operaci uzávěru

$$\int_a^b \ln \frac{\exp \left(\varphi_j(t) \right)}{\int_a^b \exp \left(\varphi_j(u) \right) \, du} \cdot \ln \frac{\exp \left(\varphi_k(t) \right)}{\int_a^b \exp \left(\varphi_k(v) \right) \, dv} \, dt$$

$$-\frac{1}{\eta} \int_a^b \ln \frac{\exp \left(\varphi_j(t) \right)}{\int_a^b \exp \left(\varphi_j(u) \right) \, du} \, dt \cdot \int_a^b \ln \frac{\exp \left(\varphi_k(s) \right)}{\int_a^b \exp \left(\varphi_k(v) \right) \, dv} \, ds.$$

Dále si vzpomeneme na pravidla počítání s logaritmy a následně zvlášť roznásobíme členy rozdílu,

$$\int_a^b \left[\varphi_j(t) - \ln \int_a^b \exp \varphi_j(u) \, du \right] \cdot \left[\varphi_k(t) - \ln \int_a^b \exp \varphi_k(v) \, dv \right] \, dt$$

$$-\frac{1}{\eta} \left[\int_a^b \varphi_j(s) \, ds - \int_a^b \left(\ln \int_a^b \exp \varphi_j(u) \, du \right) \, ds \right]$$

$$\cdot \left[\int_a^b \varphi_k(s) \, ds - \int_a^b \left(\ln \int_a^b \exp \varphi_k(v) \, dv \right) \, ds \right].$$

První člen

$$\int_a^b \varphi_j(t) \varphi_k(t) \, dt - \int_a^b \varphi_j(t) \cdot \left[\ln \int_a^b \exp \varphi_k(v) \, dv \right] \, dt$$

$$- \int_a^b \varphi_k(t) \cdot \left[\ln \int_a^b \exp \varphi_j(u) \, du \right] dt + \int_a^b \left[\ln \int_a^b \exp \varphi_j(u) \, du \cdot \ln \int_a^b \exp \varphi_k(v) \, dv \right] dt$$

upravíme do konečného tvaru

$$\int_a^b \varphi_j(t) \varphi_k(t) \, dt - \ln \int_a^b \exp \varphi_k(v) \, dv \cdot \int_a^b \varphi_j(t) \, dt - \ln \int_a^b \exp \varphi_j(u) \, du \cdot \int_a^b \varphi_k(t) \, dt$$

$$+ (b-a) \ln \int_a^b \exp \varphi_j(u) \, du \cdot \ln \int_a^b \exp \varphi_k(v) \, dv,$$

kde jsme využili toho, že konstantu lze vytknout před integrál, a toho, že $\int_a^b 1dt = b - a$. Provedeme tytéž úpravy i na druhém členu rozdílu,

$$-\frac{1}{\eta}\int_a^b \varphi_j(t)\,dt \cdot \int_a^b \varphi_k(s)\,ds + \frac{1}{\eta}\int_a^b \varphi_j(t)\,dt \cdot \int_a^b \left[\ln \int_a^b \exp\varphi_k(v)\,dv\right] ds$$

$$+\frac{1}{\eta}\int_a^b \varphi_k(s)\,ds \cdot \int_a^b \left[\ln \int_a^b \exp\varphi_j(u)\,du\right] ds$$

$$-\frac{1}{\eta}\int_a^b \left[\ln \int_a^b \exp\varphi_j(u)\,du\right] dt \cdot \int_a^b \left[\ln \int_a^b \exp\varphi_k(v)\,dv\right] dt,$$

celkově tedy pro druhý člen máme

$$-\frac{1}{\eta}\int_a^b \varphi_j(t)\,dt \cdot \int_a^b \varphi_k(s)\,ds + \frac{1}{\eta}\int_a^b \varphi_j(t)\,dt \cdot (b-a)\ln \int_a^b \exp\varphi_k(v)\,dv\,ds$$

$$+\frac{1}{\eta}\int_a^b \varphi_k(s)\,ds \cdot (b-a)\ln \int_a^b \exp\varphi_j(u)\,du\,ds$$

$$-\frac{1}{\eta}(b-a)^2 \ln \int_a^b \exp\varphi_j(u)\,du \cdot \ln \int_a^b \exp\varphi_k(v)\,dv.$$

Nyní se můžeme vrátit k výpočtu skalárního součinu; odečteme-li od sebe upravené členy, dostaneme

$$\langle \psi_j, \psi_k \rangle_B = \int_a^b \varphi_j(t)\varphi_k(t)\,dt - \frac{1}{\eta}\int_a^b \varphi_j(t)\,dt \cdot \int_a^b \varphi_k(s)\,ds = 0, \tag{6}$$

neboť φ_j a φ_k jsou prvky ortonormální báze $L^2[a,b]$ a jsou tedy kolmé a dále přitom uvážíme kolmost φ_0 a φ_j pro $j \geq 1$,

$$\langle \varphi_0, \varphi_j \rangle_2 = \int_a^b \frac{1}{\sqrt{b-a}} \cdot \varphi_j(t)\,dt = \frac{1}{\sqrt{b-a}}\int_a^b \varphi_j(t)\,dt = 0,$$

z čehož vyplývá nulovost integrálu $\int_a^b \varphi_j(t)\,dt$. Tím je kolmost ψ_j a $\psi_k \in \mathcal{B}_b^2[a,b]$ dokázána a $\{\psi_j\}_{j\geq 1}$ tvoří ortogonální množinu.

Nakonec s využitím (6) ukážeme, že ψ_j jsou normované,

$$\|\psi_j\|_B = \sqrt{\langle \psi_j, \psi_j \rangle_B} = \sqrt{\int_a^b \varphi_j(t)\varphi_j(t)\,dt - \frac{1}{\eta}\int_a^b \varphi_j(t)\,dt \cdot \int_a^b \varphi_j(s)\,ds}$$

$$= \sqrt{\int_a^b \varphi_j^2(t)\, dt} = 1.$$

Tedy $\{\psi_j\}_{j\geq 1}$ je ortonormální v $\mathcal{B}_b^2[a,b]$, což jsme chtěli dokázat. □

Věta 3.5 *Ortonormální množina $\{\psi_j\}_{j\geq 1}$ v $\mathcal{B}_b^2[a,b]$ za předpokladů z věty 3.4 je úplná.*

Důkaz: Pokud je uvažovaná množina úplná, tak nelze najít prvek z $\mathcal{B}_b^2[a,b]$ kromě $e = \frac{1}{b-a}$, který by byl kolmý ke všem prvkům z $\{\psi_j\}_{j\geq 1}$. Důkaz povedeme sporem a budeme předpokládat, že existuje $w \in \mathcal{B}_b^2[a,b]$ různý od $e = \frac{1}{b-a}$, který je kolmý na ψ_j, tj. $\langle w, \psi_j\rangle_B = 0$. Nyní využijeme faktu, že platí $\langle w, \psi_j\rangle_B = \langle \ln w, \varphi_j\rangle_2$, o čemž se přesvědčíme později. Uvážíme-li, že $\{\varphi_j\}_{j\geq 0}$ tvoří bázi v $L^2[a,b]$, tak pouze prvky $z \in L^2[a,b]$, které jsou proporcionální k φ_0, splňují kolmost k prvkům $\varphi_j, j \geq 1$, tj. $\langle z, \varphi_j\rangle_2 = \langle \alpha\varphi_0, \varphi_j\rangle_2 = 0$. Odtud

$$w = \mathcal{C}(\exp(\alpha\varphi_0)) = \frac{\exp(\alpha\varphi_0)}{\int_a^b \exp(\alpha\varphi_0)\, ds} = \frac{\exp(\alpha\varphi_0)}{\exp(\alpha\varphi_0)\int_a^b 1\, ds} = \frac{1}{b-a} = e,$$

což je spor s tím, že jsme předpokládali w různé od e. □

3.1.2. Hilbertův prostor $\mathcal{B}^2[a,b]$

Jak už víme, vynásobením spojité kompozice f konstantou $\alpha \in \mathbf{R}^+$ nedojde ke změně relativní informace. Proto za účelem získání Hilbertovy struktury prostoru $\mathcal{B}_b^2[a,b]$ budeme nyní pracovat se třídami kladných funkcí na $[a,b]$. Přitom o dvou kladných funkcích f a g řekneme, že jsou ekvivalentní, jestliže jsou proporcionální, $f = \alpha g, \alpha \in \mathbf{R}^+$. Třídy takových funkcí definujeme následovně.

Definice 3.6 *Třída funkcí f daných zobrazením $z\,[a,b] \to \mathbf{R}$ je v $\mathcal{B}^2[a,b]$, platí-li následující podmínky,*

1. $f > 0$ *na* $[a,b]$,
2. $\ln f \in L^2[a,b]$.

V případě, že $\int_a^b f(t)\,dt$ konverguje, pak za reprezentanta dané třídy bereme $\mathcal{C}(f)$ s jednotkovým integrálem. Jinak, v případě divergence integrálu, bereme takového, který splňuje $\int_a^b \ln f(t)\,dt = 0$.

Poznámka 3.2 Pro zjednodušení budeme dále třídy funkcí z definice 3.6 nazývat pouze funkcemi v $\mathcal{B}^2[a,b]$.

Nyní se pokusíme funkcím z $\mathcal{B}^2[a,b]$ přiřadit posloupnost čísel z l^2 (s vlastností konečného součtu jejich druhých mocnin). Za tímto účelem dokážeme existenci Fourierových koeficientů funkcí v $\mathcal{B}^2[a,b]$ a konvergenci součtu jejich druhých mocnin.

Věta 3.6 *Nechť $\{\psi_j\}_{j\geq 1}$ je úplná množina v $\mathcal{B}^2[a,b]$ generovaná ortonormální bází $\{\varphi_j\}_{j\geq 0}$ prostoru $L^2[a,b]$ definované v souladu s větou* 3.4. *Jestliže funkce $f \in \mathcal{B}^2[a,b]$, pak platí, že*

1. $\langle f, \psi_j\rangle_B = \langle \ln f, \varphi_j\rangle_2$,
2. $\sum_{j=1}^{\infty} \left|\langle f, \psi_j\rangle_B\right|^2 < \infty$.

Důkaz: Z toho, že je funkce $\ln f$ integrovatelná s kvadrátem na $[a,b]$, plyne, že je absolutně integrovatelná na $[a,b]$, $\ln f \in L^1[a,b]$, tj. $\int_a^b |\ln f(t)|\,dt < \infty$, a tedy $\int_a^b \ln f(t)dt < \infty$. Pak s využitím (5) rozepíšeme Bayesův skalární součin jako

$$\langle f, \psi_j\rangle_B = \int_a^b \ln f(t)\cdot \ln \psi_j(t)\,dt - \frac{1}{\eta}\int_a^b \ln f(t)\,dt \cdot \int_a^b \ln \psi_j(s)\,ds.$$

Jelikož $\int_a^b \ln f(t)\cdot \ln \psi_j(t)\,dt = \langle \ln f, \ln \psi_j\rangle_2$, provedením operace uzávěru a následnými úpravami dostáváme

$$\left\langle \ln f, \ln \frac{\exp(\varphi_j(t))}{\int_a^b \exp(\varphi_j(u))\,du}\right\rangle_2 - \frac{1}{\eta}\int_a^b \ln f(t)\,dt \cdot \int_a^b \ln \frac{\exp(\varphi_j(t))}{\int_a^b \exp(\varphi_j(u))\,du}\,dt$$

$$= \left\langle \ln f, \varphi_j - \ln \int_a^b \exp(\varphi_j(u))\,du\right\rangle_2$$

$$-\frac{1}{\eta}\left[\int_a^b \ln f(t)\,dt\right]\cdot\left[\int_a^b \varphi_j(t)\,dt-(b-a)\ln\int_a^b \exp\left(\varphi_j(u)\right)\,du\right].$$

Výše uvedený skalární součin upravíme následovně,

$$\left\langle \ln f,\varphi_j-\ln\int_a^b \exp\left(\varphi_j(u)\right)\,du\right\rangle_2$$

$$=\int_a^b\left[\ln f(t)\varphi_j(t)-\ln f(t)\cdot\ln\int_a^b \exp\left(\varphi_j(u)\right)\,du\right]\,dt$$

$$=\int_a^b \ln f(t)\varphi_j(t)\,dt-\left[\ln\int_a^b \exp\left(\varphi_j(u)\right)\,du\right]\cdot\int_a^b \ln f(t)\,dt.$$

A konečně s využitím této úpravy dostáváme

$$\langle f,\psi_j\rangle_B=\langle \ln f,\varphi_j\rangle_2-\left[\ln\int_a^b \exp\left(\varphi_j(u)\right)\,du\right]\cdot\int_a^b \ln f(t)\,dt$$

$$-\left[\int_a^b \ln f(t)\,dt\right]\cdot\left[\frac{1}{\eta}\int_a^b \varphi_j(t)\,dt-\ln\int_a^b \exp\left(\varphi_j(u)\right)\,du\right]=\langle \ln f,\varphi_j\rangle_2,$$

kdy jsme v posledním kroku využili toho, že pro $j\geq 1$ máme $\langle\varphi_0,\varphi_j\rangle_2=0$. Bayesův skalární součin $\langle f,\psi_j\rangle_B$ odpovídá Fourierovým koeficientům funkce $\ln g$ v $L^2[a,b]$ vzhledem k ortonormální bázi $\{\varphi_j\}_{j\geq 0}$, z čehož vyplývá, že součet jejich druhých mocnin konverguje. □

Přiřazení těchto Fourierových koeficientů k funkcím z $\mathcal{B}^2[a,b]$ lze definovat následovně.

Definice 3.7 *Nechť $f\in\mathcal{B}^2[a,b]$ a $\ln f(t)=\sum_{k=1}^{\infty}\alpha_k\varphi_k(t)$ s $\sum_{j=1}^{\infty}|\alpha_k|^2<\infty$. Funkci koeficientů T danou zobrazením $T:\mathcal{B}^2[a,b]\to l^2$ definujeme jako $Tf=\{\alpha_k\}_{k\geq 1}$.*

Uvědomme si, že zobrazením T přiřazujeme všem nezáporným funkcím na $[a,b]$ v rámci tříd stejnou množinu koeficientů, tedy je přiřazujeme všem třídám v $\mathcal{B}^2[a,b]$. Proto pomocí tohoto přiřazení můžeme definovat operace pertubace, mocninnou transformaci a skalární součin na $\mathcal{B}^2[a,b]$ následovně.

Definice 3.8 *Mějme dány funkce* $f, g \in \mathcal{B}^2[a,b]$ *a* $\alpha \in \mathbf{R}$*. Operace pertubace a mocninná transformace definujeme v* $\mathcal{B}^2[a,b]$ *jako*

$$f \oplus g = T^{-1}(Tf + Tg), \quad \alpha \odot f = T^{-1}(\alpha \cdot Tf),$$

kde operace $(+)$ *a* $(\cdot)$ *jsou standardními operacemi sčítaní a násobení reálnou konstantou v prostoru* l^2*. Skalární součin v* $\mathcal{B}^2[a,b]$ *obdržíme jako*

$$\langle f, g \rangle_B = \langle Tf, Tg \rangle_2,$$

kde $\langle \cdot, \cdot \rangle_2$ *odpovídá skalárnímu součinu v prostoru* l^2*. Následně tak dostaneme i normu v* $\mathcal{B}^2[a,b]$ *jako*

$$\|f\|_B = \sqrt{\langle f, f \rangle_B} = \sqrt{\langle Tf, Tf \rangle_2} = \|Tf\|_2,$$

kde $\|\cdot\|_2$ *značí standarní normu v* l^2 *(analogicky též pro vzdálenost dvou hustot z* $\mathcal{B}^2[a,b]$*.*

Zobrazení T je izomorfismus mezi $\mathcal{B}^2[a,b]$ a l^2, které zachovává algebraické operace, normu i skalární součin. Uvážením věty 1.2 můžeme zformulovat následující tvrzení.

Věta 3.7 $\mathcal{B}^2[a,b]$ *spolu s operacemi z definice* 3.8 *tvoří separabilní Hilbertův prostor.*

Úplná ortonormální množina $\{\psi_k\}_{k\geq 1}$ v $\mathcal{B}_b^2[a,b]$ je tedy ortonormální bází v $\mathcal{B}^2[a,b]$. Z toho vyplývá, že funkce f z $\mathcal{B}^2(a,b)$, jejichž logaritmus je roven součtu své příslušné Fouerierovy řady vzhledem k ortonormální bázi $\{\varphi_k\}_{k\geq 0}$ v $L^2[a,b]$, $\ln f = \sum_{k=1}^{\infty} \alpha_k \varphi_k$ s $\sum_{k=1}^{\infty} |\alpha_k|^2$, lze tedy vyjádřit jako pertubaci Fourierovy řady vzhledem k bázi $\{\psi_j\}_{j\geq 1}$, tj. $f = \bigoplus_{k=1}^{\infty} (\alpha_k \odot \psi_k)$, která konverguje vždy ve smyslu Bayesovy normy $\|\cdot\|_B$.

Závěrem lze konstatovat, že množina hustot rozdělení pravděpodobností, jejichž druhá mocnina logaritmu je integrovatelná, je zahrnuta v separabilním Hilbertově prostoru kladných funkcí, které jsou integrovatelné s kvadrátem.

4. CLR transformace

Podobně jako v případě diskrétních kompozic, chceme i zde mít transformaci, izometrické zobrazení z Bayesova prostoru do standardního L^2 prostoru, které nám převede operace pertubace a mocninnou transformaci na standardní sčítání a násobení. Bez této transformace bychom totiž nemohli pro statistické zpracování hustot použít standardní statistické metody, které byly za tohoto předpokladu vytvořeny. V této práci budeme pracovat s centrovanou log ratio (clr) transformací [8], funkcionální obdobou známé centrované logratio transformace pro diskrétní kompozice [1]; klíčem k jejímu zavedení je důkaz věty 3.4.

Funkcionální clr transformace je pro hustotu $f(t) \in \mathcal{B}^2(I)$, spojitou kladnou funkci s oborem hodnot na I a jednotkovým integrálem, definovaná jako

$$\operatorname{clr}[f(t)] = f_c(t) = \ln f(t) - \frac{1}{\eta}\int_I \ln f(t)\,dt, \tag{7}$$

kde η jako obvykle značí délku intervalu I, tj. $\eta = b - a$. Z konstrukce clr transformace plyne nulovost integrálu clr transformovaných hustot

$$\int_I \operatorname{clr}[f(t)]\;dt = \int_I \ln f(t)\,dt - \int_I \frac{1}{\eta}\int_I \ln f(s)\,ds\,dt$$

$$= \int_I \ln f(t)\,dt - \int_I \ln f(s)\,ds = 0, \tag{8}$$

která nesmí být opomíjena při jejich statistické analýze. Naštěstí se ukazuje, že dodatečná podmínka nečiní větší problémy ve statistických metodách založených na vzdálenostech, stejně tak např. ve funkcionální metodě hlavních komponent, kterou si představíme v následující kapitole. Zmíněná transformace má následující vlastnosti. Pro $f, g \in \mathcal{B}^2(I)$ a $\alpha \in \mathbf{R}$ platí

1. $\operatorname{clr}(f \oplus g)(t) = f_c(t) + g_c(t)$,
2. $\operatorname{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t)$,
3. $\langle f, g\rangle_B = \langle f_c, g_c\rangle_2 = \int_I f_c(t)g_c(t)\,dt$.

Protože clr představuje prosté zobrazení, lze její inverzní transformaci vyjádřit jako

$$\mathrm{clr}^{-1}\left[f_c(t)\right] = \frac{\exp\left[f_c(t)\right]}{\int_I \exp\left[f_c(t)\right]\ dt}, \tag{9}$$

kde, jak už víme, jmenovatel, který nám zaručuje jednotkový integrál výsledné hustoty, je zde spíše z praktických důvodů než z teoretické potřeby.

4.1. Efekt clr transformace

Abychom viděli efekt použití clr transformace, zvolíme tři rozdělení patřící do rodiny exponenciálních rozdělení [7], které budeme uvažovat vždy na nějakém uzavřeném intervalu I. Jako první se podíváme na soubor beta hustot s parametry $\alpha_i = 2 + \frac{i}{3}$ a $\beta_i = 2 + \frac{i}{3}$ pro $i = 1, \ldots, 15$ na $I = [0.1, 0.9]$,

$$f(t; \alpha_i, \beta_i) =_{B^2} t^{\alpha_i - 1}\left(1 - t\right)^{\beta_i - 1}, \quad t \in I, \tag{10}$$

kde $_{B^2}$ značí ekvivalenci v prostoru $\mathcal{B}^2(I)$. Není zde nutné uvádět normalizační konstantu vyjádřenou pomocí beta funkce B jako $\frac{1}{B(\alpha,\beta)}$ a upravenou pro uvažovaný tvar na I, neboť jejím neuvedením se nezmění relativní informace nesená jednotlivými hustotami $f(t; \alpha_i, \beta_i)$. Příslušné clr transformace získáme pomocí (7),

$$f_c(t; \alpha_i, \beta_i) = \ln\left[t^{\alpha_i - 1}\left(1 - t\right)^{\beta_i - 1}\right] - \frac{1}{\eta}\int_I \ln\left[t^{\alpha_i - 1}\left(1 - t\right)^{\beta_i - 1}\right]\ dt, \quad t \in I.$$

Původní hustoty můžeme vidět na obrázku 3a, transformované pak v 3b. Pokud bychom chtěli učinit závěry o variabilitě v datech, kterou bychom měřili ve smyslu L^2 prostorů, z obrázku 3a je patrné, že největší rozdíly v hodnotách hustot bychom spatřovali pro $t \in [0.4, 0.6]$. Ke stejnému závěru bychom dospěli i spočtením příslušné kovarianční funkce (obrázek 3c), kdy největší variabilita dat, kterou vyčteme z úhlopříčky vedoucí z levého dolního rohu do pravého horního, přísluší stejnému intervalu $[0.4, 0.6]$. Tyto závěry však zřejmě neodpovídají skutečnosti, neboť relativní variabilita jednotlivých spojitých náhodných veličin

je zde zřejmě řízena realizacemi veličin odpovídající krajním hodnotám intervalu I. Vezmeme-li při posuzování variability hustot v potaz vlastnost relativního měřítka, která je zřejmá z porovnání obrázků 3a a 3b, nárust hodnot hustot na koncích intervalu I, mající téměř nulové absolutní příspěvky na celkovou pravděpodobnost, je v jejím důsledku mnohem významnější. To je jasně zachyceno clr transformací na obrázku 3b, kdy největší rozdíl v hodnotách clr hustot můžeme vidět na koncích intervalu I, což je jasně zachyceno i kovarianční funkcí na obrázku 3d. Je tomu tak proto, že clr hustoty jsou prvky prostoru L^2, a proto může být jejich variabilita v clr prostoru posuzována standardně ve smyslu hledání největších rozdílů ve funkčních hodnotách.

Analyzujme nyní stejným způsobem soubor chí-kvadrát hustot se stupni volnosti $n_i = i + 2$ pro $i = 1, \ldots, 19$ na intervalu $I = [e^{-7}, e^{3.5}]$ (obrázek 4a),

$$f(t, n_i) =_{B^2} e^{-\frac{t}{2}} t^{\frac{n_i}{2}-1}, \quad t \in I. \tag{11}$$

Příslušné clr transformace

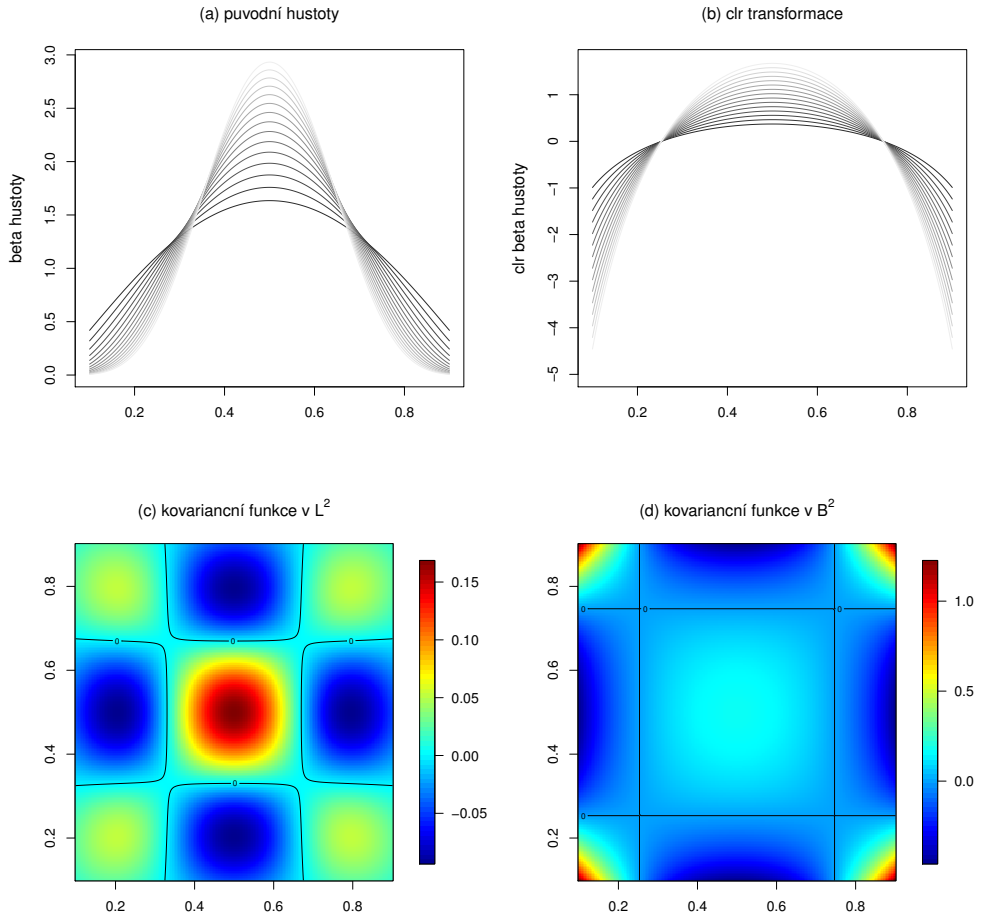
$$f_c(t; n_i) = \ln\left[e^{-\frac{t}{2}} t^{\frac{n_i}{2}-1}\right] - \frac{1}{\eta}\int_I \ln\left[e^{-\frac{t}{2}} t^{\frac{n_i}{2}-1}\right] dt, \quad t \in I.$$

pak můžeme vidět na obrázku 4b. Stejně jako v předchozím případě je variabilita ve smyslu $\mathcal{B}^2$ prostorů v důsledku vlastnosti relativního měřítka řízena malými hodnotami hustot nalevo, které jsou následkem strmějšího poklesu v jejich hodnotách v porovnání s malými hodnotami napravo daleko významnějším zdrojem relativní variability. Naopak z pohledu L^2 prostorů na tytéž hustoty je ona variabilita řízena velkými hodnotami hustot v jejich vrcholcích. To je zřejmé i z porovnaní odhadnutých kovariančních funkcí v $\mathcal{B}^2(I)$ (obrázek 4d-4e) a v $L^2(I)$ (obrázek 4c).

Nakonec se podíváme ještě na soubor exponenciálních hustot s parametrem $\lambda_i = \frac{i}{4}$ pro $i = 1, \ldots, 15$ uvažovaných na $I = [0, 10]$,

$$f(t, \lambda_i) =_{B^2} e^{-\lambda_i t}, \quad t \in I,$$

kde příslušné clr transformace (obrázek 5b) odpovídají lineárním funkcím daných



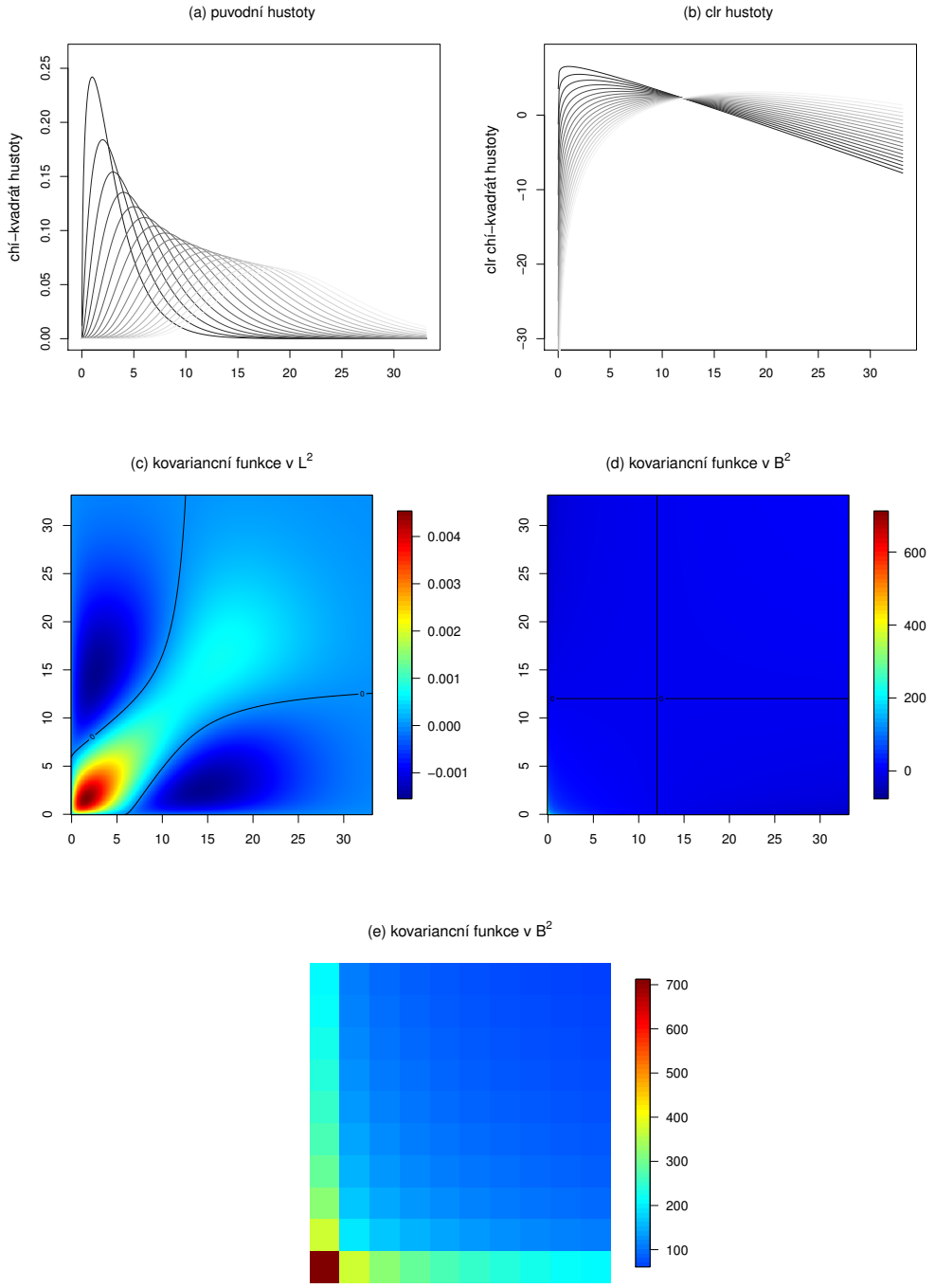
Obrázek 3: Beta hustoty na $I = [0.1, 0.9]$ s parametry $\alpha_i = \beta_i = 2 + \frac{i}{3}$ pro $i = 1, \ldots, 15$.

předpisem

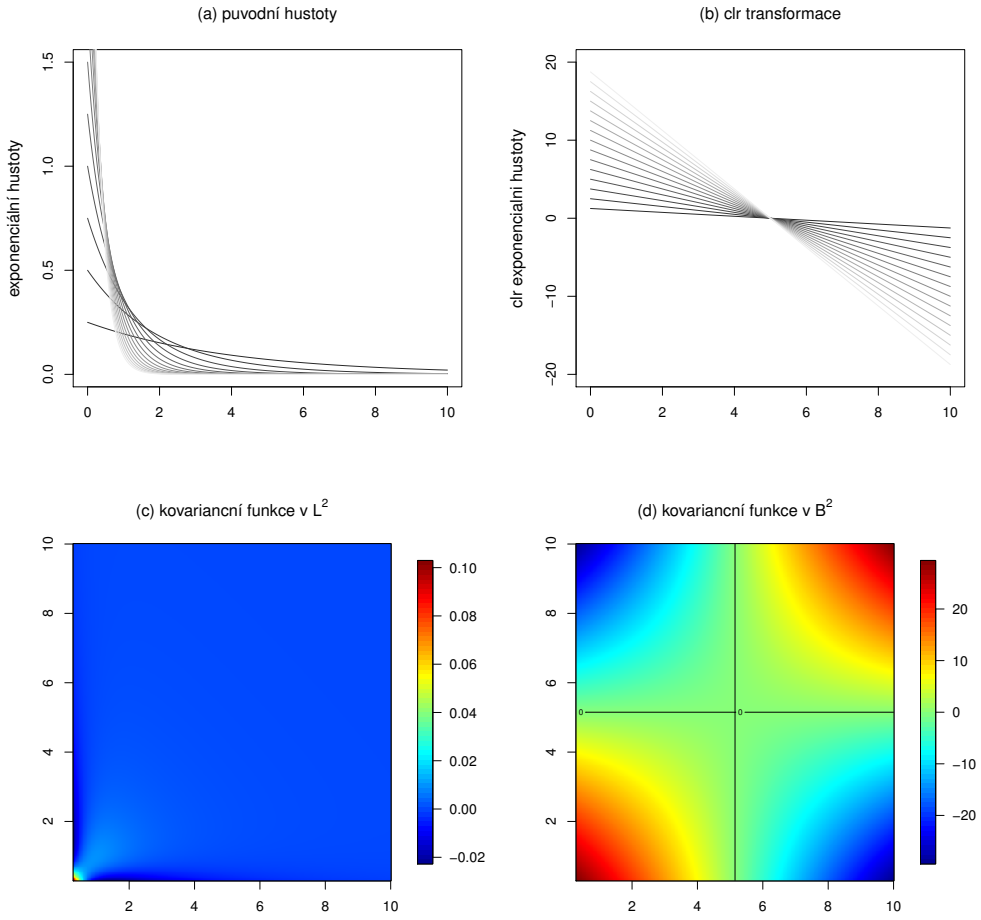
$$f_c(t; \lambda_i) = \ln\left[e^{-\lambda_i t}\right] - \frac{1}{10}\int_I \ln\left[e^{-\lambda_i t}\right]\, dt = -\lambda_i t - \frac{\lambda_i}{20}\left[t^2\right]_0^{10}$$

$$= -\lambda_i t - 5\lambda_i, \quad t \in I.$$

Zdrojem relativní variability jsou jak malé hodnoty hustot napravo, tak i velké hodnoty vlevo, kde můžeme vypozorovat jejich výrazný strmý pokles (obrázek 5d). Naopak zdrojem absolutní variability jsou hodnoty hustot odpovídající začátku intervalu I (obrázek 5e).



Obrázek 4: Chí-kvadrát hustoty na $I = [e^{-7}, e^{3.5}]$ s parametrem $n_i = i + 2$ pro $i = 1, \ldots, 19$.



Obrázek 5: Exponenciální hustoty na $I = [0, 10]$ s parametrem $\lambda_i = \frac{i}{4}$ pro $i = 1, \ldots, 15$.

V následující kapitole si představíme funkcionální metodu hlavních komponent, která bude založena na geometrii Bayesových prostorů. Ta by měla v důsledku vlastnosti relativního měřítka při aplikování na zmíněné soubory hustot zachytit relativní zdroj variability, a tedy by měla podpořit závěry, které jsme učinili v této kapitole.

Než ale přejdeme k samotné statistické metodě, je potřeba v problematice reprezentace dat, tedy clr trasformovaných hustot, zohlednit podmínku (8). Stejně jako obecná funkcionální data, tak i hustoty získáváme při měření reálných jevů v diskrétní formě. Například získaná diskrétní data, která si vykreslíme pomocí

histogramu, je potřeba dále aproximovat příslušnou funkcí. Použijeme-li při jejich reprezentaci přístupy z podkapitoly 2.2, při interpolaci či aproximaci clr trasformovaných hustot musí výsledný inteprolační či vyhlazující splajn splňovat navíc podmínku

$$\int_a^b s_k(t) = 0. \tag{12}$$

Detailní postup výpočtu B-splajnových koeficientů zohledňující (12) je popsán v [8]. Důležitým faktem, který je v tomto článku zmíněn, je nulový součet získaných B-splajnových koeficientů, tj. ve vztazích (1) a (4) platí podmínka

$$\sum_{i=-k}^{g} b_i = 0, \tag{13}$$

která je známá z clr transformace diskrétních kompozic.

5. Metoda hlavních komponent

Metoda hlavních komponent (PCA) [14], [15] patří k nejpoužívanějším mnohorozměrným statistickým metodám. Její hlavní myšlenkou je redukce dimenze dat, kdy chceme co nejvíce z původní informace, která je reprezentována p dimenzemi, zachytit pouze pomocí dimenze l, tak že $l << p$. Metoda hlavních komponent se jeví výhodná zejména při velkém počtu proměnných, které jsou vzájemně závislé. V kapitole si nejdříve představíme PCA pro mnohorozněrná data, která nám napomůže k lepšímu pochopení fungování funkcionální PCA [12].

5.1. PCA pro mnohorozměrná data

PCA vede ke vzniku nových veličin, hlavních komponent, které vysvětlují variabilitu a závislost původních proměnných. Hlavní komponenty získáme jako lineární kombinace původních proměnných, které jsou vytvářeny postupně tak, aby první hlavní komponenta vysvětlila co nejvíce z celkové variability, druhá komponenta pak nejvíce ze zbývající variability, až poslední, která za předpokladu možné redukce dimenze dat už nemusí vysvětlit žadnou. Z toho důvodu pro účely další statistické analýzy bereme pouze prvních pár komponent, zpravidla se jedná o dvě až čtyři hlavní komponenty.

Předpokládejme, že máme náhodný vektor $\mathbf{x} = (X_1, \ldots, X_n)'$ z p-rozměrného rozdělení s vektorem středních hodnot $\boldsymbol{\mu}$ a s varianční maticí $\boldsymbol{\Sigma}$. Bez újmy na obecnosti budeme předpokládat, že $\boldsymbol{\mu} = \mathbf{0}$. Prvním krokem je nalézt takový vektor koeficientů $\boldsymbol{\xi}_1$, pro který bude mít lineární kombinace původních proměnných maximální rozptyl,

$$y_1 = \xi_{11}X_1 + \xi_{12}X_2 + \ldots + \xi_{1p}X_p = \xi_1'\mathbf{x}, \tag{14}$$

přičemž je potřeba přidat omezující podmínku na koeficienty $\boldsymbol{\xi}_1'\boldsymbol{\xi}_1 = 1$. V j-tém kroku pak hledáme takovou lineární kombinaci

$$y_j = \xi_{j1}X_1 + \xi_{j2}X_2 + \ldots + \xi_{jp}X_p = \xi_j'\mathbf{x},$$

jejíž rozptyl je maximální, platí $\boldsymbol{\xi}_j'\boldsymbol{\xi}_j = 1$ a navíc $\boldsymbol{\xi}_j'\boldsymbol{\xi}_k = 0$ pro $k = 1, \ldots, j-1$.

Jelikož $\text{var}(y_1) = \text{var}\left(\boldsymbol{\xi}_1'\mathbf{x}\right) = \boldsymbol{\xi}_1'\text{var}\left(\mathbf{x}\right)\boldsymbol{\xi}_1 = \boldsymbol{\xi}_1'\boldsymbol{\Sigma}\boldsymbol{\xi}_1$, můžeme maximalizační úlohu s podmínkami (14) formulovat jako

$$\max_{\boldsymbol{\xi}_1'\boldsymbol{\xi}_1=1} \boldsymbol{\xi}_1'\boldsymbol{\Sigma}\boldsymbol{\xi}_1, \tag{15}$$

jejíž řešení získáme pomocí Lagrangeovy metody. Příslušná Lagrangeova funkce je ve tvaru

$$L_1 = \boldsymbol{\xi}_1'\boldsymbol{\Sigma}\boldsymbol{\xi}_1 - \lambda_1\left(\boldsymbol{\xi}_1'\boldsymbol{\xi}_1 - 1\right)$$

s Langrangeovým multiplikátorem λ_1. Spočteme parciální derivace funkce L_1 podle $\boldsymbol{\xi}_1$, které položíme rovny nule

$$\frac{\partial L_1}{\partial \boldsymbol{\xi}_1} = 2\boldsymbol{\Sigma}\boldsymbol{\xi}_1 - 2\lambda_1\boldsymbol{\xi}_1 = 0,$$

což můžeme následně upravit do tvaru

$$\boldsymbol{\Sigma}\boldsymbol{\xi}_1 = \lambda_1\boldsymbol{\xi}_1. \tag{16}$$

Z tvaru (16) vyplývá, že $\boldsymbol{\xi}_1$ je vlastní vektor varianční matice $\boldsymbol{\Sigma}$ odpovídající vlastnímu číslu λ_1, které s ohledem na

$$\text{var}(y_1) = \boldsymbol{\xi}_1'\boldsymbol{\Sigma}\boldsymbol{\xi}_1 = \boldsymbol{\xi}_1'\lambda_1\boldsymbol{\xi}_1 = \lambda_1\boldsymbol{\xi}_1'\boldsymbol{\xi}_1 = \lambda_1$$

bude odpovídat největšímu vlastnímu číslu. Stejně tak pro j-tou komponentu dostaneme, že $\boldsymbol{\xi}_j$ bude odpovídat vlastnímu vektoru varianční matice $\boldsymbol{\Sigma}$ příslušícímu j-tému největšímu vlastnímu číslu λ_j, tj. $\lambda_1 \geq \ldots \geq \lambda_j \geq \ldots \geq \lambda_p$. Celkový rozptyl původních proměnných je roven celkovému rozptylu hlavních komponent,

$$\text{var}\left(X_1\right) + \ldots + \text{var}\left(X_p\right) = \lambda_1 + \ldots + \lambda_p.$$

Navíc

$$\text{cov}(y_i, y_j) = \text{cov}\left(\boldsymbol{\xi}_i'\mathbf{x}, \boldsymbol{\xi}_j'\mathbf{x}\right) = \boldsymbol{\xi}_i'\text{var}\left(\mathbf{x}\right)\boldsymbol{\xi}_j = \boldsymbol{\xi}_i'\Sigma\boldsymbol{\xi}_j = \boldsymbol{\xi}_i'\lambda_j\boldsymbol{\xi}_j = \lambda_j\boldsymbol{\xi}_i'\boldsymbol{\xi}_j = 0$$

pro $i \neq j$, tedy hlavní komponenty splňují podmínku nekorelovanosti.

Uvažujme nyní redukci dimenze z p na l. Souřadnice každého z n objektů v prostoru hlavních komponent nazýváme skóry a pro i-tý objekt je určíme jako $y_{i1} = \boldsymbol{\xi}_1' \mathbf{x}_i, \ldots, y_{il} = \boldsymbol{\xi}_l' \mathbf{x}_i$. Spočtením skórů pro všechny objekty a jejich následným uspořádáním do matice získáme matici skórů $\boldsymbol{\Psi}_{n \times l}$. Vlastní vektory $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_l$ příslušné varianční matici $\boldsymbol{\Sigma}$ nazýváme zátěžemi. Ty určují směry hlavních komponent (směry největší variability v datech).

Abychom mohli řešit uvedenou maximalizační úlohu, je potřeba určit odhad $\boldsymbol{\Sigma}$. Předpokládejme, že máme k dispozici centrovanou datovou matici $\mathbf{X}_{n \times p} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, která odpovídá náhodnému výběru o rozsahu n s p proměnnými. Odhad $\boldsymbol{\Sigma}$ určíme jako výběrovou varianční matici $\mathbf{V}$ danou vztahem

$$\mathbf{V} = \frac{1}{n-1} \mathbf{X}'\mathbf{X}.$$

Maximalizační úlohu přeformulujeme následovně. Úkolem je nalézt směr největší variability $\boldsymbol{\xi}_1$ mezi všemi možnými směry $\boldsymbol{\xi}$ tak, že maximalizujeme výraz

$$\max_{\|\boldsymbol{\xi}\|^2 = 1} \boldsymbol{\xi}' \mathbf{V} \boldsymbol{\xi} = \max_{\|\boldsymbol{\xi}\|^2 = 1} \frac{1}{n-1} \boldsymbol{\xi}' \mathbf{X}' \mathbf{X} \boldsymbol{\xi} = \max_{\|\boldsymbol{\xi}\|^2 = 1} \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{\xi}' \mathbf{x}_i \right)^2$$

$$= \max_{\|\boldsymbol{\xi}\|^2 = 1} \frac{1}{n-1} \sum_{i=1}^{n} \left\langle \mathbf{x}_i, \boldsymbol{\xi} \right\rangle^2. \tag{17}$$

V j-tém kroku pak řešíme stejnou úlohu, ale navíc přidáváme podmínku na ortogonalitu $\boldsymbol{\xi}' \boldsymbol{\xi}_k = 0$, pro $k = 1, \ldots, j-1$. Jak bylo vysvětleno dříve, $\boldsymbol{\xi}_j$ získáme řešením rovnice

$$\frac{1}{n-1} \mathbf{X}'\mathbf{X} \boldsymbol{\xi} = \rho \boldsymbol{\xi}, \tag{18}$$

které odpovídá vlastnímu vektoru matice $\mathbf{V}$ příslušnému j-tému největšímu vlastnímu číslu ρ_j. Podíl vysvětlené variability j-tou komponentou určíme jako $\frac{\rho_j}{\sum_{i=1}^{p} \rho_i}$.

Obecně má matice $\mathbf{V}$ má nejvýše $\min\{p, n\}$ nenulových vlastních čísel ρ_j, v případě centrování (odečtení příslušných výběrových průměrů od sloupců $\mathbf{X}$),

pro $p > n-1$ je hodnost matice $\mathbf{X}$ je nejvýše $n-1$, a proto maximální počet nenulových vlastních čísel $\mathbf{V}$ je roven $\min\{p, n-1\}$. Existuje několik výpočetních algoritmů, které řeší (18), o těch se více můžete dočíst například v citované literatuře.

Nakonec ještě poznamenejme, že hlavní komponenty nejsou určeny jednoznačně. Lze například změnit znaménko vlastního vektoru $\boldsymbol{\xi}$ beze změny podílu vysvětlené variability příslušnou hlavní komponentou.

5.2. PCA pro funkcionální data (FPCA)

Předpokladájme, že máme k dispozici funkcionální náhodný výběr $X_1, \ldots, X_n$ v $L^2(I), I \subset \mathbf{R}$. Opět bez újmy na obecnosti budeme předpokládat, že střední funkce $\mu(t) = 0, t \in I$. Cílem FPCA je opět původní soubor pozorování (funkcí) nahradit menším počtem funkcí tak, abychom vysvětlili co možno nejvíce variability v datech. Za tímto účelem se v prvním kroce FPCA snažíme najít takovou funkci $\xi_1 \in L^2(I)$, první funkcionální hlavní komponentu, která podobně jako v (17) maximalizuje přes všechny možné $\xi \in L^2(I)$ výraz

$$\frac{1}{n-1} \sum_{i=1}^{n} \langle X_i, \xi \rangle_2^2 , \tag{19}$$

za podmínky $\|\xi\|_2^2 = \sqrt{\int_I |\xi(t)|^2 \, dt} = 1$, $t \in I$. Další hlavní komponenty $\{\xi_j\}_{j>1}$ pak mají vysvětlit co nejvíce ze zbylé variability v datech. Z toho důvodu řeší úlohu (19) s přidáním další podmínky na jejich vzájemnou ortogonalitu $\langle \xi, \xi_k \rangle_2 = \int_I \xi(t)\xi_k(s) \, dt \, ds = 0$, $t, s \in I, k < j$.

Analogicky s mnohorozměrnou PCA, funkcionální hlavní komponenty $\{\xi_j\}_{j\geq 1}$ odpovídají vlastním funkcím kovariančního operátoru $V : L^2(I) \to L^2(I)$ definovaného pro $x \in L^2(I)$ jako

$$Vx = \frac{1}{n-1} \sum_{i=1}^{n} \langle X_i, \xi \rangle_2 \, X_i,$$

nebo ekvivalentně jako

$$Vx = \int_I v(\cdot, t)x(t)\, dt,$$

kde $v(\cdot, t)$ je výběrová kovarianční funkce. Proto ξ_j získáme řešením rovnice

$$V\xi_j = \rho_j \xi_j, \tag{20}$$

kde ρ_j je vlastní číslo příslušné k vlastní funkci ξ_j s vlastností $\rho_1 \geq \ldots \geq \rho_j \geq \ldots$. Skóry pro i-té pozorování určíme pro $j = 1, 2, \ldots$ jako $\Psi_{ij} = \langle X_i, \xi_j \rangle_2$.

Opět existuje několik výpočetních algoritmů vedoucích k řešení (20). Uvedeme si zde jeden, kterým lze rovnici (20) převést na úlohu hledání vlastních vektorů příslušné matice koeficientů. Tento přístup předpokládá, že každé pozorování X_i lze rozvinout pomocí K známých bázických funkcí ϕ_k

$$X_i(\cdot) = \sum_{k=1}^{K} c_{ik}\phi_k, \tag{21}$$

kde $c_{ik} = \langle X_i, \phi_k \rangle_2$, $k = 1, \ldots, K$ jsou příslušné koeficienty i-tého pozorování. Označíme-li $\mathbf{X}(\cdot) = (X_1(\cdot), \ldots, X_n(\cdot))'$ a $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_K(\cdot))'$, lze (21) maticově vyjádřit jako $\mathbf{X}(\cdot) = \mathbf{C}\boldsymbol{\phi}(\cdot)$, kde $\mathbf{C}$ je matice koeficientů rozměru $n \times K$. Odtud výběrovou kovarianční funkci můžeme vyjádřit ve tvaru

$$v(s, t) = \frac{1}{n-1}\boldsymbol{\phi}(s)'\mathbf{C}'\mathbf{C}\boldsymbol{\phi}(t), \quad s, t \in I. \tag{22}$$

Dále předpokládejme, že lze vlastní funkce $\xi_j, j \geq 1$ rozvinout pomocí ϕ_k,

$$\xi_j(\cdot) = \sum_{k=1}^{K} b_{jk}\phi_k, \tag{23}$$

kde $b_{jk} = \langle \xi_j, \phi_k \rangle_2, k = 1, \ldots, K$, maticově $\xi_j(\cdot) = \boldsymbol{\phi}(\cdot)'\mathbf{b}_j$. S využitím (22) a maticového vyjádření (23) získáme

$$V\xi_j(\cdot) = \int_I v(\cdot, t)\xi_j(t)\, dt = \int_I \frac{1}{n-1}\boldsymbol{\phi}(\cdot)'\mathbf{C}'\mathbf{C}\boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'\mathbf{b}_j$$

$$= \boldsymbol{\phi}(\cdot)' \frac{1}{n-1} \mathbf{C}'\mathbf{C} \left(\int_I \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'\, dt \right) \mathbf{b}_j = \boldsymbol{\phi}(\cdot)' \frac{1}{n-1} \mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b}_j,$$

kde $\mathbf{W}$ je symetrická matice řádu K s prvky $W_{kl} = \langle \phi_k, \phi_l \rangle_2$. Konečně rovnici (20) můžeme nahradit rovnicí ve tvaru

$$\frac{1}{n-1} \mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b}_j = \rho_j \mathbf{b}_j,$$

kdy následným řešením této lineární soustavy získáme $\mathbf{b}_j = (b_{jk})$ z (23). V případě, že systém bázických funkcí je ortonormální, $\mathbf{W} = \mathbf{I}$, a tato úloha přechází v úlohu mnohorozměrné PCA, kdy hledáme vlastní čísla a vlastní vektory matice $\frac{1}{n-1}\mathbf{C}'\mathbf{C}$.

Je potřeba si uvědomit, že podmínka $\|\xi\|_2 = \|\boldsymbol{\phi}'\mathbf{b}\|_2 = 1$ implikuje $\mathbf{b}'\mathbf{W}\mathbf{b} = 1$, a z podmínky ortogonality dvou funkcí ξ_j a ξ_k vyplývá, že $\mathbf{b}_j'\mathbf{W}\mathbf{b}_k = 0$. Abychom dostali funkcionální hlavní komponenty splňující tyto požadavky, definujeme $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{b}$ a řešíme ekvivalentní úlohu ve tvaru

$$\frac{1}{n-1} \mathbf{W}^{1/2}\mathbf{C}'\mathbf{C}\mathbf{W}^{1/2}\mathbf{u} = \rho_j \mathbf{u}$$

za podmínek $\mathbf{u}'\mathbf{u} = 1$. V každém kroku pak spočteme vektor koeficientů $\mathbf{b} = \mathbf{W}^{-1/2}\mathbf{u}$ a určíme $\xi(\cdot)$.

Na rozdíl od mnohorozměrné PCA, kdy počet proměnných p je obvykle menší jak počet pozorování n, v FPCA je počet funkčních hodnot u každého pozorování roven nekonečnu. Z konstrukce $X_j(\cdot)$ však vyplývá, že jsou vzniklé vektory koeficientů lineárně nezávislé, hodnost kovariančního operátoru V je proto $n-1$, což implikuje $n-1$ nenulových vlastních čísel. Oblíbeným nástrojem pro určení optimálního počtu funkcionálních komponent je scree plot, který zobrazuje podíl vysvětlené variability jednotlivými funkcionálními komponentami, v němž hledáme bod zlomu od rychlého klesání k pozvolnému. Případně o jejich optimálním počtu můžeme rozhodnout pomocí jednoduchého pravidla, založeného na zkušenostech z praxe (expert knowledge), kdy vyžadujeme například nejméně 80% vysvětlené variability. Tyto funkcionální komponenty pak zobrazíme do grafu

pomocí střední funkce společně se dvěma funkcemi, které odpovídají násobku příslušné funkcionální komponenty, který je ke střední funkci přičtena a odečtena. Je to z toho důvodu, že funkcionální komponenty vysvětlují variabilitu kolem střední funkce. Způsob, kterým lze určit příslušný násobek, je popsán v [12]. Zobrazením skórů pro dvojice funkcionálních hlavních komponent získáme užitečnou informaci o shlukování jednotlivých pozorování v prostoru vytvořeném komponentami. Nakonec ještě poznamenejme, že mnohorozměná PCA často vychází z korelační matice z důvodu odstranění vlivu nestejného měřítka, neboť na statistických jednotkách měříme p různých vlastností mající často odlišné jednotky. Namísto toho v FPCA získáme soubor pozorování měřením jedné veličiny ve stejné jednotce, a proto vždy vychází z kovarianční funkce.

5.3. PCA pro hustoty (SFPCA)

Nyní již můžeme přejít k odvození postupu výpočtu funkcionálních hlavních komponent (SFPC) [4] pro soubor pozorování, který bude tvořen hustotami. Uvažujme proto centrovaný náhodný výběr $X_1, \ldots, X_n$ v $\mathcal{B}^2(I)$, který získáme z původního výběru $\widetilde{X}_1, \ldots, \widetilde{X}_n$ v $\mathcal{B}^2(I)$ aplikací operace $\ominus$ na každý $\widetilde{X}_i, i = 1, \ldots, n$ způsobem $X_i = \widetilde{X}_i \ominus \overline{X}$, kde $\overline{X} = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \widetilde{X}_i$ značí výběrový průměr (jedná se vlastně o geometrický průměr hustot). Formulujme maximalizační úlohu následoně. Cílem je najít takové funkcionální hlavní komponenty v $\mathcal{B}^2(I)$, prvky $\{\zeta_j\}_{j\geq 1} \in \mathcal{B}^2(I)$, které budou přes všechny možné $\zeta \in \mathcal{B}^2(I)$ maximalizovat výraz

$$\frac{1}{n-1} \sum_{i=1}^{n} \langle X_i, \zeta \rangle_B \text{ za podmínek } \|\zeta\|_B = 1 \text{ a } \langle \zeta, \zeta_k \rangle_B = 0 \text{ pro } k < j, \tag{24}$$

kdy podmínky vzájemné ortogonality $\langle \zeta, \zeta_k \rangle_B = 0$ pro $k < j$ přidáváme pro $j > 1$.

Stejně jako v FPCA získáme ζ_j řešením rovnice

$$V\zeta_j = \lambda_j \odot \zeta_j,$$

kde ζ_j je vlastní funkce příslušná vlastnímu číslu λ_j kovariančního operátoru

$V : \mathcal{B}^2(I) \longrightarrow \mathcal{B}^2(I)$ definovaného pro $x \in \mathcal{B}^2(I)$ jako

$$Vx = \frac{1}{n-1} \odot \bigoplus_{i=1}^{n} \langle X_i, x \rangle_B \odot X_i.$$

Pro praktické výpočty je výhodné úlohu (24) s použitím clr transformace převést z $\mathcal{B}^2(I)$ do $L^2(I)$, kde se provedou potřebné výpočty, které se pak následně prostřednictvím inverzní clr transformují zpět do $\mathcal{B}^2(I)$. Aplikováním clr transformace na (24) dostaneme maximalizaci přes $\zeta \in \mathcal{B}^2(I)$ výrazu

$$\frac{1}{n-1} \sum_{i=1}^{n} \langle \mathrm{clr}\,(X_i)\,, \mathrm{clr}\,(\zeta) \rangle_2 \text{ za podmínek } \|\mathrm{clr}\,(\zeta)\|_2 = 1;$$

$$\langle \mathrm{clr}\,(\zeta)\,, \mathrm{clr}\,(\zeta_k) \rangle_2 = 0 \text{ pro } k < j.$$

Odtud pro $j \geq 1$ lze (24) ekvivalentně formulovat jako

$$\frac{1}{n-1} \sum_{i=1}^{n} \langle \mathrm{clr}\,(X_i)\,, \xi \rangle_2 \text{ za podmínek } \|\xi\|_2 = 1; \langle \xi, \xi_k \rangle_2 = 0 \text{ pro } k < j \text{ a } \int_I \xi = 0, \tag{25}$$

kde opět podmínky ortogonality přidáváme pro $j > 1$ a navíc též podmínku $\int_I \xi = 0$, která plyne z vlastnosti clr tranformace.

Tato úloha by se shodovala s úlohou (19) s danými podmínkami, kdyby zde nebyla dodatečná podmínka na nulovost integrálu funkcí ξ_j. Abychom ukázali, že řešením této úlohy jsou vlastní funkce $\{\xi_j\}_{j \geq 1}$ výběrového kovariančního operátoru $V_{\mathrm{clr}} : L^2(I) \to L^2(I)$ definovaného pro $x \in L^2(I)$ způsobem

$$V_{\mathrm{clr}} x = \frac{1}{n-1} \sum_{i=1}^{n} \langle \mathrm{clr}\,(X_i)\,, x \rangle_2 \,\mathrm{clr}(X_i),$$

je potřeba ukázat, že navíc splňují $\int_I \xi_j = 0$ pro všechna $j \geq 1$. K tomu je potřeba si uvědomit, že důsledkem clr transformace výběru $\mathrm{clr}\,(X_1), \ldots, \mathrm{clr}\,(X_1)$ V_{clr} připouští nulové vlastní číslo ρ_0, které je generované konstantní vlastní funkcí $\xi_0 \equiv \frac{1}{b-a}$,

$$V_{\mathrm{clr}} \xi_0 = \frac{1}{n-1} \sum_{i=1}^{n} \left\langle \mathrm{clr}\,(X_i)\,, \frac{1}{b-a} \right\rangle_2 \mathrm{clr}(X_i)$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{b-a} \left[\int_I \mathrm{clr}\,(X_i) \right] \mathrm{clr}(X_i) \equiv 0.$$

Protože vlastní funkce $\{\xi_j\}_{j\geq 1}$ odpovídají zbylým vlastním číslům ρ_j, musí splňovat podmínku ortogonality k vlastní funkci ξ_0, tj. pro $j \geq 1$ máme $\langle \xi_0, \xi_j \rangle_2 = \frac{1}{b-a} \int_I \xi_j = 0$, z čehož vyplývá splnění podmínky $\int_I \xi_j = 0$. Aplikováním inverzní clr transformace na vlastní funkce $\{\xi_j\}_{j\geq 1}$ příslušné k nenulovým vlastním číslům $\{\rho_j\}_{j\geq 1}$ operátoru V_{clr} získáme hledané SFPC, tj. pro $\xi_j \neq \xi_0$ máme $\zeta_j = \mathrm{clr}^{-1}(\xi_j) =_B \exp(\xi_j)$.

Při výpočtu vlastních funkcí ξ_j budeme opět předpokládat, že každé z $\mathrm{clr}\,(X_i)$ pro $i = 1, \ldots, n$ a $\xi_j, j \geq 1$ lze vyjádřit jako lineární kombinaci K známých bázových funkcí,

$$\mathrm{clr}\,(X_i)\,(\cdot) = \sum_{k=1}^{K} c_{ik}\phi_k(\cdot), \quad \xi_j(\cdot) = \sum_{k=1}^{K} b_{jk}\phi_k(\cdot).$$

Při volbě B-splajnové báze musí být ve výpočtu B-splajnových koeficientů zohledněna podmínka (12), tj. hledané koeficienty budou navíc splňovat (13), neboli

$$\sum_{k=1}^{K} c_{ik} = 0, \quad \sum_{k=1}^{K} b_{jk} = 0. \tag{26}$$

Hledání vlastních funkcí ξ_j přechází stejným způsobem jako u FPCA na úlohu hledání vlastních vektorů $\mathbf{b}_j = (b_{jk})$ řešením soustavy

$$\frac{1}{n-1}\mathbf{C}'\mathbf{C}\mathbf{W}\mathbf{b}_j = \rho_j \mathbf{b}_j,$$

kdy v důsledku první podmínky v (26) dostáváme nulové součty řádků a sloupců matice $\mathbf{C}'\mathbf{C}$, které nečiní žádné výpočetní problémy, a proto jim není potřeba věnovat speciální pozornost.

Stejně jako u FPCA, důležitými nástroji pro interpretaci SFPC je graf skórů pro dvojice hlavních komponent a graf, kde společně se střední funkcí jsou zde vyneseny i pertubace střední funkce j-tou komponentou ζ_j, které jsou mocněny

příslušnou konstantou prostřednictvím operace mocninné tranformace. Pro určení správné dimenze dat a počtu hlavních komponent můžeme opět využít scree plotu nebo stanovené dolní hranice pro podíl vysvětlené variability.

5.4. Simulační studie - SFPCA pro hustoty patřící do rodiny exponenciálních hustot

Cílem simulační studie je porovnat výstupy FPCA a SFPCA na již představených souborech hustot z podkapitoly 4.1. Jedná se o rozdělení patřící do rodiny exponenciálních rozdělení [2], která mají tu vlastnost, že počet parametrů daného rozdělení odpovídá horní hranici dimenze afinního podprostoru Bayesova prostoru, jenž tomuto rozdělení odpovídá [4]. Přitom k-parametrickou exponenciální rodinou rozdělení na I, značíme $Exp_{\mathcal{B}^2(I)}(g, \mathbf{T}, \boldsymbol{\vartheta})$, nazveme množinu hustot

$$f(t, \boldsymbol{\theta}) =_{B^2} g(t) \cdot \exp\left\{\sum_{j=1}^{k} \vartheta_j(\boldsymbol{\theta}) T_j(t)\right\} =_{B^2} g(t) \cdot \prod_{j=1}^{k} \left[\exp\left\{T_j(t)\right\}\right]^{\vartheta_j(\boldsymbol{\theta})}, \quad t \in I, \tag{27}$$

kde $\boldsymbol{\theta}$ je vektor z k-dimenzionálního parametrického prostoru Θ a $g : I \to \mathbf{R}$, $\vartheta_j : \Theta \to \mathbf{R}$ a $T_j : I \to \mathbf{R}$ pro $j = 1, \ldots, k$ jsou borelovsky měřitelné funkce. Protože $Exp_{\mathcal{B}^2(I)}(g, \mathbf{T}, \boldsymbol{\vartheta})$ tvoří konečně dimenzionální afinní podprostor $\mathcal{B}^2(I)$, lze každou hustotu z $Exp_{\mathcal{B}^2(I)}(g, \mathbf{T}, \boldsymbol{\vartheta})$ vyjádřit jako lineární kombinaci v $\mathcal{B}^2(I)$,

$$f(t, \boldsymbol{\theta}) =_{B^2} g(t) \oplus \bigoplus_{j=1}^{k} \left[\vartheta_j(\boldsymbol{\theta}) \odot \exp\left\{T_j(t)\right\}\right], \quad t \in I,$$

s příslušnou clr transformací ve tvaru

$$f_c(t, \boldsymbol{\theta}) = g_c(t) + \sum_{j=1}^{k} \left[\vartheta_j(\boldsymbol{\theta}) \cdot \mathrm{clr}\left(\exp\left\{T_j(t)\right\}\right)\right], \quad t \in I.$$

Použitím SFPCA na soubor hustot, které tvoří k-parametrickou exponenciální rodinu hustot při počtu $k_0 \leq k$ měnících se parametrů, získáme hlavní komponenty, které budou odhady ortonormální báze odpovídajícího k-dimenzionálního

afinního prostoru v $\mathcal{B}^2(I)$. Tyto komponenty budou příslušet $k_0 \leq k$ nenulovým vlastním číslům, kde k_0 bude odpovídat počtu proměnlivých parametrů.

Uvažujme dva soubory hustot. Jako první uvedeme soubor beta hustot s parametry $\alpha_i = 2 + \frac{i}{2}$ a $\beta_i = 2 + \frac{i}{2}$ pro $i = 1, \ldots, 15$ na $I = [0.1, 0.9]$ s hustotami (10), které upravíme do tvaru (27),

$$f(t, \alpha, \beta) =_{B^2} 1 \cdot \exp\left\{(\alpha - 1)\ln(t) + (\beta - 1)\ln(1 - t)\right\}.$$

Z něj vyčteme, že beta rozdělení na I patří k 2-parametrické rodině exponenciálních hustot s $\boldsymbol{\theta} = (\alpha, \beta)$, $\vartheta_1(\boldsymbol{\theta}) = \alpha - 1$, $g(t) = 1$, $\vartheta_2(\boldsymbol{\theta}) = \beta - 1$, $T_1(t) = \ln(t)$ a $T_2(t) = \ln(1 - t)$.

Jako druhý budeme uvažovat soubor hustot chí-kvadrát rozdělení s parametrem $n_i = i + 2$ pro $i = 1, \ldots, 19$ na $I = [e^{-7}, e^{3.5}]$ s hustotami (10). Z následné úpravy těchto hustot podle (27),

$$f(t, n) =_{B^2} e^{-\frac{t}{2}} \cdot \exp\left\{\left(\frac{n}{2} - 1\right)\ln(t)\right\},$$

lze vyčíst, že chí-kvadrát rozdělení $\chi^2(n)$ na I patří k 1-parametrické exponenciální rodině rozdělení s $\theta = n, \vartheta_1(\theta) = \frac{n}{2} - 1$, $g(t) = e^{-\frac{t}{2}}$ a $T_1(t) = \ln(t)$.

Použitím SFPCA jako statistické metody na redukci dimenze dat očekáváme, že pro soubor beta hustot povede na snížení dimenze na $k = 2$, pro druhý soubor hustot pak na $k = 1$. V případě beta hustot si ukážeme, že je potřeba zohlednit též směr, ve kterém se pohybujeme při měnících se parametrech. Na souboru hustot chí-kvadrát rozdělení porovnáme výstupy SFPCA a FPCA. Postup pro první soubor hustot s využitím softwaru R [6] je následující:

1. načteme knihovny `fda` a `robCompositions`,
2. nasimulujeme 15 clr transformovaných beta hustot,
3. diskretizujeme tyto hustoty, čímž získáme 200 funkčních hodnot pro každou z nich,

4. `fdobj`, funkcionální datový objekt, získáme pomocí interpolace bázovými splajny. V našem případě konkrétní volba okrajových podmínek povede na interpolaci přirozenými kubickými splajny,

5. použijeme příkaz `pca.fd` na vytvořený funkcionální datový objekt sestávající z interpolovaných clr hustot. Výstupem funkce `pca.fd` je seznam, který obsahuje:
   `harmonics`: funkcionální datový objekt pro vlastní funkce ξ_j,
   `values`: vlastní čísla ρ_j,
   `scores`: matici skórů $\mathbf{\Psi}$,
   `varprop`: vektor obsahující podíly vysvětlené variability každou vlastní funkcí ξ_j, tj. $\frac{\rho_j}{\sum_j \rho_j}$,
   `meanfd`: funkcionální datový objekt odhadu střední funkce $\mu(t)$, tj. $\bar{x}(t)$.

6. získané vlastní funkce ξ_j v clr prostoru převedeme pomocí inverzní clr do původního prostoru $\mathcal{B}^2(I)$,

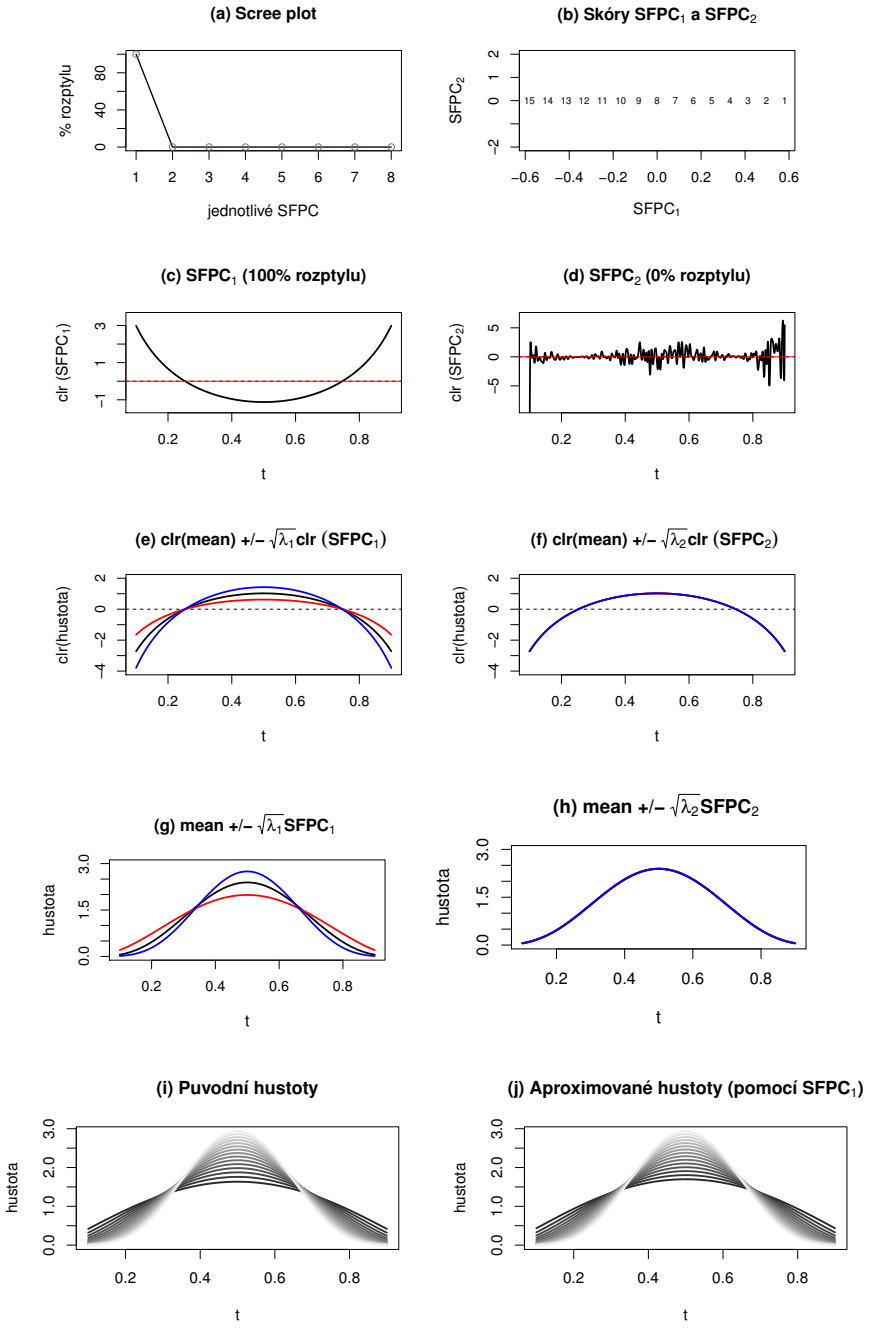
7. kroky 1-5 provedeme i na původní hustoty v případě hustot chí-kvadrát rozdělení a výstupy obou přístupů porovnáme.

Výstup SFPCA pro soubor hustot beta rozdělení můžeme vidět na obrázku 6. První komponenta neboli vlastní funkce (SFPC) vysvětluje veškerou variabilitu dat, a to zejména na koncích I (obrázek 6c). To je jasně zachyceno na obrázku 6e, stejně jako nulový podíl vysvětlené variability druhou SFPC na obrázku 6f, kde je k clr střední funkci přičtena a odečtena $\mathrm{clr}(\mathrm{SFPC})_1$, resp. $\mathrm{clr}(\mathrm{SFPC})_2$, která je násobena příslušnou směrodatnou odchylkou $\sqrt{\lambda_1}$, resp. $\sqrt{\lambda_2}$. Vysoké hodnoty skórů podél SFPC_1 odpovídají vysokým hodnotám směrodatné odchylky příslušných hustot a naopak malé skóry jsou spojeny s nízkými hodnotami směrodatných odchylek hustot. Skóry jsou zobrazeny v obrázku 6b, kde indexy $1, \ldots, 15$ odpovídají hustotám (10) se směrodatnými odchylkami $\sigma_1 = 0.2100 > \sigma_2 = 0.1987 > \ldots > \sigma_{15} = 0.1291$.

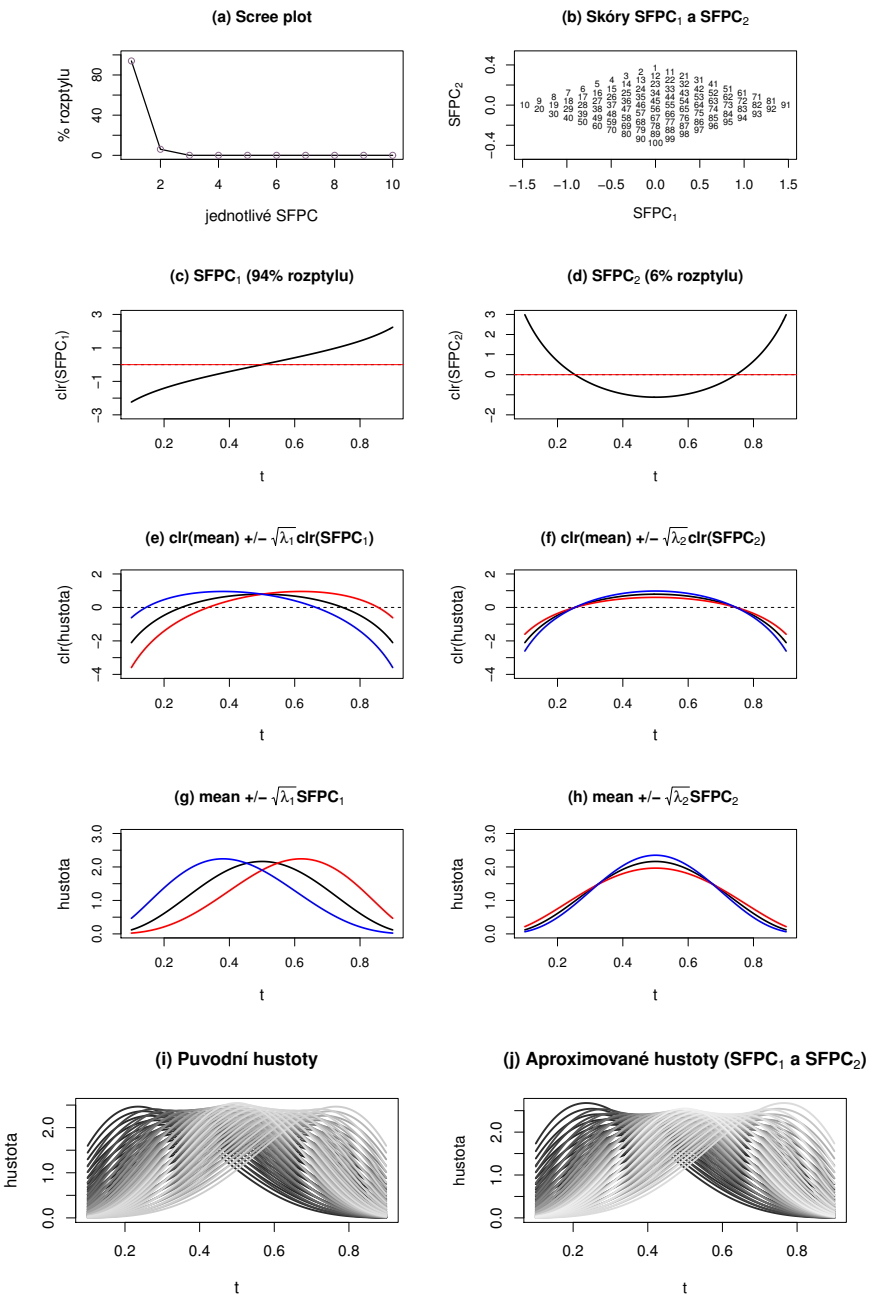
Ačkoli jsme očekávali dvě nenulové SFPC, získali jsme pouze jednu, a to z toho důvodu, že při změně parametrů beta rozdělení docházelo ke změně ve

tvaru hustot pouze v jednom směru. Očekavané dvě nenulové SFPC můžeme vidět na obrázku 7c-7d, které jsme získali pro soubor beta hustot (10) s parametry $\alpha_i = 2 + \frac{i}{3}$ a $\beta_j = 2 + \frac{j}{3}$ pro $i, j = 1, \ldots, 10$. V případě souboru chí-kvadrát hustot je celková variabilita zachycena právě jednou SFPC a jak jsme očekávali, největší variabilitu vysvětluje v levé části distribuce. Podotýkáme, že opět SFPC_1 nám jednotlivé hustoty uspořádala podle jejich příslušných rozptylů, a to od nejmenšího, příslušného $f(t, n_1)$, po největší, příslušného $f(t, n_{19})$.

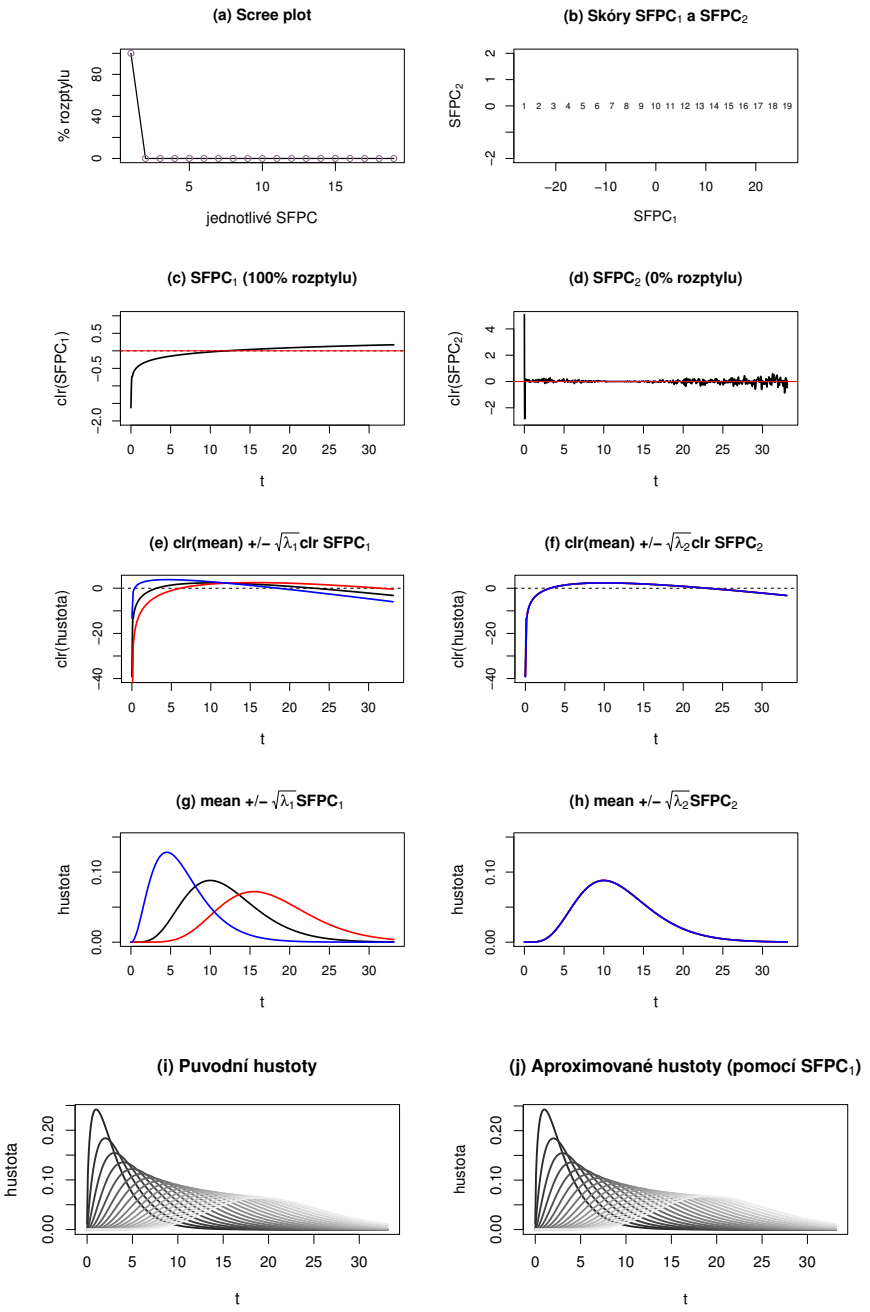
Z výše uvedeného je zřejmé, že SFPCA určuje správnou dimenzi afinního prostoru $\mathcal{B}^2(I)$ uvažovaných exponenciálních rodin a aproximace dat pomocí SFPC se v těchto případech ukazuje jako vhodná. Odlišné výsledky získáme při nerespektování přirozených vlastností hustot použitím standardní FPCA v $L^2(I)$ na soubor chí-kvadrát hustot. Navržená redukce dimenze pomocí scree plotu (obrázek 9a) nerespektuje dimenzi dat, stejně tak si můžeme všimnout nelineárního vztahu mezi skóry (obrázek 9b). Následným porovnáním obrázků 9g a 9h vidíme, že aproximace hustot pomocí prvních dvou FPC se zde jeví zcela nevhodná, neboť navíc obsahuje i záporné hodnoty.



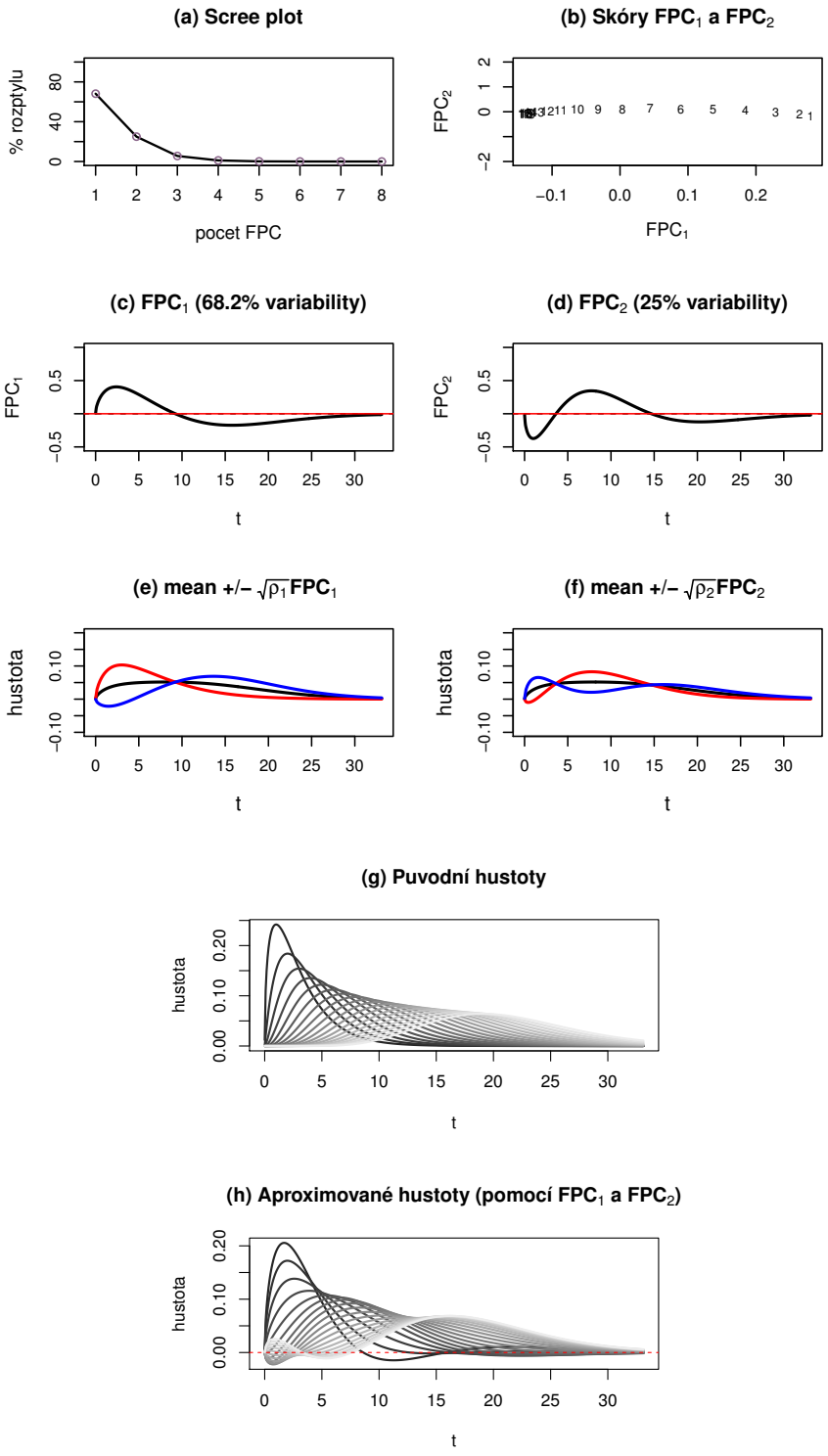
Obrázek 6: SFPCA beta hustot na $I = [0.1, 0.9]$ s $\alpha_i = \beta_i = 2 + \frac{i}{3}$ pro $i = 1, \ldots, 15$. V (e)-(f) černá křivka odpovídá střední funkci, červená střední funkci plus SFPC a modrá střední funkci minus SFPC v clr prostoru. V (g)-(h) černá křivka je střední funkce, červená střední funkce $\oplus$ SFPC, modrá pak střední funkce $\ominus$ SFPC.



Obrázek 7: SFPCA beta hustot na $I = [0.1, 0.9]$ s parametry $\alpha_i = 2+\frac{i}{3}$ a $\beta_j = 2+\frac{j}{3}$ pro $i, j = 1, \ldots, 10$.



Obrázek 8: SFPCA chí-kvadrát hustot na $I = [e^{-7}, e^{3.5}]$ s parametrem $n_i = i + 2$ pro $i = 1, \ldots, 19$.



Obrázek 9: FPCA chí-kvadrát hustot na $I = [e^{-7}, e^{3.5}]$ s $n_i = i + 2$ pro $i = 1, \ldots, 19$.

6. Reálný příklad

Nyní se podíváme, jak lze získané teoretické poznatky uplatnit při práci s reálnými daty. Provedeme SFPCA na souboru dat, který je tvořen rozdělením příjmů domácností v Itálii, jež jsme převzali z článku [8]. Tato data vzešla z oficiálního výzkumu příjmu domácností (SHIW), provedeného centrální italskou bankou. Naším cílem je pomocí snížení dimenze dat získat souhrnnou informaci o rozdělení příjmů v jednotlivých italských regionech, která umožní naše data charakterizovat, přičemž budeme brát v potaz jejich relativní charakter (respektováním geometrie Bayesových prostorů). Výstup SFPCA, který by měl zohlednit průmyslovější sever Itálie oproti zemědělskému jihu, bude následně porovnán se standardním přístupem FPCA.

6.1. Představení a reprezentace dat

K dispozici máme rozdělení příjmů domácností (v EUR) ve všech dvaceti italských regionech, které vzešly z okolo 8000 údajů o výši jejich příjmů. V tabulce 1 můžeme vidět jednotlivá zastoupení příjmů pro každý z regionů, které jsou navíc rozděleny do tří skupin podle geografického uspořádání na sever (N), střed (M) a jih (J), kam byly zařazeny i ostrovy v souladu s tříděním podle italského národního statistického úřadu (ISTAT). Byl určen optimální počet devíti příjmových tříd se středy $x_i, i = 1, \dots, 9$ a s rozsahem $[6574, 110709]$. Hodnoty $f_i, i = 1, \dots, 9$ v tabulce představují podíly jednotlivých tříd na celkovém rozdělení příjmů v rámci jednotlivých regionů. Dále byla na tyto proporce použita diskrétní verze clr transformace [1], tj.

$$z_i = \ln \frac{f_i}{g(f_1, \dots, f_9)},$$

kde g značí geometrický průměr z příslušných složek. Tyto clr transformované hodnoty byly vyhlazeny metodou vyhlazovacích splajnů s nosičem na $[0, 110709]$ při volbě sítě čtyř uzlů ve výši příjmů $0, 30000, 70000, 110709$. Spočtené výsledné vektory B-koeficientů (tabulka 2) pro každý z 20 vyhlazovacích splajnů splňujících

region	poloha	proporce tříd příjmů, $f_1, \ldots, f_9$								
1 Piemonte	N	0.067	0.385	0.323	0.134	0.052	0.022	0.009	0.005	0.003
2 Valle d'Aosta	N	0.042	0.340	0.319	0.212	0.042	0.016	0.006	0.016	0.006
3 Lombardia	N	0.089	0.275	0.269	0.151	0.107	0.056	0.022	0.018	0.012
4 Trentino	N	0.058	0.320	0.279	0.127	0.134	0.029	0.041	0.006	0.005
5 Veneto	N	0.103	0.329	0.255	0.177	0.081	0.022	0.015	0.010	0.007
6 Friuli	N	0.084	0.320	0.232	0.168	0.088	0.068	0.028	0.008	0.004
7 Liguria	N	0.081	0.362	0.207	0.213	0.081	0.026	0.026	0.003	0.002
8 Emilia Romagna	N	0.065	0.303	0.275	0.189	0.085	0.045	0.017	0.015	0.006
9 Toscana	M	0.043	0.283	0.293	0.188	0.105	0.052	0.015	0.015	0.007
10 Umbria	M	0.052	0.351	0.337	0.157	0.056	0.026	0.015	0.004	0.002
11 Marche	M	0.115	0.401	0.219	0.153	0.058	0.032	0.014	0.006	0.003
12 Lazio	M	0.120	0.349	0.260	0.150	0.066	0.032	0.012	0.007	0.002
13 Abruzzo	S	0.100	0.368	0.294	0.144	0.045	0.030	0.004	0.010	0.005
14 Molise	S	0.131	0.349	0.277	0.109	0.080	0.022	0.022	0.006	0.004
15 Campania	S	0.238	0.485	0.167	0.066	0.019	0.016	0.006	0.002	0.001
16 Puglia	S	0.238	0.441	0.201	0.068	0.025	0.009	0.011	0.003	0.002
17 Basilicata	S	0.246	0.385	0.169	0.115	0.038	0.031	0.006	0.006	0.003
18 Calabria	S	0.240	0.408	0.209	0.084	0.037	0.005	0.010	0.004	0.003
19 Sicilia	S	0.255	0.473	0.161	0.053	0.029	0.014	0.012	0.002	0.001
20 Sardegna	S	0.167	0.425	0.217	0.123	0.044	0.015	0.006	0.003	0.002
středy intervalů		6574	19591	32608	45625	58641	71658	84675	97692	110709

Tabulka 1: Proporce tříd příjmů ve 20 italských regionech.

podmínku (13) nyní využijeme pro určení clr transformovaných hustot. Začneme nastavením cesty ke zdrojovým datům v počítači a načtením doplňkové knihovny:

```
> setwd("Cesta k datům")
> library(fda).
```

Následně načteme soubor `bkoef.csv`, v jehož řádcích jsou B-koeficienty pro jednotlivé regiony:

```
> B = read.csv("bkoef.csv", header = FALSE, sep = "", dec=".").
```

Sestrojíme B-splajnovou bázi kubických splajnů na $I = [0, 110709]$ podobně jako v příkladu 2, ale s tím rozdílem, že přidáváme parametr `breaks`, neboť body interpolace se zde liší od zvolených uzlů splajnů:

```
> x.int = c(0,110709)
> nbasis = 6
> norder = 4
> breaks = c(0,30000,70000,110709)
> baze = create.bspline.basis(x.int,nbasis,norder,breaks).
```

Výsledné vyhlazující splajny získáme pomocí funkce `fd`:

```
clr.data.fd=fd(B,baze)
```

a pomocí příkazu `plot`, případně `matplot` je vykreslíme. Na obrázku 10 nahoře

region	poloha	splajnové koeficienty b_i^*, $i = -3, \ldots, 2$					
1 Piemonte	N	−1.972	2.650	2.376	−0.907	−2.202	−2.734
2 Valle d'Aosta	N	−1.801	1.192	3.979	−2.791	−1.048	−1.877
3 Lombardia	N	−1.362	1.609	1.413	−0.129	−1.788	−1.708
4 Trentino	N	−2.686	2.512	1.128	0.519	−2.250	−2.357
5 Veneto	N	−0.660	1.602	2.370	−1.079	−1.872	−2.067
6 Friuli	N	−2.065	2.531	0.742	0.828	−2.205	−2.728
7 Liguria	N	−1.756	2.628	1.467	0.487	−2.643	−3.299
8 Emilia Romagna	N	−1.806	1.609	2.080	−0.666	−1.506	−2.297
9 Toscana	M	−2.595	1.615	2.038	−0.254	−1.823	−2.101
10 Umbria	M	−2.517	2.625	2.226	−0.444	−2.144	−3.254
11 Marche	M	−1.260	2.816	1.418	−0.160	−2.239	−2.896
12 Lazio	M	−0.750	2.247	1.889	−0.555	−2.015	−2.944
13 Abruzzo	S	−0.872	2.184	2.542	−1.268	−2.280	−2.056
14 Molise	S	−0.827	2.474	1.541	−0.448	−2.073	−2.508
15 Campania	S	−0.275	4.574	0.753	−0.177	−2.878	−3.523
16 Puglia	S	0.372	3.304	1.843	−1.926	−1.449	−2.856
17 Basilicata	S	0.476	2.974	1.174	−0.267	−2.672	−2.633
18 Calabria	S	1.041	2.570	2.561	−2.272	−1.802	−2.400
19 Sicilia	S	−0.360	4.563	0.260	0.097	−2.864	−2.874
20 Sardegna	S	−0.124	3.085	1.937	−0.529	−3.097	−2.903

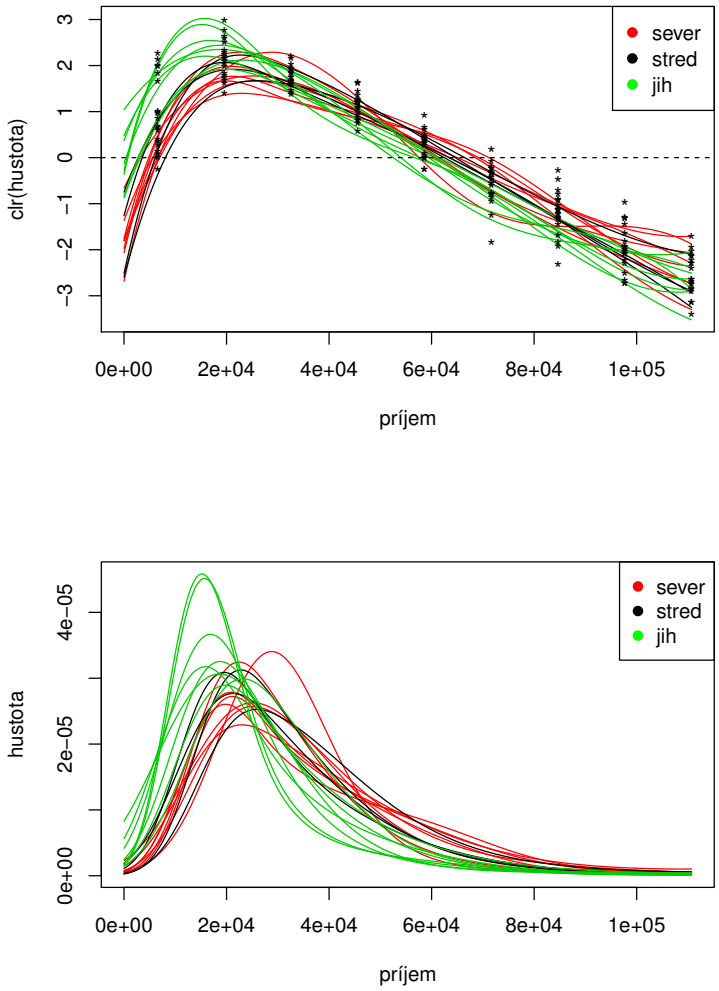
Tabulka 2: B-koeficienty vyhlazujících splajnů pro clr transformované hustoty 20 italských regionů.

můžeme vidět soubor dvaceti vyhlazených clr hustot společně s clr transformacemi původních dat, dole pak původní hustoty, které jsme získali pomocí inverzní clr transformace (9), v obou případech jsou barevně rozlišeny podle geografického uspořádání. Je zde pěkně vidět, že se distribuce příjmů v severní a střední části Itálie liší od těch na jihu země, kde převažují výrazně nižší příjmy domácností.

6.2. SFPCA - aplikace na reálných datech

Na hustoty zachycené na obrázku 10 dole se budeme dívat jako na spojité kompozice v $\mathcal{B}^2(I)$. Z toho důvodu se uchylujeme k SFPCA a následné clr-transformaci (7), která nám umožní provést samotné výpočty v $L^2(I)$.

Na obrázku 11 vidíme výstup SFPCA, který získáme použitím funkce `pca.fd` na funkcionální datový objekt `clr.data.fd`. Zobrazené dvě hlavní komponenty (SFPC) (obrázek 11d-11e) nám vysvětlují postupně 68.7% a 17.3% celkové variability v datech. První komponenta představuje kontrast mezi levou částí distribuce, odpovídající nízkým příjmům, a pravou částí distribuce. To je zřejmý důsledek relativního měřítka, zachyceného ve formě absolutního měřítka clr tramsformací v obrázku 10 nahoře, jež nám umožňuje zachytit relativní zdroj variability dat. Ta je určena především malými hodnotami hustot distribucí v jejich levé



Obrázek 10: Hustoty příjmů v Itálii a jejich clr transformace.

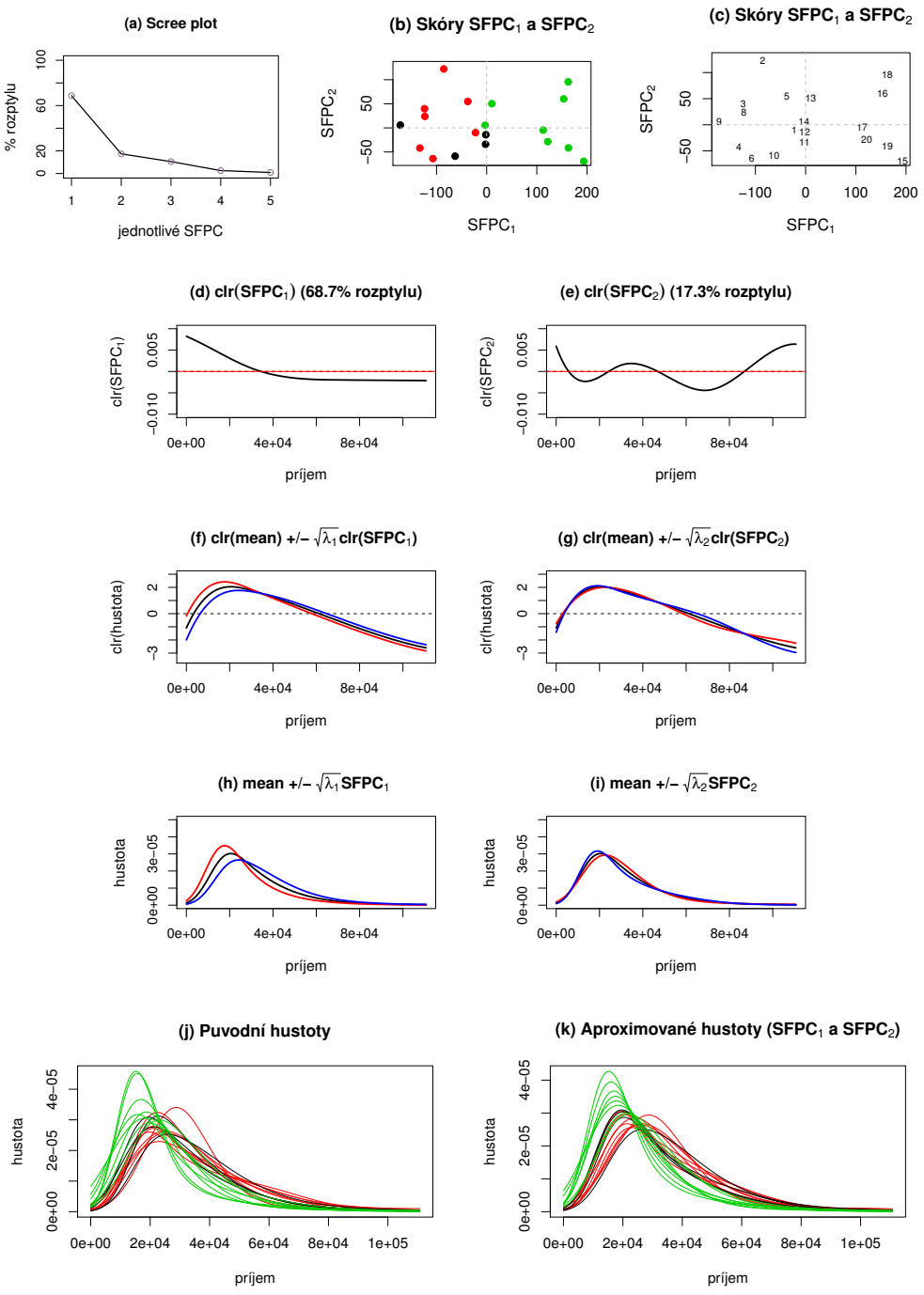
části, kde můžeme sledovat strmější pokles, na rozdíl od malých hodnot napravo, kde je jejich pokles pozvolnější. To nám odráží rozdíly v příjmech domácností mezi jihem země, kde jsou příjmy nižší, a zbytkem Itálie, kde převažují vyšší příjmy. To lze vyčíst i z obrázku 11f, kde je největší variabilita zachycena právě v nižších příjmech.

Podíváme-li se na skóry na obrázku 11b, vidíme, že první SFPC nám rozděluje soubor dat na dvě skupiny, na sever země společně se středem a na jižní část

země. Vysoké skóry podél první SFPC odpovídají velkému (relativnímu) výskytu nízkých příjmů na celkových příjmech, což odpovídá tomu, že na jihu země podíl domácností s nižšími příjmy výrazně převažuje nad domácnostmi s vyššími příjmy. Naopak nízké hodnoty skórů odpovídají malému výskytu nižších příjmů, a tedy vypovídají o převažujícím podílu domácností s vyššími příjmy. U nízkých skórů převládá červená a černá barva, a tedy odpovídají regionům na severu a středu Itálie. Porovnáme-li tyto závěry s údaji z tabulky 1, vidíme, že nejnižší skór má region Toscana (obrázek 11c), ve kterém skutečně podíly nižších příjmů patří mezi nejnižší a převažují v něm vyšší příjmy domácností. Sami se můžete přesvedčit, že naopak vysoké skóry u jižních regionů odráží skutečnost, že je tato část Itálie zaměřena na zemědělství, a proto neoplývá takovým bohatstvím v porovnání se zbytkem země. Naše závěry jsou tedy v souladu s realitou, do které se promítá industriální struktura země.

Druhé hlavní komponentě, vysvětlující část ze zbývající variability, nelze přiřadit žádnou intuitivní intepretaci, nicméně zachycuje variabilitu zejména ve středních a vyšších příjmech.

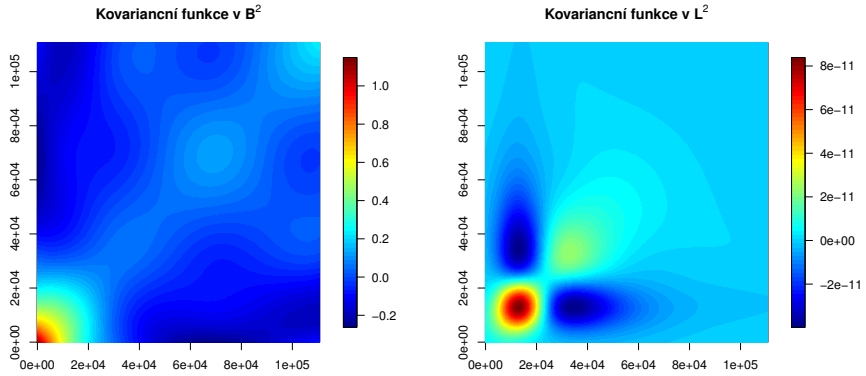
Navržená redukce dimenze pomocí prvních dvou hlavních komponent na základě scree plotu (obrázek 11a) by měla určovat tu správnou dimenzi dat. Aproximaci souboru hustot, kterou získáme projekcí vyhlazených dat do prostoru prvních dvou SFPC, můžeme vidět na obrázku 11k. Mohlo by se zdát, že neodrážejí zcela adekvátně povahu dat zejména ve vrcholcích hustot, nicméně z pohledu hustot je důležitější, že dostatečně přesně charakterizují relativní variabilitu dat.



Obrázek 11: SFPCA hustot příjmů.

Srovnejme nyní výstupy SFPCA a FPCA. FPCA byla aplikována na hustoty, které jsou zobrazeny na obrázku 10 dole. Podíváme-li se na odhadnutou kovarianční funkci v $\mathcal{B}^2(I)$ (obrázek 12 vlevo), vidíme, že nám podpoří dosažené závěry, které jsme učinili pomocí SFPCA. Potvrzuje největší zdroj variability řízený levým koncem distribuce, kde jsou nižší příjmy. Naopak odhadnutá kovarianční funkce v L^2 prostoru (na obrázku 12 vpravo) zachycuje absolutní variabilitu dat, největší rozdíly v příjmech mezi 15000 až 20000 EUR, v nichž hustoty dosahují svých nejvyšších hodnot. Posuzování variability z pohledu L^2 prostorů tak neodpovídá relativnímu charakteru hustot.

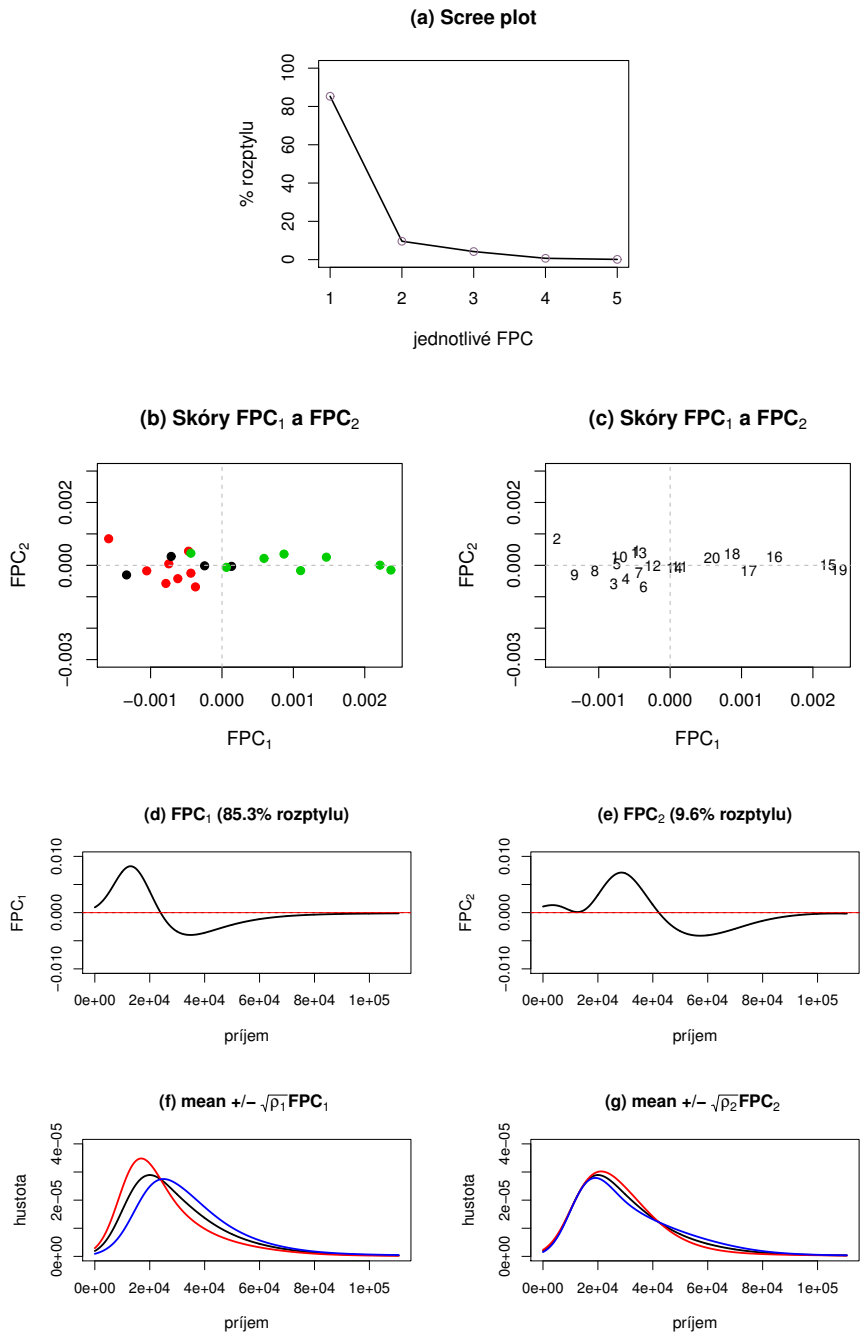
První komponenta vysvětluje 85.3% rozptylu řízená zejména vrcholky hustot a částečně též středními příjmy (obrázek 13d). Komponenta klasifikuje objekty podobně jako v SFPCA do dvou shluků (obrázek 13b-c), na chudší jih Itálie proti bohatší severní a střední část země. Druhá komponenta pak zachycuje zbývající variabilitu ve středních a vyšších příjmech, celkem 9.6% (obrázek 13e).



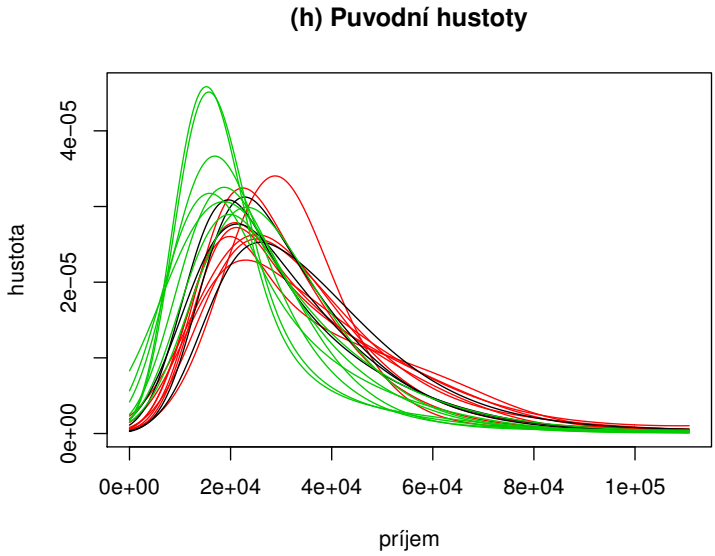
Obrázek 12: Kovarianční funkce v $\mathcal{B}^2(I)$ (vlevo) a kovarianční funkce v $L^2(I)$ (vpravo).

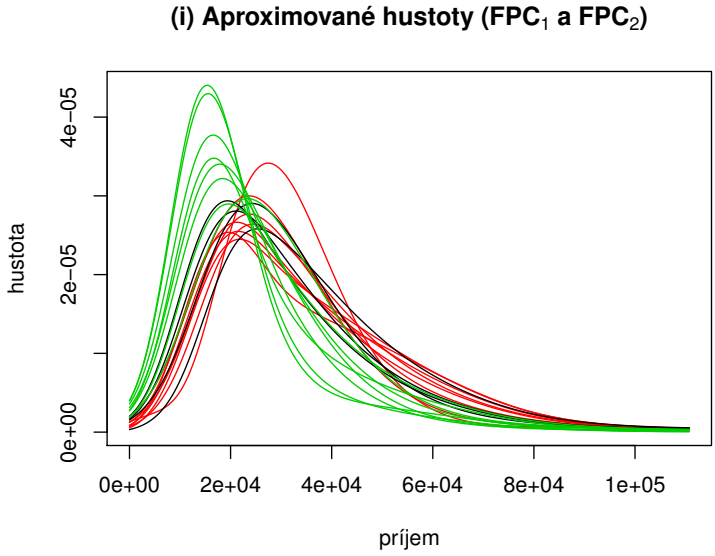
Aproximaci původních dat pomocí prvních dvou hlavních komponent můžeme vidět na obrázku 14 dole. Kvalita výsledné aproximace i přes zachycení většího množství informace se nejeví výrazně lepší v porovnání s aproximací získanou použitím SPFCA. Přestože FPCA díky primárnímu zaměření na zpracování absolutní informace věrohodně charakterizuje vrcholky hustot a v tomto konkrétním

případě dokonce pomocí prvních dvou komponent vysvětlila více celkové variability než SFPCA, její použití na statistické zpracování hustot pravděpodobnosti se z metodického hlediska nejeví jako dostatečně relevantní.



Obrázek 13: FPCA hustot příjmů.





Obrázek 14: Původní hustoty (nahoře), aproximované hustoty (dole).

Závěr

Pro statistické zpracování funkcionálních dat byla navržena ucelená metodika, která nabízí široké spektrum nástrojů pro jejich statistickou analýzu. Při práci se souborem hustot rozdělení pravděpodobností je ovšem stěžejní porozumění jejich geometrickým vlastnostem, které přímé použití těchto standardních statistických metod neumožňuje. V této práci jsem se snažila vytvořit ucelený pohled na tuto problematiku, kdy za stěžejní část mojí práce vnímám zohlednění relativního charakteru hustot ve funkcionální metodě hlavních komponent, která z praktického hlediska představuje jednu z nejpopulárnějších metod funkcionální analýzy.

Studium problematiky statistického zpracování hustot a s tím souvisejících matematických disciplín pro mě bylo jednoznačně velice přínosné. Prvně jsem se setkala se základními poznatky z funkcionální analýzy a přesvědčila jsem se o důležitosti a potřebě analýzy funkcionálních dat v současnosti. Při psaní této práce jsem narazila hned na několik překážek, které se mnohdy pro mě zdály býti nepřekonatelné. Ať už to bylo pochopení struktury Bayesova prostoru, popsaného ve třetí kapitole, nebo problematika interpolace funkcionálních dat. V obou případech se jednalo o relativně novou problematiku, ke které nebylo dostatek literatury. Doufám, že nejenom simulační studie, ale pak i samotné zpracování reálných dat pomocí SFPCA čtenáře přesvědčilo o úskalích použití standardních statistických metod pro statistické zpracování hustot.

Věřím, že by tato práce povede čtenáře k zamyšlení, motivaci a základnímu pochopení principů práce s funkcemi, které jsou hustotami rozdělení pravděpodobností. Na tomto místě bych ještě jednou chtěla poděkovat svému vedoucímu, který mě o výběru představeného tématu přesvědčil a umožnil mi tak seznámit se s touto zajímavou problematikou.

Literatura

[1] Aitchison, J., *The statistical analysis of compositional data*, London, Chapman and Hall, 1986.

[2] Anděl, J., *Základy matematické statistiky*. 2. vydání. Praha: MATFYZPRESS, 2007.

[3] Díaz-Barrero, J.L., Egozcue, J.J., Pawlowsky-Glahn, V., *Hilbert space of probability density functions based on Aitchison geometry*. Acta Mathematica Sinica, English Series 22: 1175-1182, 2006.

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

[5] Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V., *Bayes Hilber spaces*. Australian & New Zeland Journal of Statistics, English series 56(2): 171-194, 2014.

[6] Graves, S., Hooker, G., Ramsay, J.O., *Functional data analysis with R and matlab*. Springer, New York, 2009.

[7] Hron, K., Kunderová, P., *Základy pravděpodobnosti a matematické statistiky*. 1. vydání. Olomouc, vydavatelství Univerzity Palackého v Olomouci, 2013.

[8] Hron, K., Machalová, J., Monti, G.S., *Preprocessing of centred logratio transformed density functions using smoothing splines*. arXiv:1501.07047v1, 2014.

[9] Kobza, J, *Splajny*. 1. vydání. Olomouc, vydavatelství Univerzity Palackého v Olomouci, 1993.

[10] Lukeš, J. *Zápisky z funkcionální analýzy*. Praha, Karolinum, 2012.

[11] Hilbertovy prostory [online], dostupné z: http://www.karlin.mff.cuni.cz/~rokyta/vyuka/1112/zs/F_apl_mat/ApMat_Kap_19_tisk.pdf, [citováno 4. 1. 2015].

[12] Ramsay, J.O., Silverman, B.W., *Functional data analysis*. 2. vydání. Springer, New York, 2005.

[13] *The R Project for Statistical Computing* [online], dostupné z: http://www.r-project.org/, [citováno 10. 11. 2014].

[14] Wehrens, R., *Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences*. Springer, New York, 2011.

[15] Everitt, B., Hothorn, T., *An introduction to applied multivariate analysis with R*. Springer, New York, 2011.