UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Statistická analýza dvourozměrných hustot v Bayesových prostorech

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **prof. RNDr. Karel Hron, Ph.D.**
Vypracoval(a): **Bc. Adéla Czolková**
Studijní program: N0541A170026 Aplikovaná matematika
Studijní obor: Aplikovaná matematika
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Adéla Czolková

Název práce: Statistická analýza dvourozměrných hustot v Bayesových prostorech

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2024

Abstrakt: Reprezentace hustot rozdělení pravděpodobnosti v Bayesových prostorech umožňuje jejich ortogonální rozklad na nezávislou a interakční část, a tím otevírá nový pohled na jejich statistickou analýzu. Cílem diplomové práce je popsat a rozvinout dosavadní poznatky v této oblasti, resp. zpracovat reálná data užitím vhodných metod analýzy funkcionálních dat.

Klíčová slova: Bayesovy prostory, dvourozměrné hustoty, nezávislá a interakční část, funkcionální metoda hlavních komponent

Počet stran: 62

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Adéla Czolková

Title: Statistical analysis of bivariate densities in Bayes spaces

Type of thesis: Master's

Department: Department of Mathematical Analysis and Applications of Mathematics

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

The year of presentation: 2024

Abstract: Representation of probability densities in Bayes spaces enables their orthogonal decomposition into independent and interactive parts and thus opens a new perspective to their statistical analysis. The aim of the Master's thesis is to describe and develop the existing knowledge in this area, respectively analyse an empirical data set using appropriate methods of functional data analysis.

Key words: Bayes spaces, bivariate densities, independent and interactive part, functional principal component analysis

Number of pages: 62

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana prof. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala prof. RNDr. Karlu Hronovi, PhD. za odborné vedení při zpracovávání této práce, cenné rady, trpělivost a čas, který mi věnoval během konzultací.

Úvod

Tématem mé diplomové práce je analýza dvourozměrných hustot rozdělení pravděpodobnosti, na které nahlížíme jako na prvky Bayesova prostoru, což nám umožňuje provedení jejich ortogonálního rozkladu. Tento přístup nabízí nové možnosti analýzy hustot jako funkcionálních dat nesoucích relativní informaci.

Toto téma jsem si zvolila hned ze dvou důvodů. Ve své bakalářské práci jsem se věnovala kompozičním datům a jejich analýze, která právě úzce souvisí s analýzou hustot. Kompoziční data můžeme totiž považovat za data reprezentující diskrétní rozdělení pravděpodobnosti, a proto lze metodologii používanou pro kompoziční data zobecnit pro hustoty, které popisují spojité pravděpodobnostní rozdělení. Druhým důvodem bylo to, že jsem se už se základní teorií Bayesových prostorů seznámila během své letní stáže na Humboldtově univerzitě v Berlíně v roce 2022, a tak jsem v této práci mohla uplatnit své nově získané znalosti, dále je rozvíjet a využít je při práci s reálnými daty.

1. Hustoty rozdělení pravděpodobnosti jako prvky Bayesova prostoru

V první kapitole nejprve definujeme Bayesův prostor, jehož prvky jsou hustoty rozdělení pravděpodobnosti. Popíšeme jeho některé vlastnosti a ukážeme si, jak můžeme hustoty takové, že druhé mocniny jejich logaritmů mají konečný integrál, transformovat z Bayesova prostoru do prostoru reálných funkcí. Tato transformace umožňuje jejich analýzu pomocí klasických funkcionálních statistických metod. Následně se zaměříme na dva přístupy k definici marginálních hustot a představíme si tzv. aritmetické a geometrické marginální hustoty. Dále uvidíme, že každou hustotu lze rozložit na dvě navzájem ortogonální části – nezávislou a interakční. A nakonec se zaměříme na interakční část hustoty, kterou lze intuitivně ještě dále rozložit.

1.1. Bayesův prostor

Bayesovy prostory [2, 4] poskytují geometrickou reprezentaci pro hustoty rozdělení pravděpodobnosti, jejichž charakteristickou vlastností je invariance na změnu měřítka, což znamená, že máme-li dán definiční obor Ω, pak kladné funkce (hustoty) $f(x)$ a $g(x)$, $x \in \Omega$, jsou proporcionální, jestliže nesou stejnou relativní informaci, tj. platí $g(x) = cf(x)$ pro nějaké reálné číslo $c > 0$. Tato vlastnost je přímým důsledkem invariance odpovídajících měr na změnu měřítka.

Uvažujme libovolný měřitelný prostor $(\Omega, \mathcal{A})$, kde $\mathcal{A}$ je σ-algebra na Ω, a na něm definovanou σ-konečnou kladnou reálnou míru λ (Lebesgueovu míru), tzv. referenční míru. Dále nechť $\mathcal{M}(\lambda)$ je množina všech σ-konečných

kladných reálných měr μ na $(\Omega, \mathcal{A})$ takových, že

$$\forall A \in \mathcal{A} : \ \mu(A) = 0 \ \Leftrightarrow \ \lambda(A) = 0.$$

Nyní můžeme definovat relaci ekvivalence $=_{\mathcal{B}}$ na $\mathcal{M}(\lambda)$ následovně

$$\mu, \nu \in \mathcal{M}(\lambda) : \ \mu =_{\mathcal{B}} \nu \quad \Leftrightarrow \quad \exists c \in (0, +\infty) \ \forall A \in \mathcal{A} : \ \nu(A) = c\mu(A).$$

O těchto mírách říkáme, že jsou proporcionální.

Bayesův prostor $\mathcal{B}(\lambda)$ potom definujeme jako tzv. kvocientový prostor $\mathcal{M}(\lambda)| =_{\mathcal{B}}$, tedy prostor tříd ekvivalence měr v $\mathcal{M}(\lambda)$ vzhledem k relaci $=_{\mathcal{B}}$.

Jak již bylo naznačeno výše, existuje vztah mezi mírami a hustotami. Každé míře $\mu \in \mathcal{M}(\lambda)$ můžeme přiřadit hustotu f vzhledem ke zvolené referenční míře λ, neboť platí vztah $f = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ (Radonova–Nikodymova derivace). Díky tomuto vztahu se můžeme dívat na Bayesův prostor $\mathcal{B}(\lambda)$ jako na prostor tříd ekvivalence hustot měr v $\mathcal{M}(\lambda)$, a místo měr tak můžeme pracovat přímo s hustotami.

Přestože jsme dosud jako referenční míru uvažovali pouze Lebesgueovu míru λ, v praxi si můžeme zvolit referenční míru libovolně. Na její volbě však záleží, neboť určuje váhy v definičním oboru Ω [6]. Někdy se jako referenční míra volí pravděpodobnostní míra, tedy míra P taková, že $\mathsf{P}(\Omega) = 1$. Obecně platí, že použijeme-li referenční míru P, a následně míru $c\mathsf{P}$, $c > 0$, budou se výsledky lišit pouze v měřítku a nedojde ke změně relativní informace. Referenční míru můžeme jednoduše změnit z λ na libovolnou míru P s kladnou hustotou $p = \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}\lambda}$ pomocí řetězového pravidla. Pro libovolnou míru μ tedy platí

$$\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \, \mathrm{d}\lambda = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \frac{\mathrm{d}\lambda}{\mathrm{d}\mathsf{P}} \, \mathrm{d}\mathsf{P} = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\lambda} \frac{1}{p} \, \mathrm{d}\mathsf{P}, \quad A \in \mathcal{A}.$$

Vzhledem k tomu, že v této práci nebudeme potřebovat vážený definiční obor (tj. všechny části definičního oboru budou mít stejnou váhu) ani pravděpodobnostní míru, budeme používat Lebesgueovu referenční míru λ.

Definujme nyní dvě operace – perturbaci $\oplus$ a mocnění $\odot$. Nechť $f, g \in \mathcal{B}(\lambda)$ a $\alpha \in \mathbb{R}$, pak

$$(f \oplus g) =_{\mathcal{B}} f \cdot g, \quad (\alpha \odot f) =_{\mathcal{B}} f^{\alpha},$$

kde místo rovnosti používáme ekvivalenci $=_{\mathcal{B}}$, protože pravé strany vztahů mohou být libovolně přeškálovány, aniž by došlo ke změně relativní informace, která bude obsažena ve výsledné hustotě.

$(f \oplus g)$ a $(\alpha \odot f)$ jsou hustoty, které také náleží do $\mathcal{B}(\lambda)$, a trojice $(\mathcal{B}(\lambda), \oplus, \odot)$ tvoří vektorový prostor.

Dále předpokládejme, že λ je konečná míra, a uvažujme podprostor $\mathcal{B}^2(\lambda)$ prostoru $\mathcal{B}(\lambda)$, jehož prvky jsou hustoty, pro které platí, že druhá mocnina jejich logaritmu má konečný integrál, tj.

$$\mathcal{B}^2(\lambda) = \left\{ f \in \mathcal{B}(\lambda) \;\middle|\; \int_{\Omega} |\ln(f)|^2 \,\mathrm{d}\lambda < +\infty \right\}.$$

Trojice $(\mathcal{B}^2(\lambda), \oplus, \odot)$ tvoří vektorový podprostor prostoru $(\mathcal{B}(\lambda), \oplus, \odot)$, na $\mathcal{B}^2(\lambda)$ můžeme definovat skalární součin

$$\langle f, g \rangle_{\mathcal{B}^2(\lambda)} = \frac{1}{2\lambda(\Omega)} \int_{\Omega} \int_{\Omega} \ln \frac{f(s)}{f(t)} \ln \frac{g(s)}{g(t)} \,\mathrm{d}\lambda(s) \,\mathrm{d}\lambda(t),$$

a získáme tak (separabilní) Hilbertův prostor $(\mathcal{B}^2(\lambda), \langle \cdot, \cdot \rangle_{\mathcal{B}^2(\lambda)})$.

Pomocí skalárního součinu můžeme ještě dodefinovat normu a vzdálenost

$$\|f\|_{\mathcal{B}^2(\lambda)} = \sqrt{\langle f, f \rangle_{\mathcal{B}^2(\lambda)}}, \qquad \mathrm{d}_{\mathcal{B}^2(\lambda)}(f, g) = \|f \ominus g\|_{\mathcal{B}^2(\lambda)},$$

kde $f \ominus g = f \oplus [(-1) \odot g]$ je perturbační odčítání hustot.

Hustoty však neanalyzujeme přímo v Bayesově prostoru $\mathcal{B}^2(\lambda)$, ale zobrazíme je do prostoru reálných funkcí, jejichž druhé mocniny mají konečný integrál, což je

$$L^2(\lambda) = \left\{ F : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega F^2 \,\mathrm{d}\lambda < +\infty \right\}.$$

Za tímto účelem definujeme zobrazení clr: $\mathcal{B}^2(\lambda) \to L^2(\lambda)$, známé jako clr (centrovaná logpodílová, anglicky centered logratio) transformace, následovně

$$\mathrm{clr}(f) = \ln f - \frac{1}{\lambda(\Omega)} \int_\Omega \ln f \,\mathrm{d}\lambda.$$

Clr transformací hustoty $f \in \mathcal{B}^2(\lambda)$ tedy získáme funkci $\mathrm{clr}(f)$: $\Omega \to \mathbb{R}$, pro kterou navíc platí

$$\int_\Omega \mathrm{clr}(f) \,\mathrm{d}\lambda = 0,$$

a proto funkce $\mathrm{clr}(f)$ leží v podprostoru

$$L_0^2(\lambda) = \left\{ F : \Omega \to \mathbb{R} \;\middle|\; \int_\Omega F^2 \,\mathrm{d}\lambda < +\infty, \int_\Omega F \,\mathrm{d}\lambda = 0 \right\}$$

prostoru $L^2(\lambda)$, kterému v kontextu Bayesových prostorů říkáme clr prostor.

Clr transformace je izomorfismus [2] – bijektivní zobrazení, které zachovává všechny vlastnosti prostoru $\mathcal{B}^2(\lambda)$. Například víme, že clr transformace je lineární, tj. pro všechny $f, g \in \mathcal{B}^2(\lambda)$ a $\alpha \in \mathbb{R}$ platí

$$\mathrm{clr}(f \oplus g) = \mathrm{clr}(f) + \mathrm{clr}(g), \quad \mathrm{clr}(\alpha \odot f) = \alpha \cdot \mathrm{clr}(f).$$

Transformované hustoty můžeme také snadno zobrazit z prostoru $L_0^2(\lambda)$

zpět do $\mathcal{B}^2(\lambda)$ pomocí inverzní clr transformace, tj.

$$f = \mathrm{clr}^{-1}(F) =_{\mathcal{B}} \exp\{F\} \in \mathcal{B}^2(\lambda), \quad F = \mathrm{clr}(f) \in L_0^2(\lambda).$$

1.2. Aritmetické a geometrické marginální hustoty

V této kapitole si ukážeme dvě různé definice marginálních hustot. Zaměříme se sice pouze na dvourozměrné hustoty, nicméně vše lze zobecnit i pro případ vícerozměrných hustot [2, 4].

V případě dvourozměrných hustot platí, že definiční obor Ω je kartézským součinem definičních oborů Ω_X a Ω_Y. Referenční míra λ je pak součinová míra. Uvažujeme tedy dva měřitelné prostory $(\Omega_X, \mathcal{A}_X)$ a $(\Omega_Y, \mathcal{A}_Y)$ s konečnými kladnými reálnými mírami λ_X a λ_Y, jejich součinem pak získáme měřitelný prostor $(\Omega, \mathcal{A})$ s mírou λ, kde

$$\Omega = \Omega_X \times \Omega_Y, \quad \mathcal{A} = \mathcal{A}_X \times \mathcal{A}_Y, \quad \lambda = \lambda_X \times \lambda_Y.$$

Na prostorech $(\Omega, \mathcal{A})$, $(\Omega_X, \mathcal{A}_X)$, $(\Omega_Y, \mathcal{A}_Y)$ můžeme definovat Bayesovy prostory $\mathcal{B}(\lambda)$, $\mathcal{B}(\lambda_X)$, $\mathcal{B}(\lambda_Y)$ a jejich podprostory $\mathcal{B}^2(\lambda)$, $\mathcal{B}^2(\lambda_X)$, $\mathcal{B}^2(\lambda_Y)$ stejným způsobem jako výše (kapitola 1.1). Stejně budeme postupovat i v případě odpovídajících clr prostorů $L^2(\lambda)$, $L^2(\lambda_X)$, $L^2(\lambda_Y)$ a jejich podprostorů $L_0^2(\lambda)$, $L_0^2(\lambda_X)$, $L_0^2(\lambda_Y)$.

Prostory $\mathcal{B}^2(\lambda_X)$, $\mathcal{B}^2(\lambda_Y)$ můžeme jednoduše vnořit do $\mathcal{B}^2(\lambda)$. Libovolnou míru $\mu_X \in \mathcal{B}^2(\lambda_X)$ zobrazíme na součinovou míru $\mu_X \times \lambda_Y$ a analogicky $\mu_Y \in \mathcal{B}^2(\lambda_Y)$ zobrazíme na $\lambda_X \times \mu_Y$, získáme tak podprostory $\mathcal{B}_X^2(\lambda)$ a $\mathcal{B}_Y^2(\lambda)$ prostoru $\mathcal{B}^2(\lambda)$. Pro hustoty $f \in \mathcal{B}_X^2(\lambda)$ a $g \in \mathcal{B}_Y^2(\lambda)$ pak platí vztahy

$$f =_{\mathcal{B}} f_X \cdot 1, \quad g =_{\mathcal{B}} 1 \cdot g_Y,$$

kde $f_X \in \mathcal{B}^2(\lambda_X)$, $g_Y \in \mathcal{B}^2(\lambda_Y)$ a 1 je neutrální prvek odpovídajícího Bayesova prostoru.

Lze dokázat [2], že prostory $\mathcal{B}^2_X(\lambda)$ a $\mathcal{B}^2_Y(\lambda)$ jsou ortogonální, což je vlastnost důležitá pro ortogonální rozklad dvourozměrné hustoty, jak uvidíme později.

Nyní se již zaměříme na definice marginálních hustot. Nejprve definujeme clr marginální hustoty

$$\begin{aligned}
\operatorname{clr}(f_{X,g})(x) &= \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \operatorname{clr}(f)(x,y)\,\mathrm{d}\lambda_Y = \\
&= \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \ln f(x,y)\,\mathrm{d}\lambda_Y - \frac{1}{\lambda(\Omega)} \int_{\Omega_X}\int_{\Omega_Y} \ln f(x,y)\,\mathrm{d}\lambda_X\,\mathrm{d}\lambda_Y \\
\operatorname{clr}(f_{Y,g})(y) &= \frac{1}{\lambda_X(\Omega_X)} \int_{\Omega_X} \operatorname{clr}(f)(x,y)\,\mathrm{d}\lambda_X = \\
&= \frac{1}{\lambda_X(\Omega_X)} \int_{\Omega_X} \ln f(x,y)\,\mathrm{d}\lambda_X - \frac{1}{\lambda(\Omega)} \int_{\Omega_X}\int_{\Omega_Y} \ln f(x,y)\,\mathrm{d}\lambda_X\,\mathrm{d}\lambda_Y,
\end{aligned}$$

kde $x \in \Omega_X$ a $y \in \Omega_Y$.

Všimneme si, že platí

$$\int_{\Omega_X} \operatorname{clr}(f_{X,g})(x)\ \mathrm{d}\lambda_X = 0, \quad \int_{\Omega_Y} \operatorname{clr}(f_{Y,g})(y)\ \mathrm{d}\lambda_Y = 0,$$

a tedy $\operatorname{clr}(f_{X,g}) \in L^2_0(\lambda_X)$ a $\operatorname{clr}(f_{Y,g}) \in L^2_0(\lambda_Y)$.

Prostory $L^2_0(\lambda_X)$ a $L^2_0(\lambda_Y)$ můžeme opět jednoduše vnořit do prostoru $L^2_0(\lambda)$ tak, že je zobrazíme na

$$L^2_{0,X}(\lambda) = \left\{ F \in L^2_0(\lambda) \;\middle|\; \exists F_X \in L^2_0(\lambda_X)\ \forall (x,y) \in \Omega : F(x,y) = F_X(x) \right\}$$

a

$$L^2_{0,Y}(\lambda) = \left\{ F \in L^2_0(\lambda) \;\middle|\; \exists F_Y \in L^2_0(\lambda_Y)\ \forall (x,y) \in \Omega : F(x,y) = F_Y(y) \right\},$$

což jsou podprostory prostoru $L^2_0(\lambda)$.

Pomocí clr marginálních hustot dále definujeme geometrické marginální hustoty $f_{X,g}$ a $f_{Y,g}$ jako prvky $\mathcal{B}^2(\lambda_X)$ a $\mathcal{B}^2(\lambda_Y)$ a zároveň $\mathcal{B}^2_X(\lambda)$ a $\mathcal{B}^2_Y(\lambda)$, což jsou podprostory prostoru $\mathcal{B}^2(\lambda)$, tedy

$$f_{X,g}(x,y) \equiv f_{X,g}(x) =_{\mathcal{B}} \exp\left\{\mathrm{clr}(f_{X,g})(x)\right\} =_{\mathcal{B}} \exp\left\{ \frac{1}{\lambda_Y(\Omega_Y)} \int\limits_{\Omega_Y} \ln f(x,y)\,\mathrm{d}\lambda_Y \right\},$$

$$f_{Y,g}(x,y) \equiv f_{Y,g}(y) =_{\mathcal{B}} \exp\left\{\mathrm{clr}(f_{Y,g})(y)\right\} =_{\mathcal{B}} \exp\left\{ \frac{1}{\lambda_X(\Omega_X)} \int\limits_{\Omega_X} \ln f(x,y)\,\mathrm{d}\lambda_X \right\},$$

kde $x \in \Omega_X$ a $y \in \Omega_Y$.

Takto definované geometrické marginální hustoty můžeme interpretovat jako ortogonální projekce hustoty $f \in \mathcal{B}^2(\lambda)$ na $\mathcal{B}^2_X(\lambda)$, respektive na $\mathcal{B}^2_Y(\lambda)$. Analogické tvrzení platí i pro clr marginální hustoty, které jsou ortogonálními projekcemi $\mathrm{clr}(f) \in L^2_0(\lambda)$ na $L^2_{0,X}(\lambda)$, respektive na $L^2_{0,Y}(\lambda)$.

V závěru této kapitoly si ještě připomeneme definici marginálních hustot známou ze základního kurzu pravděpodobnosti, tj.

$$f_{X,a}(x) = \int\limits_{\Omega_Y} f(x,y)\,\mathrm{d}\lambda_Y, \quad f_{Y,a}(y) = \int\limits_{\Omega_X} f(x,y)\,\mathrm{d}\lambda_X,$$

tyto hustoty v kontextu Bayesových prostorů nazýváme aritmetické marginální hustoty. Ty se obecně liší od těch geometrických, neboť aritmetické marginální hustoty jsou obsaženy v těch geometrických, které obsahují i jednorozměrnou informaci ze závislostní struktury mnohorozměrné hustoty [4].

1.3. Ortogonální rozklad dvourozměrné hustoty

Pomocí výše definovaných geometrických marginálních hustot definujeme nezávislou část hustoty jako první složku ortogonálního rozkladu dvourozměrné hustoty, druhou složku tohoto rozkladu pak bude tvořit interakční část hustoty [2, 4].

Nejprve definujeme tzv. nezávislý prostor

$$\mathcal{B}^2_{\text{ind}}(\lambda) = \left\{ f \in \mathcal{B}^2(\lambda) \;\middle|\; \exists f_X \in \mathcal{B}^2_X(\lambda)\ \exists f_Y \in \mathcal{B}^2_Y(\lambda) : f = f_X \oplus f_Y \right\}.$$

Nezávislá část hustoty $f \in \mathcal{B}^2(\lambda)$ je pak prvkem právě definovaného prostoru $\mathcal{B}^2_{\text{ind}}(\lambda)$ a platí pro ni

$$f_{\text{ind}}(x, y) = f_{X,g}(x) \oplus f_{Y,g}(y) =_{\mathcal{B}} f_{X,g}(x) \cdot f_{Y,g}(y),$$

kde $(x, y) \in \Omega = \Omega_X \times \Omega_Y$. Z tohoto zápisu je zřejmé, že nezávislá část hustoty je perturbací jejích geometrických marginálních hustot v Bayesově prostoru, což odpovídá i intuitivní interpretaci z pohledu teorie pravděpodobnosti. Nezávislý prostor $\mathcal{B}^2_{\text{ind}}(\lambda)$ je tedy součinový prostor prostorů $\mathcal{B}^2(\lambda_X)$ a $\mathcal{B}^2(\lambda_Y)$, a také platí

$$\mathcal{B}^2_{\text{ind}}(\lambda) = \mathcal{B}^2_X(\lambda) \oplus \mathcal{B}^2_Y(\lambda).$$

Dále lze dokázat [2, 4], že pro libovolnou hustotu $f \in \mathcal{B}^2(\lambda)$ je nezávislá část f_{ind} její ortogonální projekcí na $\mathcal{B}^2_{\text{ind}}(\lambda)$.

Nyní definujeme i druhou složku ortogonálního rozkladu, a to interakční část hustoty, využijeme při tom nezávislou část a geometrické marginální

hustoty, tedy

$$f_{\text{int}}(x,y) = f(x,y) \ominus f_{\text{ind}}(x,y) =_{\mathcal{B}} \frac{f(x,y)}{f_{X,g}(x) \cdot f_{Y,g}(y)},$$

kde $(x,y) \in \Omega = \Omega_X \times \Omega_Y$.

Podobně jako jsme definovali nezávislý prostor, můžeme definovat i interakční prostor

$$\mathcal{B}^2_{\text{int}}(\lambda) = \left\{ f \in \mathcal{B}^2(\lambda) \;\middle|\; \forall g \in \mathcal{B}^2_{\text{ind}}(\lambda) : \langle f, g \rangle_{\mathcal{B}^2(\lambda)} = 0 \right\},$$

který je ortogonálním doplňkem prostoru $\mathcal{B}^2_{\text{ind}}(\lambda)$.

Stejně jako pro nezávislou část hustoty, pro její interakční část platí, že f_{int} je ortogonální projekcí hustoty f z $\mathcal{B}^2(\lambda)$ na $\mathcal{B}^2_{\text{int}}(\lambda)$.

Na základě výše uvedených vztahů tedy můžeme libovolnou hustotu $f \in \mathcal{B}^2(\lambda)$ rozložit následovně

$$f = f_{\text{ind}} \oplus f_{\text{int}} = f_{X,g} \oplus f_{Y,g} \oplus f_{\text{int}}.$$

Uveďme ještě několik tvrzení, které platí pro každou hustotu $f \in \mathcal{B}^2(\lambda)$ [2, 4].

- Pythagorova věta:

$$\begin{aligned} \|f\|^2_{\mathcal{B}^2(\lambda)} &= \|f_{\text{ind}}\|^2_{\mathcal{B}^2(\lambda)} + \|f_{\text{int}}\|^2_{\mathcal{B}^2(\lambda)} = \\ &= \|f_{X,g}\|^2_{\mathcal{B}^2(\lambda)} + \|f_{Y,g}\|^2_{\mathcal{B}^2(\lambda)} + \|f_{\text{int}}\|^2_{\mathcal{B}^2(\lambda)}. \end{aligned}$$

- Geometrické marginální hustoty interakční části f_{int} jsou

$$f_{\text{int},X,g} =_{\mathcal{B}} 1 \quad \text{a} \quad f_{\text{int},Y,g} =_{\mathcal{B}} 1.$$

- Pokud $f = f_X \cdot f_Y \in \mathcal{B}^2(\lambda)$, kde $f_X \in \mathcal{B}^2(\lambda_X)$ a $f_Y \in \mathcal{B}^2(\lambda_Y)$, pak $f_{X,g} =_{\mathcal{B}} f_X$ a $f_{Y,g} =_{\mathcal{B}} f_Y$. Také

$$f \in \mathcal{B}^2_{\text{ind}}(\lambda) \quad \Leftrightarrow \quad f =_{\mathcal{B}} f_{X,g} \oplus f_{Y,g} \quad \Leftrightarrow \quad f_{\text{int}} =_{\mathcal{B}} 1.$$

- Nechť $g =_{\mathcal{B}} g_X \cdot g_Y \in \mathcal{B}^2_{\text{ind}}$, $g_X \in \mathcal{B}^2(\lambda_X)$, $g_Y \in \mathcal{B}^2(\lambda_Y)$ a uvažujeme $h = f \oplus g$, $f \in \mathcal{B}^2(\lambda)$. Potom

$$h_{\text{ind}} = f_{\text{ind}} \oplus g, \quad h_{\text{int}} = f_{\text{int}}$$

  a geometrické marginální hustoty jsou

$$h_{X,g} =_{\mathcal{B}} f_{X,g} \cdot g_X \quad \text{a} \quad h_{Y,g} =_{\mathcal{B}} f_{Y,g} \cdot g_Y.$$

Ortogonální rozklad lze provést i v clr prostoru $L^2_0(\lambda)$, kde

$$\text{clr}(f_{\text{ind}}) = \text{clr}(f_{X,g}) + \text{clr}(f_{Y,g}) \quad \text{a} \quad \text{clr}(f_{\text{int}}) = \text{clr}(f) - \text{clr}(f_{X,g}) - \text{clr}(f_{Y,g}).$$

Zůstává v platnosti také například Pythagorova věta, tedy

$$\begin{aligned} \|\text{clr}(f)\|^2_{L^2(\lambda)} &= \|\text{clr}(f_{\text{ind}})\|^2_{L^2(\lambda)} + \|\text{clr}(f_{\text{int}})\|^2_{L^2(\lambda)} \\ &= \|\text{clr}(f_{X,g})\|^2_{L^2(\lambda)} + \|\text{clr}(f_{Y,g})\|^2_{L^2(\lambda)} + \|\text{clr}(f_{\text{int}})\|^2_{L^2(\lambda)}. \end{aligned}$$

Na závěr se ještě krátce vrátíme k aritmetickým marginálním hustotám. Jestliže $f \in \mathcal{B}^2_{\text{ind}}$, tj. $f = f_X \cdot f_Y$, $f_X \in \mathcal{B}^2(\lambda_X)$, $f_Y \in \mathcal{B}^2(\lambda_Y)$, potom její aritmetické a geometrické marginální hustoty splývají, což znamená, že

$$f_{X,g} =_{\mathcal{B}} f_X = f_{X,a} \quad \text{a} \quad f_{Y,g} =_{\mathcal{B}} f_Y = f_{Y,a}.$$

1.4. Rozklad interakční části hustoty

V této kapitole se podíváme, jak bychom dále mohli rozložit interakční část hustoty $f_{\text{int}} \in \mathcal{B}^2_{\text{int}}(\lambda)$ tím, že provedeme její singulární rozklad.

Nejprve definujeme jádrové funkce v clr prostoru pro referenční míru $\lambda = \lambda_X \times \lambda_Y$ následovně

$$k_1(x_1, x_2) = \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \operatorname{clr}(f_{\text{int}})(x_1, y) \operatorname{clr}(f_{\text{int}})(x_2, y) \, \mathrm{d}\lambda_Y(y),$$
$$k_2(y_1, y_2) = \frac{1}{\lambda_X(\Omega_X)} \int_{\Omega_X} \operatorname{clr}(f_{\text{int}})(x, y_1) \operatorname{clr}(f_{\text{int}})(x, y_2) \, \mathrm{d}\lambda_X(x).$$

Jedná se vlastně o kovarianční funkce, které udávají závislost mezi dvojicemi bodů definičního oboru Ω_X, respektive Ω_Y. Navíc lze ukázat, že jejich integrály jsou nulové, neboť clr marginální hustoty interakční části hustoty jsou také nulové, protože $f_{\text{int},X,g} =_{\mathcal{B}} 1$, $f_{\text{int},Y,g} =_{\mathcal{B}} 1$, $\ln 1 = 0$ a

$$\int_{\Omega_Y} \operatorname{clr}(f_{\text{int}})(x, y) \, \mathrm{d}\lambda_Y(y) = 0, \quad \int_{\Omega_X} \operatorname{clr}(f_{\text{int}})(x, y) \, \mathrm{d}\lambda_X(x) = 0,$$

pak tedy platí

$$\begin{aligned}
&\int_{\Omega_X} k_1(x_1, x_2) \, \mathrm{d}\lambda_X(x_1) = \\
&\quad = \int_{\Omega_X} \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \operatorname{clr}(f_{\text{int}})(x_1, y) \operatorname{clr}(f_{\text{int}})(x_2, y) \, \mathrm{d}\lambda_Y(y) \, \mathrm{d}\lambda_X(x_1) = \\
&\quad = \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \operatorname{clr}(f_{\text{int}})(x_2, y) \left[\int_{\Omega_X} \operatorname{clr}(f_{\text{int}})(x_1, y) \, \mathrm{d}\lambda_X(x_1) \right] \mathrm{d}\lambda_Y(y) = 0,
\end{aligned}$$

$$\int_{\Omega_X} k_1(x_1, x_2)\ \mathrm{d}\lambda_X(x_2) =$$

$$= \int_{\Omega_X} \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \mathrm{clr}(f_{\mathrm{int}})(x_1, y)\, \mathrm{clr}(f_{\mathrm{int}})(x_2, y)\ \mathrm{d}\lambda_Y(y)\, \mathrm{d}\lambda_X(x_2) =$$

$$= \frac{1}{\lambda_Y(\Omega_Y)} \int_{\Omega_Y} \mathrm{clr}(f_{\mathrm{int}})(x_1, y) \left[\int_{\Omega_X} \mathrm{clr}(f_{\mathrm{int}})(x_2, y)\ \mathrm{d}\lambda_X(x_2) \right] \mathrm{d}\lambda_Y(y) = 0$$

a analogicky pro $k_2(y_1, y_2)$.

Proto pro libovolnou vlastní funkci φ jádrové funkce k_1 odpovídající vlastnímu číslu γ, tj.

$$\int_{\Omega_X} k_1(x, t)\, \varphi(t)\ \mathrm{d}\lambda_X(t) = \gamma\varphi(x),$$

platí, že $\varphi \in L_0^2(\lambda_X)$. Podobně pro libovolnou vlastní funkci ψ jádrové funkce k_2 platí, že $\psi \in L_0^2(\lambda_Y)$.

Dále podle [3] existují nerostoucí posloupnost kladných vlastních čísel $\{\gamma_i\}$ jádrových funkcí k_1 a k_2, ortonormální posloupnost $\{\varphi_i\}$, $\varphi_i \in L_0^2(\lambda_X)$ vlastních funkcí funkce k_1 a ortonormální posloupnost $\{\psi_i\}$, $\psi_i \in L_0^2(\lambda_Y)$ vlastních funkcí funkce k_2 takové, že pro $x_1, x_2, x \in \Omega_X$, $y_1, y_2, y \in \Omega_Y$ platí

$$k_1(x_1, x_2) = \sum_{i=1}^{\infty} \gamma_i \varphi_i(x_1) \varphi_i(x_2),$$
$$k_2(y_1, y_2) = \sum_{i=1}^{\infty} \gamma_i \psi_i(y_1) \psi_i(y_2)$$

a

$$\mathrm{clr}(f_{\mathrm{int}})(x, y) = \sum_{i=1}^{\infty} \gamma_i^{1/2} \varphi_i(x) \psi_i(y),$$

a navíc všechny výše uvedené řady konvergují ve smyslu L^2 normy.

Poslední vztah lze vyjádřit také přímo v prostoru $\mathcal{B}^2(\lambda)$ (v jeho podprostoru $\mathcal{B}^2_{\text{int}}(\lambda)$) jako

$$f_{\text{int}}(x,y) = \bigoplus_{i=1}^{\infty} \gamma_i^{1/2} \odot \exp\left\{\varphi_i(x)\psi_i(y)\right\}.$$

Funkce φ_i a ψ_i bychom chtěli použít k podrobnějšímu zkoumání struktury závislosti. Podobně jako v jiných metodách redukce dimenze hodnoty γ_i určují důležitost jednotlivých komponent (vlastních funkcí) φ_i a ψ_i, pomocí tzv. scree plotu můžeme zvolit vhodný počet komponent pro další analýzu a vytvořit pomocí nich dobrou aproximaci zkoumané interakční části hustoty f_{int}.

2. Metody analýzy dvourozměrných hustot

Tato kapitola bude věnována metodám analýzy dvourozměrných hustot. Hustoty analyzujeme jako funkcionální data, která jsou nekonečné dimenze, a proto cílem zde představených metod je redukce dimenze hustot, která umožňuje přehlednější vizualizaci dat a snadnější interpretaci výsledků jejich analýzy. Asi nejznámější metodou redukce dimenze je metoda hlavních komponent, a právě té se zde budeme věnovat. Nejprve si představíme simplexovou funkcionální metodu hlavních komponent (anglicky simplical functional principal component analysis, SFPCA) pro jednorozměrné hustoty, poté si ukážeme její mnohorozměrnou verzi, která slouží k analýze mnohorozměrných funkcionálních kompozic (tedy „vektorů“, jejichž složky jsou jednorozměrné hustoty), a nakonec se podíváme, jak lze metodu hlavních komponent použít k analýze dvourozměrných hustot.

2.1. SFPCA

Začněme tedy s jednorozměrnou SFPCA, což je upravená verze funkcionální metody hlavních komponent (anglicky functional principal component analysis, FPCA) vhodná pro analýzu hustot, tj. dat nesoucích relativní informaci (FPCA nezohledňuje invarianci na změnu měřítka a relativní měřítko) [5].

Nechť $\widetilde{X}_1, \ldots, \widetilde{X}_N$ je náhodný výběr jednorozměrných hustot v $\mathcal{B}^2(\lambda)$ (dále jen $\mathcal{B}^2$), což je v tomto případě Bayesův prostor na intervalu I s referenční mírou λ. Tento výběr budeme centrovat, tj. pro $i = 1, \ldots, N$ získáme $X_i = \widetilde{X}_i \ominus \overline{X}$, kde $\overline{X} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \widetilde{X}_i$ je výběrový průměr (průměrná hustota). Nyní budeme hledat směry největší variability v prostoru $\mathcal{B}^2$, tedy

hlavní komponenty $\{\zeta_j\}_{j\leq 1},\ \zeta_j \in \mathcal{B}^2$, které maximalizují výraz

$$\frac{1}{N}\sum_{i=1}^{N}\langle X_i,\zeta\rangle_{\mathcal{B}^2}^2 \quad \text{za podmínek} \quad \|\zeta\|_{\mathcal{B}^2}=1 \quad \text{a} \quad \langle\zeta,\zeta_k\rangle_{\mathcal{B}^2}=0,\ k<j,$$

kde podmínka ortogonality $\langle\zeta,\zeta_k\rangle_{\mathcal{B}^2}=0,\ k<j$ má smysl pouze pro $j\geq 2$.

Platí, že j-tá komponenta ζ_j je řešením rovnice

$$V\zeta_j=\gamma_j\odot\zeta_j,$$

kde γ_j je vlastní číslo odpovídající vlastní funkci ζ_j výběrového kovariančního operátoru $V\colon\mathcal{B}^2\to\mathcal{B}^2$, který hustotě $f\in\mathcal{B}^2$ přiřazuje hustotu

$$Vf=\frac{1}{N}\odot\bigoplus_{i=1}^{N}\langle X_i,f\rangle_{\mathcal{B}^2}\odot X_i.$$

Vlastní čísla γ_j a vlastní funkce ζ_j však nebudeme hledat přímo v prostoru $\mathcal{B}^2$, ale využijeme clr transformaci a přeneseme tento problém do prostoru $L_0^2(\lambda)$ (dále opět jen L_0^2). Maximalizujeme tedy výraz

$$\frac{1}{N}\sum_{i=1}^{N}\langle\mathrm{clr}(X_i),\mathrm{clr}(\zeta)\rangle_{L_0^2}^2$$
$$\text{za podmínek} \quad \|\mathrm{clr}(\zeta)\|_{L_0^2}=1 \quad \text{a} \quad \langle\mathrm{clr}(\zeta),\mathrm{clr}(\zeta_k)\rangle_{L_0^2}=0,\ k<j.$$

Hledáme tedy funkci $\nu=\mathrm{clr}(\zeta)\in L_0^2$, tj. $\nu\in L^2,\ \int_I\nu\,\mathrm{d}\lambda=0$, která maximalizuje výraz

$$\frac{1}{N}\sum_{i=1}^{N}\langle\mathrm{clr}(X_i),\nu\rangle_{L_0^2}^2 \quad \text{za podmínek} \quad \|\nu\|_{L_0^2}=1 \quad \text{a} \quad \langle\nu,\nu_k\rangle_{L_0^2}=0,\ k<j,$$

podmínka ortogonality $\langle\nu,\nu_k\rangle_{L_0^2}=0,\ k<j$ má opět smysl pouze pro $j\geq 2$.

Řešením jsou v tomto případě vlastní funkce $\{\xi_j\}_{j\geq 1}$ výběrového kovariančního operátoru $V_{\mathrm{clr}} : L_0^2 \to L_0^2$ transformovaného náhodného výběru $\mathrm{clr}(X_1), \ldots, \mathrm{clr}(X_N)$ takového, že pro $F \in L_0^2$ je

$$V_{\mathrm{clr}} F = \frac{1}{N} \sum_{i=1}^{N} \langle \mathrm{clr}(X_i), F \rangle_{L_0^2} \cdot \mathrm{clr}(X_i).$$

V prostoru L_0^2 můžeme také použít ekvivalentní zápis operátoru V_{clr}:

$$V_{\mathrm{clr}} F = \int_I v(\cdot, t) F(t) \, \mathrm{d}\lambda(t) = \int_I v(\cdot, t) F(t) \, \mathrm{d}t,$$

kde $v \colon I \times I \to \mathbb{R}$ je výběrová kovarianční funkce

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

Vyřešením rovnice

$$V_{\mathrm{clr}} \, \xi_j = \gamma_j \, \xi_j,$$

kde $\gamma_j > 0$ je vlastní číslo odpovídající vlastní funkci ξ_j, získáme tak j-tou funkcionální hlavní komponentu ξ_j a dopočítáme skóry

$$z_{ij} = \langle \mathrm{clr}(X_i), \xi_j \rangle_{L_0^2}, \quad i = 1, \ldots, N.$$

Platí také, že vlastní čísla $\gamma_1, \gamma_2, \ldots$ operátoru V_{clr} jsou stejná jako vlastní čísla operátoru V. Stejně jako u klasické (mnohorozměrné) metody hlavních komponent podíl $\frac{\gamma_j}{\sum \gamma_j}$ udává, kolik z celkové variability je vysvětleno pomocí j-té komponenty ξ_j, resp. ζ_j.

Hledané hlavní komponenty v $\mathcal{B}^2$ pak získáme pomocí inverzní clr trans-

formace, tj.

$$\zeta_j = \mathrm{clr}^{-1}(\xi_j) =_{\mathcal{B}} \exp(\xi_j),\ j \geq 1,$$

které jsou určeny jednoznačně až na mocnění čísly ± 1, a skóry vypočteme jako

$$z_{ij} = \langle X_i, \zeta_j \rangle_{\mathcal{B}^2}, \quad i = 1, \ldots, N,\ j = 1, 2, \ldots.$$

Snadno navíc ukážeme, že skóry v $\mathcal{B}^2$ jsou totožné jako skóry v L_0^2, neboť

$$\begin{aligned}
&\langle \mathrm{clr}(X_i), \xi_j \rangle_{L_0^2} = \int_I \mathrm{clr}(X_i)(x)\, \xi_j(x)\, \mathrm{d}x = \int_I \mathrm{clr}(X_i)(x)\, \mathrm{clr}(\zeta_j)(x)\, \mathrm{d}x = \\
&= \int_I \left(\ln X_i(x) - \frac{1}{\lambda(I)} \int_I \ln X_i(t)\, \mathrm{d}t \right) \left(\ln \zeta_j(x) - \frac{1}{\lambda(I)} \int_I \ln \zeta_j(t)\, \mathrm{d}t \right) \mathrm{d}x = \\
&= \int_I \ln X_i(x)\ \ln \zeta_j(x)\, \mathrm{d}x - \frac{1}{\lambda(I)} \int_I \ln X_i(x)\, \mathrm{d}x \int_I \ln \zeta_j(t)\, \mathrm{d}t - \\
&\quad - \frac{1}{\lambda(I)} \int_I \ln \zeta_j(x)\, \mathrm{d}x \int_I \ln X_i(t)\, \mathrm{d}t\ + \frac{1}{\lambda(I)} \int_I \ln \zeta_j(t)\, \mathrm{d}t \int_I \ln X_i(t)\, \mathrm{d}t = \\
&= \int_I \ln X_i(x)\ \ln \zeta_j(x)\, \mathrm{d}x - \frac{1}{\lambda(I)} \int_I \ln X_i(x)\, \mathrm{d}x \int_I \ln \zeta_j(x)\, \mathrm{d}x,
\end{aligned}$$

kde $\lambda(I)$ je délka intervalu I,

$$\begin{aligned}
\langle X_i, \zeta_j \rangle_{\mathcal{B}^2} &= \frac{1}{2\lambda(I)} \int_I \int_I \ln\left(\frac{X_i(s)}{X_i(t)}\right) \ln\left(\frac{\zeta_j(s)}{\zeta_j(t)}\right) \mathrm{d}s\,\mathrm{d}t = \\
&= \frac{1}{2\lambda(I)} \int_I \int_I \Big[\ln X_i(s) - \ln X_i(t)\Big] \Big[\ln \zeta_j(s) - \ln \zeta_j(t)\Big] \mathrm{d}s\,\mathrm{d}t = \\
&= \frac{1}{\lambda(I)} \int_I \int_I \ln X_i(s)\ \ln \zeta_j(s)\,\mathrm{d}s\,\mathrm{d}t - \frac{1}{\lambda(I)} \int_I \int_I \ln X_i(s)\ \ln \zeta_j(t)\,\mathrm{d}s\,\mathrm{d}t = \\
&= \int_I \ln X_i(x)\ \ln \zeta_j(x)\,\mathrm{d}x - \frac{1}{\lambda(I)} \int_I \ln X_i(x)\,\mathrm{d}x \int_I \ln \zeta_j(x)\,\mathrm{d}x,
\end{aligned}$$

tj. oba skalární součiny dokážeme upravit do stejného tvaru, a proto jsou si rovny.

2.2. Mnohorozměrná SFPCA

Jak už bylo zmíněno výše, mnohorozměrná SFPCA se využívá při analýze mnohorozměrných funkcionálních kompozic, to jsou K–dimenzionální vektory (kompozice), jejichž prvky jsou jednorozměrné hustoty, tedy prvky Bayesova prostoru $\mathcal{B}^2$ na intervalu I s referenční mírou λ. Může to být například kompozice, jejíž prvky jsou (jednorozměrné) marginální hustoty nějaké mnohorozměrné hustoty [1].

Mnohorozměrné funkcionální kompozice $\boldsymbol{f} = (f_1, \ldots, f_K)',\ f_i \in \mathcal{B}^2,\ i = 1, \ldots, K$ můžeme považovat za prvky prostoru $[\mathcal{B}^2]^K = \mathcal{B}^2 \times \cdots \times \mathcal{B}^2$, který je separabilním Hilbertovým prostorem, jestliže na něm pro $\boldsymbol{f}, \boldsymbol{g} \in [\mathcal{B}^2]^K$ a $\alpha \in \mathbb{R}$ definujeme následující operace:

- perturbaci $\boldsymbol{f} \oplus \boldsymbol{g} = (f_1 \oplus g_1, \ldots, f_K \oplus g_K)'$,
- mocnění $\alpha \odot \boldsymbol{f} = (\alpha \odot f_1, \ldots, \alpha \odot f_K)'$,
- skalární součin $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{[\mathcal{B}^2]^K} = \sum_{i=1}^K \langle f_i, g_i \rangle_{\mathcal{B}^2}$.

Uvažujeme náhodný výběr mnohorozměrných funkcionálních kompozic $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, předpokládáme, že je centrovaný (každý výběr lze vždy centrovat), a hledáme hlavní směry jeho variability, tedy $\{\boldsymbol{\zeta}_j\}_{j\geq 1}$, $\boldsymbol{\zeta}_j \in [\mathcal{B}^2]^K$, které maximalizují funkcionál

$$\frac{1}{N}\sum_{i=1}^{N}\langle \boldsymbol{X}_i, \boldsymbol{\zeta}\rangle^2_{[\mathcal{B}^2]^K}$$

$$\text{za podmínek} \quad \|\boldsymbol{\zeta}\|_{[\mathcal{B}^2]^K} = 1 \quad \text{a} \quad \langle \boldsymbol{\zeta}, \boldsymbol{\zeta}_k\rangle_{[\mathcal{B}^2]^K} = 0,\ k < j,$$

kde podmínka ortogonality $\langle \boldsymbol{\zeta}, \boldsymbol{\zeta}_k\rangle_{[\mathcal{B}^2]^K} = 0,\ k < j$ má smysl pouze pro $j \geq 2$.

Hlavní komponenty $\boldsymbol{\zeta}_j$, $j = 1, 2, \ldots$ nalezneme jako vlastní funkce výběrového kovariančního operátoru $V\colon [\mathcal{B}^2]^K \to [\mathcal{B}^2]^K$ takového, že

$$V\boldsymbol{f} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \langle \boldsymbol{X}_i, \boldsymbol{f}\rangle_{[\mathcal{B}^2]^K} \odot \boldsymbol{X}_i \quad \text{pro} \quad \boldsymbol{f} \in [\mathcal{B}^2]^K.$$

Tento operátor má $N-1$ nenulových vlastních čísel, $\gamma_1 > \cdots > \gamma_{N-1}$ [1], která udávají míru variability ve směrech komponent $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{N-1}$.

Při výpočtu vlastní čísel a funkcí budeme postupovat podobně jako u jednorozměrné SFPCA, použijeme tedy clr transformaci. V tomto případě

$$\mathbf{clr}(\boldsymbol{f}) = (\mathrm{clr}(f_i)) \in [L_0^2]^K, \quad \text{pro } \boldsymbol{f} = (f_i) \in [\mathcal{B}^2]^K.$$

Pro transformovaný výběr $\mathbf{clr}(\boldsymbol{X}_1), \ldots, \mathbf{clr}(\boldsymbol{X}_N)$ pak hledáme mnohorozměrné funkcionální hlavní komponenty $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{N-1} \in [L_0^2]^K$ odpovídající nenulovým vlastním číslům a skóry

$$z_{ij}^{MV} = \langle \mathbf{clr}(\boldsymbol{X}_i), \boldsymbol{\xi}_j\rangle_{[L_0^2]^K} = \sum_{k=1}^{K} \langle \mathrm{clr}(X_{ik}), \xi_{jk}\rangle_{L_0^2},$$

$i = 1, \ldots, N,\ j = 1, \ldots, N-1.$

Získané komponenty pak transformujeme zpět do prostoru $[\mathcal{B}^2]^K$ pomocí inverzní clr transformace (definované opět po složkách), tj.

$$\boldsymbol{\zeta}_j = \mathbf{clr}^{-1}(\boldsymbol{\xi}_j) = (\mathrm{clr}^{-1}(\xi_{j1}), \ldots, \mathrm{clr}^{-1}(\xi_{jK}))', \quad j = 1, \ldots, N-1,$$

a vypočteme skóry

$$z_{ij}^{MV} = \langle \boldsymbol{X}_i, \boldsymbol{\zeta}_j \rangle_{[\mathcal{B}^2]^K} = \sum_{k=1}^{K} \langle X_{ik}, \zeta_{jk} \rangle_{\mathcal{B}^2},$$

$i = 1, \ldots, N,\ j = 1, \ldots, N-1.$

Opět zde platí, že skóry v $[\mathcal{B}^2]^K$ jsou totožné jako skóry v $[L_0^2]^K$.

Mnohorozměrnou SFPCA budeme aplikovat na geometrické marginální hustoty dvourozměrných hustot, tj. $\boldsymbol{X}_i \in [\mathcal{B}^2]^2$.

2.3. FPCA pro dvourozměrné hustoty

Dosud jsme vždy uvažovali pouze jednorozměrné hustoty, nyní se však zaměříme na ty dvourozměrné, představíme si funkcionální metodu hlavních komponent pro dvourozměrná funkcionální data [7] a aplikujeme ji na clr transformace dvourozměrných hustot.

Mějme náhodný výběr dvourozměrných funkcí $\widetilde{F}_1, \ldots, \widetilde{F}_N \in L^2$ definovaných na kompaktní množině $\Omega = \Omega_X \times \Omega_Y \subseteq \mathbb{R}^2$. Tento výběr centrujeme a pro $i = 1, \ldots, N$ dostaneme $F_i = \widetilde{F}_i - \overline{F}$, kde $\overline{F} = \frac{1}{N} \sum_{i=1}^{N} \widetilde{F}_i$. Dále definujeme výběrovou kovarianční funkci

$$K(x_1, y_1; x_2, y_2) = \frac{1}{N} \sum_{i=1}^{N} F_i(x_1, y_1)\, F_i(x_2, y_2), \quad (x_1, y_1), (x_2, y_2) \in \Omega$$

a předpokládáme, že tato funkce je spojitá na $\Omega \times \Omega$ a její druhé mocniny mají konečný integrál.

Uvažujeme funkci $\phi(x, y)$ na Ω takovou, že $\|\phi\|^2 = \int_\Omega \phi^2(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$. Projekce $F_i(x, y)$ na $\phi(x, y)$ označíme $z_i = \int_\Omega F_i(x, y)\,\phi(x, y)\,\mathrm{d}x\,\mathrm{d}y$, $i = 1, \ldots, N$, a jejich výběrový rozptyl můžeme spočítat jako

$$\int_\Omega \int_\Omega \phi(x_1, y_1)\, K(x_1, y_1; x_2, y_2)\, \phi(x_2, y_2)\ \mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2.$$

První hlavní komponentu hledáme jako funkci ϕ_1, která maximalizuje tento rozptyl, tj.

$$\phi_1 = \operatorname*{arg\,max}_{\phi:\|\phi\|=1} \int_\Omega \int_\Omega \phi(x_1, y_1)\, K(x_1, y_1; x_2, y_2)\, \phi(x_2, y_2)\ \mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2.$$

Další komponenty získáme podobným způsobem, musí pro ně však navíc platit

$$\int_\Omega \phi_j(x, y)\, \phi_k(x, y)\,\mathrm{d}x\,\mathrm{d}y = 0 \quad \text{pro } j \neq k.$$

Tyto hlavní komponenty můžeme jednoduše získat pomocí singulárního rozkladu kovarianční funkce $K(x_1, y_1; x_2, y_2)$ jako ortonormální posloupnost funkcí $\{\phi_j\}_{j\geq 1}$, k níž existuje nerostoucí posloupnost kladných čísel $\{\kappa_j\}_{j\geq 1}$ taková, že

$$\int_\Omega K(x_1, y_1; x_2, y_2)\, \phi_j(x_2, y_2)\,\mathrm{d}x_2\,\mathrm{d}y_2 = \kappa_j\, \phi_j(x_1, y_1)$$

a

$$K(x_1, y_1; x_2, y_2) = \sum_{j=1}^{\infty} \kappa_j \, \phi_j(x_1, y_1) \, \phi_j(x_2, y_2),$$

kde posloupnosti $\{\kappa_j\}_{j\geq 1}$ a $\{\phi_j\}_{j\geq 1}$ jsou vlastní čísla a vlastní funkce kovarianční funkce $K(x_1, y_1; x_2, y_2)$.

Skóry pak určíme jako projekce funkcí F_i na jednotlivé hlavní komponenty ϕ_j, tj.

$$z_{ij}^{BD} = \int_{\Omega} F_i(x, y) \, \phi_j(x, y) \, \mathrm{d}x \, \mathrm{d}y, \quad i = 1, \ldots, N, \; j = 1, 2, \ldots.$$

Máme-li náhodný výběr dvourozměrných hustot $X_1, \ldots, X_N \in \mathcal{B}^2(\lambda)$ ($\lambda = \lambda_X \times \lambda_Y$), pak stačí položit $F_i = \mathrm{clr}(X_i), \; i = 1, \ldots, N$ a použít výše popsaný postup.

Integrál kovarianční funkce $K(x_1, y_1; x_2, y_2)$ je v případě clr hustot nulový, neboť

$$\begin{aligned}
&\int_{\Omega}\int_{\Omega} K(x_1, y_1; x_2, y_2) \; \mathrm{d}x_1 \, \mathrm{d}y_1 \, \mathrm{d}x_2 \, \mathrm{d}y_2 = \\
&\quad = \int_{\Omega}\int_{\Omega} \frac{1}{N} \sum_{i=1}^{N} \mathrm{clr}(X_i)(x_1, y_1) \, \mathrm{clr}(X_i)(x_2, y_2) \; \mathrm{d}x_1 \, \mathrm{d}y_1 \, \mathrm{d}x_2 \, \mathrm{d}y_2 = \\
&\quad = \frac{1}{N} \sum_{i=1}^{N} \left[\int_{\Omega} \mathrm{clr}(X_i)(x_1, y_1) \; \mathrm{d}x_1 \, \mathrm{d}y_1 \right] \left[\int_{\Omega} \mathrm{clr}(X_i)(x_2, y_2) \; \mathrm{d}x_2 \, \mathrm{d}y_2 \right] = 0,
\end{aligned}$$

tedy $K \in L_0^2$ na $\Omega \times \Omega$. Proto vlastní funkce $\phi_j, j = 1, 2, \ldots$ mají také nulový integrál, tj. $\phi_j \in L_0^2$ na Ω.

V této práci budeme pomocí dvourozměrné FPCA analyzovat nezávislé a interakční části hustot.

3. Zpracování reálných dat

Nyní si ukážeme, jak lze výše popsané metody aplikovat na konkrétní soubor dat. Představíme si data vybraná k analýze, použijeme je k vytvoření dvourozměrných hustot, a ty pak budeme rozkládat a analyzovat. Všechny kódy v `R`, které byly k analýze dat použity, jsou nahrány na GitHub[1].

3.1. Popis datového souboru a jeho úprava

Analyzovaná data se týkají letů vypravených v roce 2013 ze tří newyorských letišť:

- Newark Liberty International Airport (EWR),
- John F. Kennedy International Airport (JFK),
- LaGuardia Airport (LGA)

a jsou volně k dispozici v balíčku `nycflights13` v softwaru `R`.

Budou nás zajímat následující informace:

- zpoždění při odletu z New Yorku a při příletu do cílové destinace (hodnoty v intervalech $[-43, 1301]$ a $[-86, 1272]$ minut),
- měsíc, ve kterém let proběhl (leden až prosinec 2013),
- vzdálenost mezi newyorským letištěm a cílovou destinací (hodnoty v intervalu $[80, 4983]$ mil) – pro zjednodušení si vytvoříme čtyři vzdálenostní kategorie: 0–500, 501–1000, 1001–2000, 2001+ mil (jen pro představu: 1 míle je přibližně 1.6093 kilometrů),
- název letiště, ze kterého byl let vypraven (letiště EWR, JFK, LGA).

[1] `https://github.com/adelaczolkova/AnalyzaLetu`

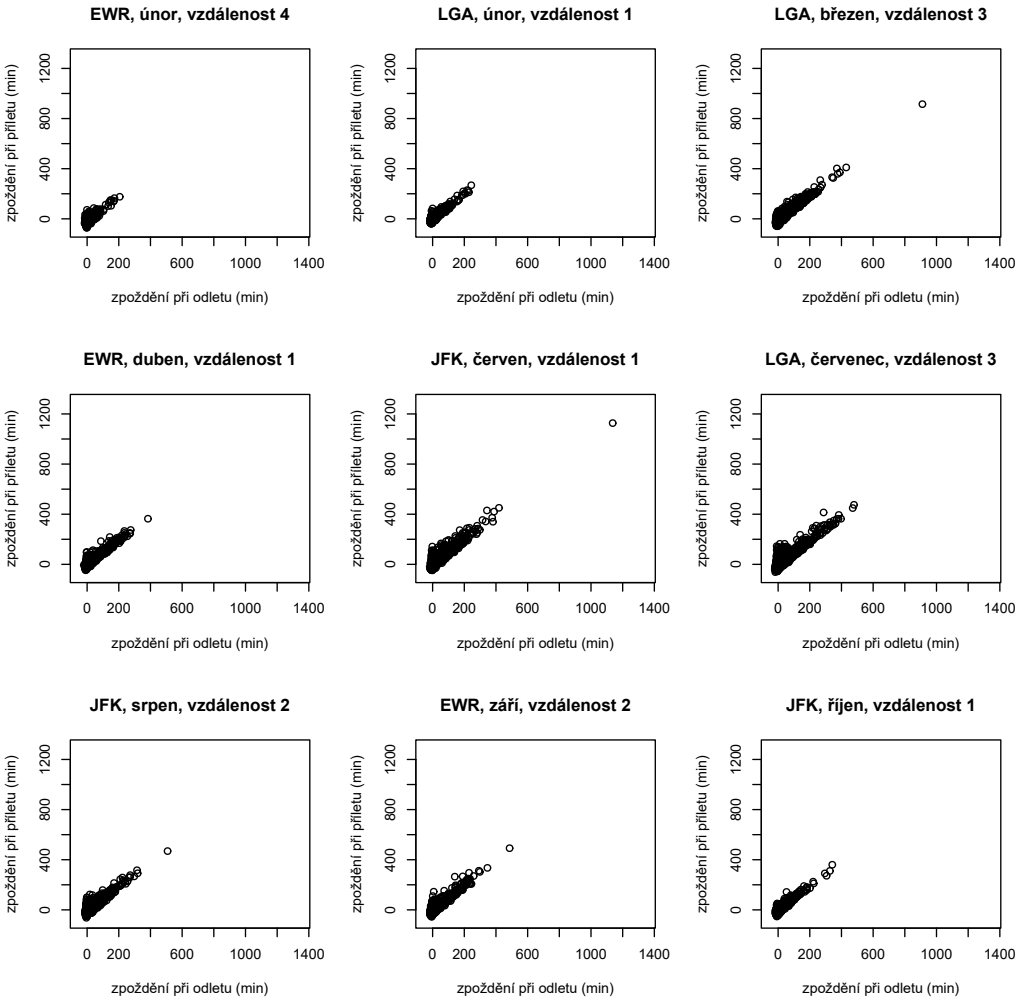
V závislosti na měsíci, vzdálenostní kategorii a letišti si lety rozdělíme do 144 skupin. Ve skutečnosti však budeme pracovat pouze se 132 skupinami, neboť z letiště LaGuardia nebyly v žádném měsíci vypraveny lety ve čtvrté vzdáleností kategorii (2001+ mil), tj. $144 - 12 = 132$.

Vzhledem k tomu, že kdyby zde bylo zobrazeno 132 grafů, byla by tato práce velmi nepřehledná, vybereme jen grafy pro několik skupin, na které se zde podíváme. Vybranými skupinami jsou odlety z letišť:

- Newark v únoru do destinací ve 4. vzdálenostní kategorii (EWR, únor, vzálenost 4),
- LaGuardia v únoru do destinací v 1. vzdálenostní kategorii (LGA, únor, vzdálenost 1),
- LaGuardia v březnu do destinací ve 3. vzdálenostní kategorii (LGA, březen, vzdálenost 3),
- Newark v dubnu do destinací v 1. vzdálenostní kategorii (EWR, duben, vzdálenost 1),
- JFK v červnu do destinací v 1. vzdálenostní kategorii (JFK, červen, vzdálenost 1),
- LaGuardia v červenci do destinací ve 3. vzdálenostní kategorii (LGA, červenec, vzdálenost 3),
- JFK v srpnu do destinací ve 2. vzdálenostní kategorii (JFK, srpen, vzdálenost 2),
- Newark v září do destinací ve 2. vzdálenostní kategorii (EWR, září, vzádlenost 2),

- JFK v říjnu do destinací v 1. vzdálenostní kategorii (JFK, říjen, vzdálenost 1).

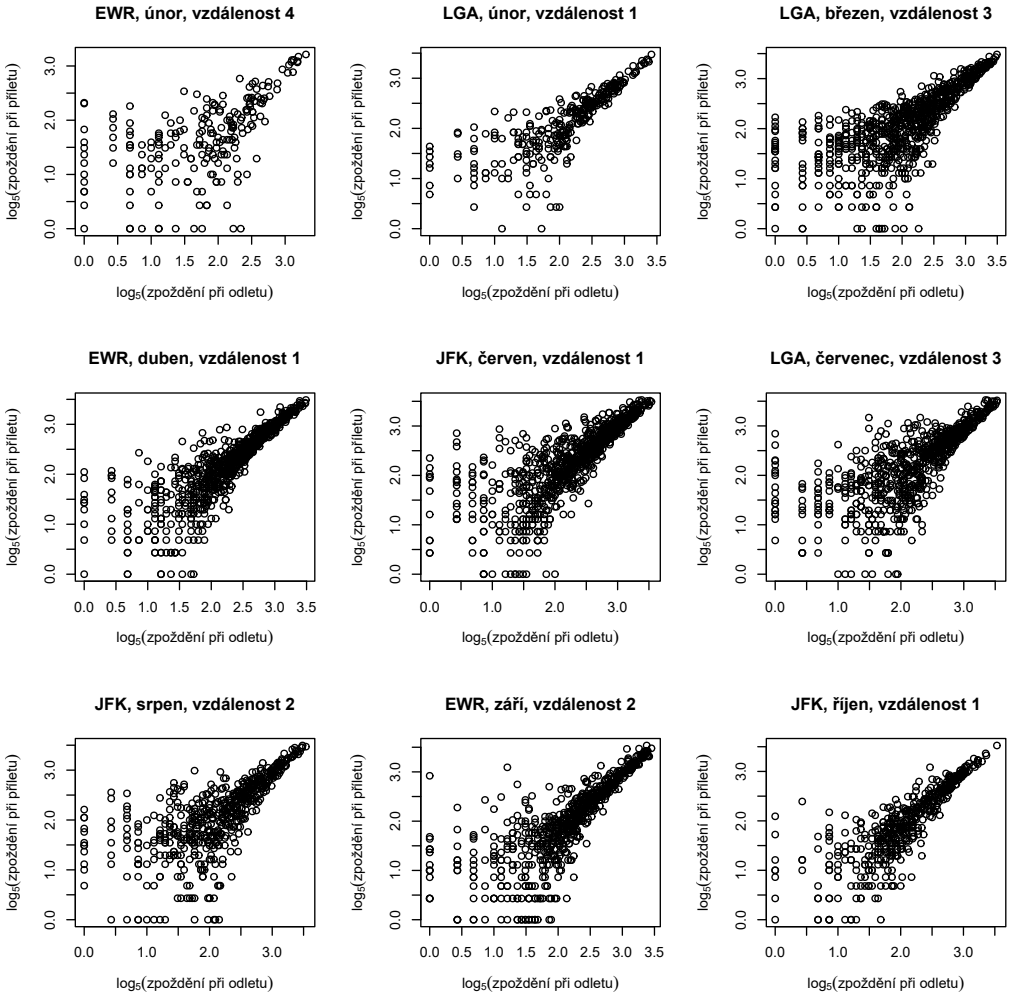
Zobrazíme-li si zpoždění při odletu a příletu v jednotlivých skupinách, uvidíme, že v datech je málo velmi opožděných letů (několikahodinové zpoždění při odletu/příletu), a také že mezi zpožděními je vysoká korelace (obrázek 1).



Obrázek 1: Všechna zpoždění ve vybraných skupinách.

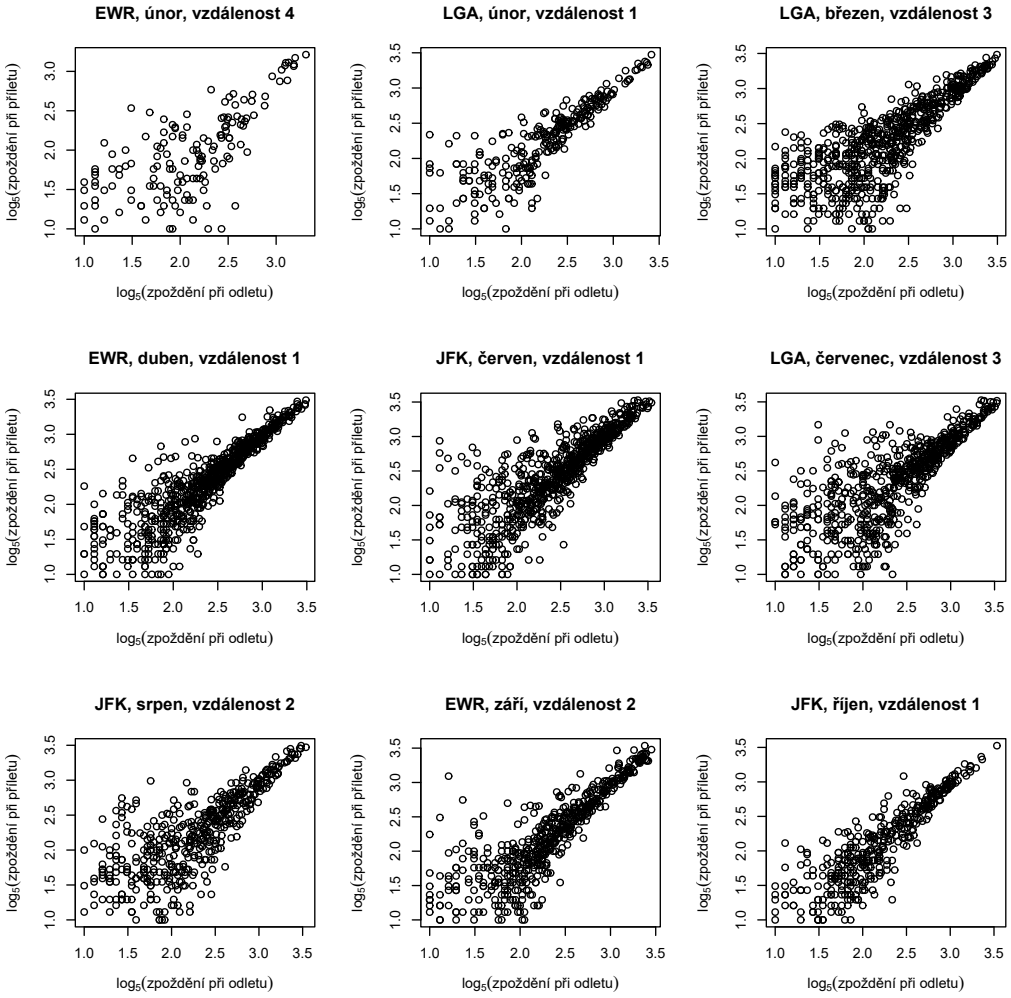
Abychom mohli s daty lépe pracovat, vynecháme lety se zpožděním při od-

letu nebo příletu větším než 300 minut (tj. 5 hodin). Další úpravou, kterou provedeme, je logaritmizace zpoždění. Abychom ji však mohli provést, musíme nejprve ještě vynechat lety se záporným a nulovým zpožděním (dřívější a včasné odlety/přílety, kterých v datech příliš mnoho není). Na obrázku 2 můžeme vidět skoky mezi zlogaritmovanými hodnotami malých zpoždění, které působí poněkud nepřirozeně a během odhadu hustot by mohly způsobit problémy. Vynecháme proto ještě lety se zpožděním menším než pět minut.



Obrázek 2: Zlogaritmovaná zpoždění v intervalu 1 až 300 minut.

Logaritmus se základem pět použijeme proto, aby hodnoty zlogaritmovaného zpoždění začínaly jedničkou (mohli bychom samozřejmě zvolit logaritmus s jakýmkoli jiným základem), výsledná zlogaritmovaná zpoždění, se kterými budeme dále pracovat, tedy budou v intervalu $[1, 3.54]$ a vidíme je na obrázku 3.



Obrázek 3: Zlogaritmovaná zpoždění v intervalu 5 až 300 minut.

Můžeme si všimnout, že se počty pozorovaných zpoždění v jednotlivých zobrazených skupinách liší. Také vidíme, že se zpoždění během letu může

změnit, v našem případě to je dobře vidět u menších zpoždění, protože jsme hodnoty zlogaritmovali. Obecně ale můžeme říct, že mezi zpožděními je kladná korelace, tj. čím větší zpoždění při odletu, tím větší zpoždění při příletu.

3.2. Vytvoření (odhad) hustot a jejich clr transformace

Jakmile máme data upravená, můžeme se pustit do odhadu dvourozměrných hustot pravděpodobnosti pro zpoždění při odletu a příletu v jednotlivých skupinách (132 skupin). Použijeme k tomu jádrový odhad hustoty s Gaussovským jádrem (normální rozdělení).

Označíme-li $\mathbf{x}_1, \ldots, \mathbf{x}_n$ naše data o zpoždění jednotlivých letů (ve zvolené skupině), kde $\mathbf{x}_i = (x_{i1}, x_{i2})'$, x_{i1} je zpoždění při odletu a x_{i2} při příletu, $i = 1, \ldots, n$. Odhad hustoty f pro zpoždění $\mathbf{x} = (x_1, x_2)'$ pak získáme jako

$$\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right),$$

kde matice $\mathbf{H} > \mathbf{0}$ řádu 2 je vyhlazovací parametr, pro jednoduchost budeme předpokládat, že se jedná od diagonální matici s hodnotami $h_X, h_Y > 0$ na hlavní diagonále, potom determinant $|\mathbf{H}| = h_X h_Y$. K je (nezáporná) jádrová funkce, v našem případě hustota dvourozměrného normovaného normálního rozdělení, tj.

$$K(\mathbf{x}) = \frac{1}{2\pi} e^{-\frac{1}{2}\mathbf{x}'\mathbf{x}}, \quad \mathbf{x} \in \mathbb{R}^2,$$

proto

$$K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right) = \frac{1}{2\pi\sqrt{h_X h_Y}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)'\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)}.$$

V softwaru `R` lze pro výpočet jádrového odhadu hustoty použít funkci `bkde2D()` z balíčku `KernSmooth`. Tato funkce má několik parametrů:

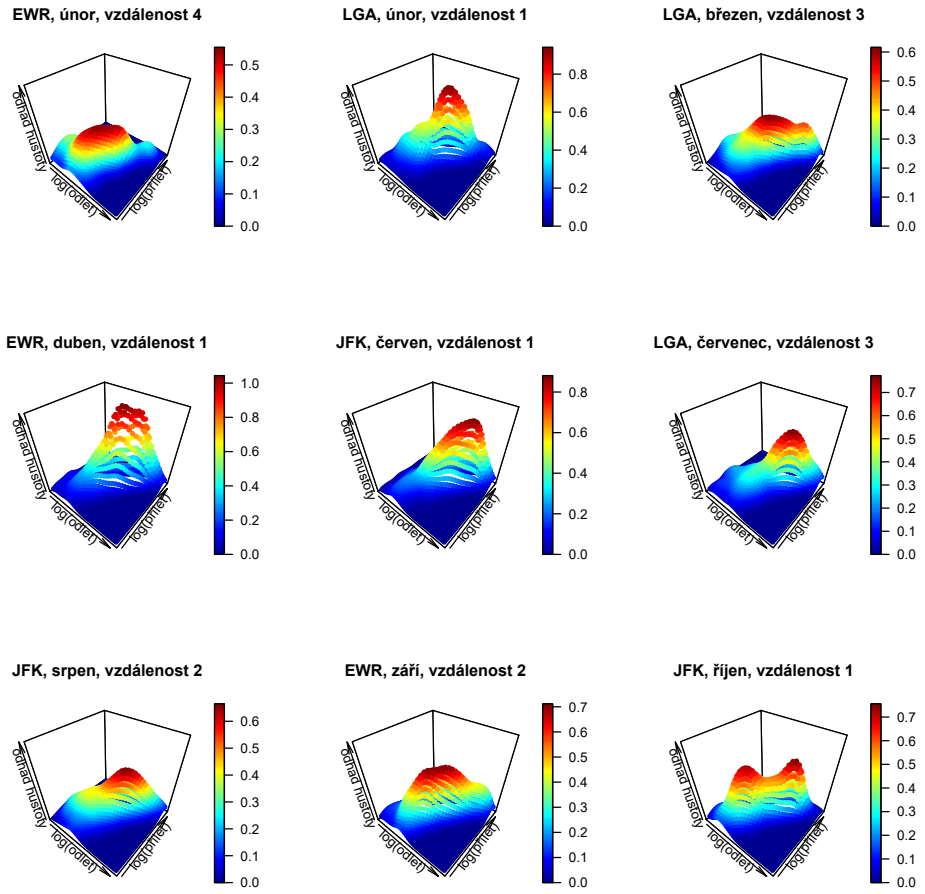
- `x` – datová matice, která má dva sloupce (zpoždění při odletu a při příletu),
- `bandwidth` – dvousložkový vektor vyhlazovacích parametrů h_X, h_Y, tzv. šířka okna pro každou souřadnicovou osu (zpoždění při odletu/příletu), jeho volbě se budeme věnovat níže,
- `gridsize` – vektor počtů bodů mřížky pro každou souřadnicovou osu (zvolíme mřížku 40x40, tj. na každé ose bude rovnoměrně rozmístěno 40 bodů, ve kterých budou spočítány odhady hustoty),
- `range.x` – seznam obsahující dva dvousložkové vektory, které určují minimální a maximální hodnotu bodů mřížky na každé souřadnicové ose (obě zpoždění jsou v intervalu $[1, 3.54]$),
- `truncate` – logický parametr, pokud je jeho hodnota `TRUE`, pak jsou data mimo rozsah `range.x` ignorována (v našem případě pracujeme pouze z daty, která jsou v požadovaném rozsahu, tudíž žádná nebudou vynechána).

K volbě parametru `bandwidth` využijeme v `R` funkci `bw.nrd()`, která optimální hodnotu parametru h pro zvolenou souřadnicovou osu určuje jako

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06 \cdot \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) n^{-1/5},$$

kde $\hat{\sigma}$ je odhad směrodatné odchylky na dané ose a n je počet pozorovaných hodnot (viz výše). Hodnotu h tedy počítáme zvlášť pro zpoždění při odletu (h_X) a při příletu (h_Y).

Hustoty jsou sice spojité funkce, pomocí funkce `bkde2D()` však dostaneme pouze jejich hodnoty na zvolené mřížce, takže dále budeme pracovat pouze s těmito hodnotami – diskretizovanými hustotami (obrázek 4). Kdybychom chtěli pracovat se spojitými odhady hustot, mohli bychom je reprezentovat pomocí splajnů, těm se však zde věnovat nebudeme.



Obrázek 4: Diskretizované hustoty pro zlogaritmovaná zpoždění v intervalu 5–300 minut ve vybraných skupinách.

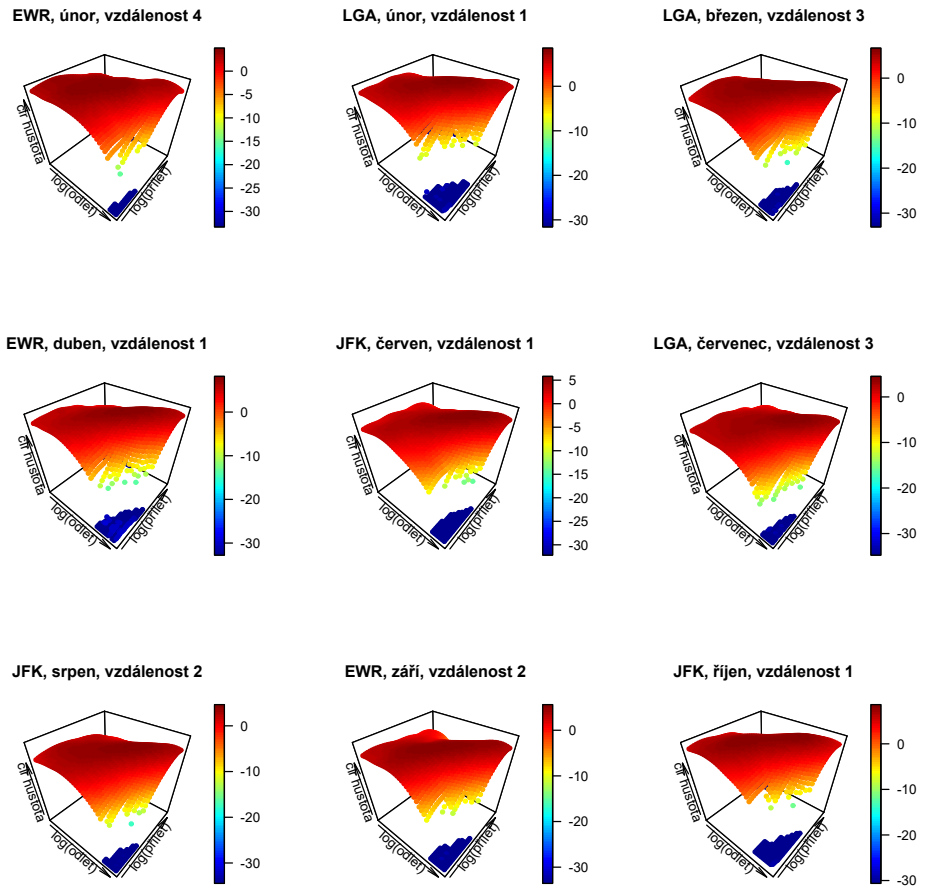
Na obrázku 4 jsou zobrazeny diskretizované hustoty pro vybrané skupiny letů. Na vodorovných osách jsou zpoždění při odletu a při příletu, na svislé

ose jsou hodnoty odhadovaných hustot, hodnoty na osách rostou ve směru šipek a jejich rozsahy jsou pro všechny skupiny stejné, naopak barevné škály jsou pro jednotlivé skupiny různé (toto bude platit i pro další podobné grafy).

Na první pohled si můžeme všimnout, že se jednotlivé hustoty mezi sebou více či méně liší, přesto všechny dobře zachycují již dříve zmíněnou kladnou korelaci mezi zpožděními. Vidíme také například podobnost mezi hustotami pro lety do destinací v 1. vzdálenostní kategorii, které jsou oproti ostatním špičatější. Naopak asi nejméně špičatá je hustota pro lety z letiště Newark v únoru do destinací ve 4. vzdálenostní kategorii, což by mohlo znamenat rovnoměrnější rozdělení zpoždění letů v této skupině, avšak může to být také důsledek malého počtu pozorování, která v této skupině máme. Dále se podívejme třeba na hustotu pro lety z letiště JFK v říjnu do destinací v 1. vzdálenostní kategorii, která má na rozdíl od ostatních dva vrcholy, dalo by se tedy říct, že v této skupině se nejčastěji vyskytují dva typy letů – buďto lety s malým (několikaminutovým) zpožděním při odletu i příletu, nebo lety s velkým (více než hodinovým) zpožděním při odletu i příletu, a toto zpoždění se během letu výrazně nezmění.

Diskretizované hustoty bychom dále chtěli transformovat z Bayesova prostoru $\mathcal{B}^2$ do prostoru L_0^2 pomocí clr transformace uvedené v kapitole 1.1, kterou přizpůsobíme tomu, že pracujeme s diskretizovanými hustotami (viz níže). Před provedením transformace se však musíme ještě zaměřit na samotné odhadnuté hodnoty hustot v bodech mřížky, protože některé tyto hodnoty jsou velmi blízké nule (řádu 10^{-16} a nižšího), pravděpodobně se ve skutečnosti jedná o nuly, které nejsou nulami kvůli zaokrouhlování během výpočtů, a mezi těmito hodnotami a dalšími skutečně nenulovými hodnotami hustot (řádu 10^{-12} a vyššího) je skok, který by dále ovlivnil podobu clr hustot (obrázek 5). Musíme se proto zamyslet, jak zdánlivě nenulové hodnoty

upravit, aby neměly výrazný vliv na podobu clr hustot.



Obrázek 5: Diskretizované clr hustoty bez dalších úprav pro zlogaritmovaná zpoždění v intervalu 5–300 minut ve vybraných skupinách.

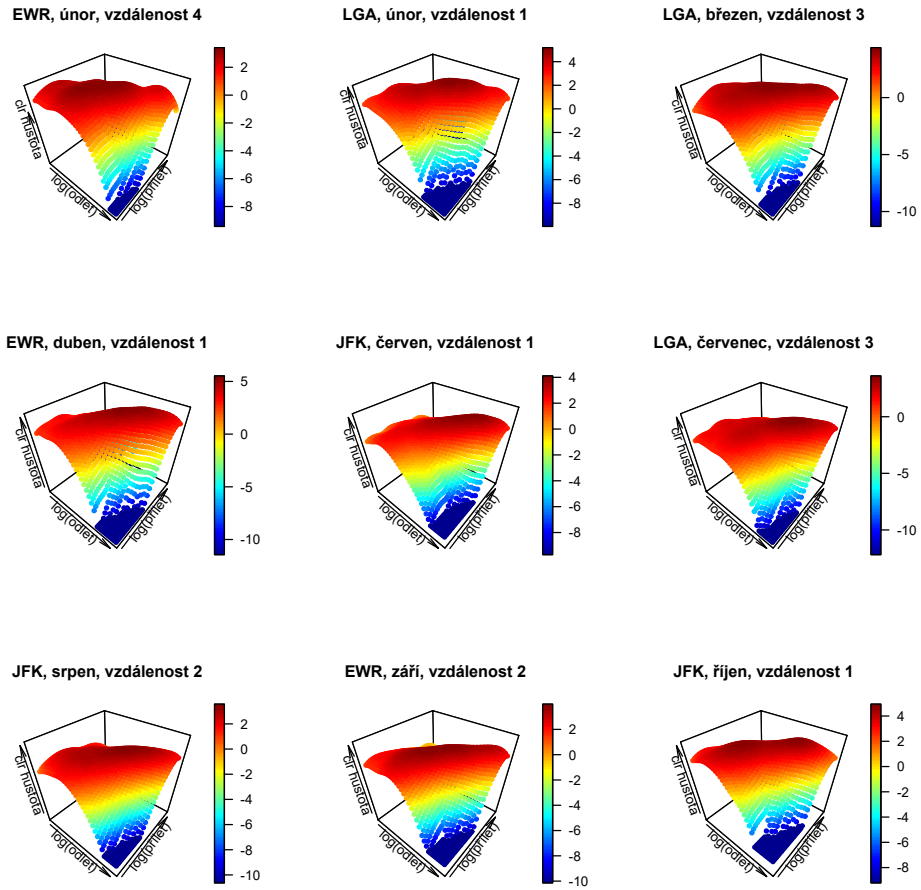
Heuristickou metodou jsme se rozhodli vyřešit tento problém tak, že pro každou ze 132 hustot spočítáme průměr 20 jejích nejmenších hodnot větších než 10^{-13} a hodnoty hustoty menší než tento vypočtený průměr nahradíme právě tímto průměrem.

Nyní už můžeme bez obav provést clr transformaci, označíme-li hodnoty

hustot v bodech mřížky $\widehat{f}_{jk},\ j,k = 1,2,\ldots,40$, potom

$$\mathrm{clr}(\widehat{f}_{jk}) = \ln \widehat{f}_{jk} - \frac{1}{40 \cdot 40} \sum_{r=1}^{40} \sum_{s=1}^{40} \ln \widehat{f}_{rs}, \quad j,k = 1,\ldots,40,$$

a získáme tak diskretizované clr hustoty, které budeme dále analyzovat.



Obrázek 6: Diskretizované clr hustoty po nahrazení zdánlivě nenulových hodnot.

Na obrázku 6 vidíme, že po nahrazení zdánlivě nenulových hodnot zmizel výrazný skok mezi hodnotami clr hustot, který jsme pozorovali na obrázku

5. Také si všimneme, že všechny clr hustoty jsou si na první pohled velmi podobné, a to i přesto, že některé původní (netransformované) hustoty se od sebe výrazně lišily (obrázek 4). I v zde však můžeme stále dobře vidět, že mezi zpožděními je kladná korelace.

3.3. Ortogonální rozklad a rozklad interakční hustoty

Ještě než přistoupíme k samotným rozkladům dvourozměrných hustot, podíváme se podrobněji na jednotlivé typy marginálních hustot popsané v kapitole 1.2 a ukážeme si, že se aritmetické a geometrické marginální hustoty (prvky Bayesových prostorů) skutečně liší.

Nejprve vypočítáme clr marginální hustoty jednotlivých clr hustot, pro diskretizované hustoty máme

$$\mathrm{clr}(\widehat{f}_{X,g,j}) = \frac{1}{40}\sum_{k=1}^{40}\mathrm{clr}(\widehat{f}_{jk}), \quad j = 1,\ldots,40,$$
$$\mathrm{clr}(\widehat{f}_{Y,g,k}) = \frac{1}{40}\sum_{j=1}^{40}\mathrm{clr}(\widehat{f}_{jk}), \quad k = 1,\ldots,40,$$

kde X používáme pro zpoždění při odletu a Y pro zpoždění při příletu, jak už bylo naznačeno v kapitole 3.2.

Odhad geometrických marginálních hustot v Bayesově prostoru $\mathcal{B}^2$ získáme jako

$$\widehat{f}_{X,g,j} = \exp\{\mathrm{clr}(\widehat{f}_{X,g,j})\}\Big/\sum_{l=1}^{40} h\cdot\exp\{\mathrm{clr}(\widehat{f}_{X,g,l})\}, \quad j = 1,\ldots,40,$$
$$\widehat{f}_{Y,g,k} = \exp\{\mathrm{clr}(\widehat{f}_{Y,g,k})\}\Big/\sum_{l=1}^{40} h\cdot\exp\{\mathrm{clr}(\widehat{f}_{Y,g,l})\}, \quad k = 1,\ldots,40,$$

kde h je vzdálenost mezi body mřížky (na obou osách stejná).

Aritmetické marginální hustoty odhadneme s pomocí jádrové funkce, zde se bude jednat o hustotu jednorozměrného normovaného normálního rozdělení, tj.

$$K(x) = \frac{1}{2\pi} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R},$$

potom odhad jednorozměrné hustoty pro zpoždění x (při odletu, nebo příletu)

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right),$$

kde $h > 0$ je vyhlazovací parametr. Za x_i, $i = 1, \ldots, n$, dosazujeme buď zpoždění při odletu x_{i1}, nebo zpoždění při příletu x_{i2} (viz značení v kapitole 3.2). V `R` pro odhad využijeme funkci `bkde()` z balíčku `KernSmooth`, která pracuje podobně jako již dříve popsaná funkce `bkde2D()`, a k odhadu vyhlazovacího parametru h použijeme opět funkci `bw.nrd()`. Stejně jako v případě dvourozměrných hustot získáme diskretizované hustoty

- $\widehat{f}_{X,a,j},\ j = 1, \ldots, 40$ pro zpoždění při odletu,
- $\widehat{f}_{Y,a,k},\ k = 1, \ldots, 40$ pro zpoždění při příletu.

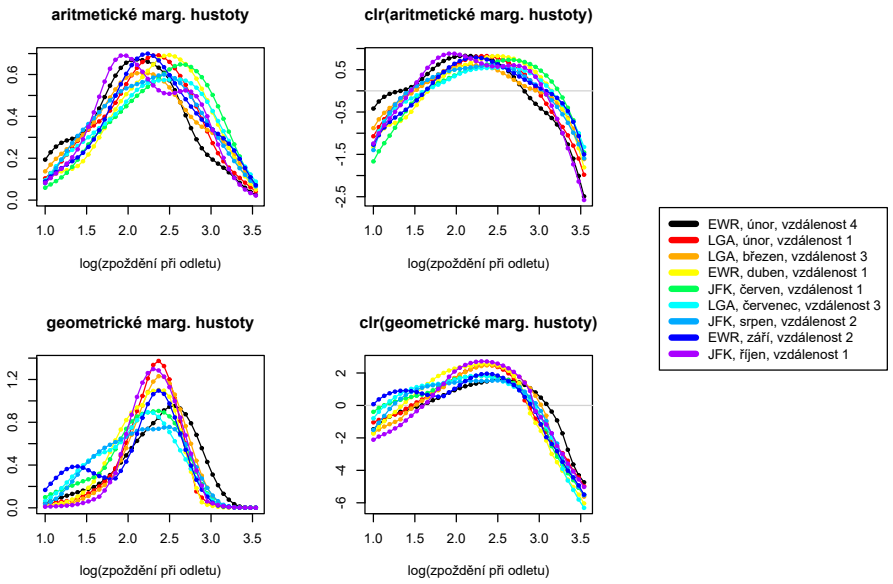
I tyto odhady aritmetických marginálních hustot můžeme transformovat pomocí clr transformace, tedy

$$\mathrm{clr}(\widehat{f}_{X,a,j}) = \ln \widehat{f}_{X,a,j} - \frac{1}{40} \sum_{l=1}^{40} \ln \widehat{f}_{X,a,l}, \quad j = 1, \ldots, 40,$$

$$\mathrm{clr}(\widehat{f}_{Y,a,k}) = \ln \widehat{f}_{Y,a,k} - \frac{1}{40} \sum_{l=1}^{40} \ln \widehat{f}_{Y,a,l}, \quad k = 1, \ldots, 40.$$

Odhadnuté diskretizované aritmetické i geometrické marginální hustoty si zobrazíme spolu s jejich clr transformacemi; z obrázků 7 a 8 je zřejmé, že

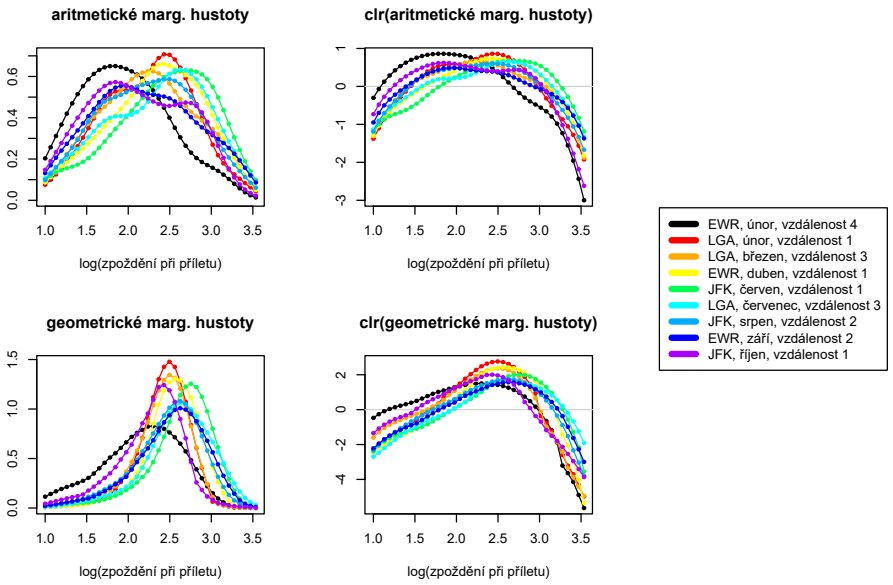
jsou tyto dva typy marginálních hustot opravdu odlišné a liší se i jejich clr hustoty. Obecně lze říct, že aritmetické marginální mají těžší chvosty, zatímco ty geometrické jsou špičatější, a navíc u nich můžeme vidět výraznější rozdíly ve výškách jednotlivých hustot, po clr transformaci je už však nepozorujeme.



Obrázek 7: Aritmetické a geometrické marginální hustoty a jejich clr transformace pro zpoždění při odletu pro vybrané skupiny letů.

Lze také porovnat aritmetické a geometrické marginální hustoty pro konkrétní skupinu letů. Například si všimněme (obrázek 7), že aritmetická marginální hustota zpoždění při odletu pro lety z letiště JFK v říjnu do destinací v 1. vzdálenostní kategorii (fialová) je mírně zvlněná, to se projeví i po clr transformaci, i když ne příliš výrazně. Podíváme-li se na odpovídající geometrickou marginální hustotu, žádné zvlnění u ní nepozorujeme. Tento efekt je pak vidět i u zpoždění při příletu, avšak zde není tak výrazný (obrázek 8). Opačná situace nastala u letů z letiště Newark v září do destinací ve 2. vzdálenostní kategorii (tmavě modrá), zde pozorujeme zvlnění u geometrické marginální hustoty pro zpoždění při odletu a nikoliv u té aritmetické (obrázek

7), u zpoždění při příletu ale už nic takového nepozorujeme (obrázek 8).



Obrázek 8: Aritmetické a geometrické marginální hustoty a jejich clr transformace pro zpoždění při příletu pro vybrané skupiny letů.

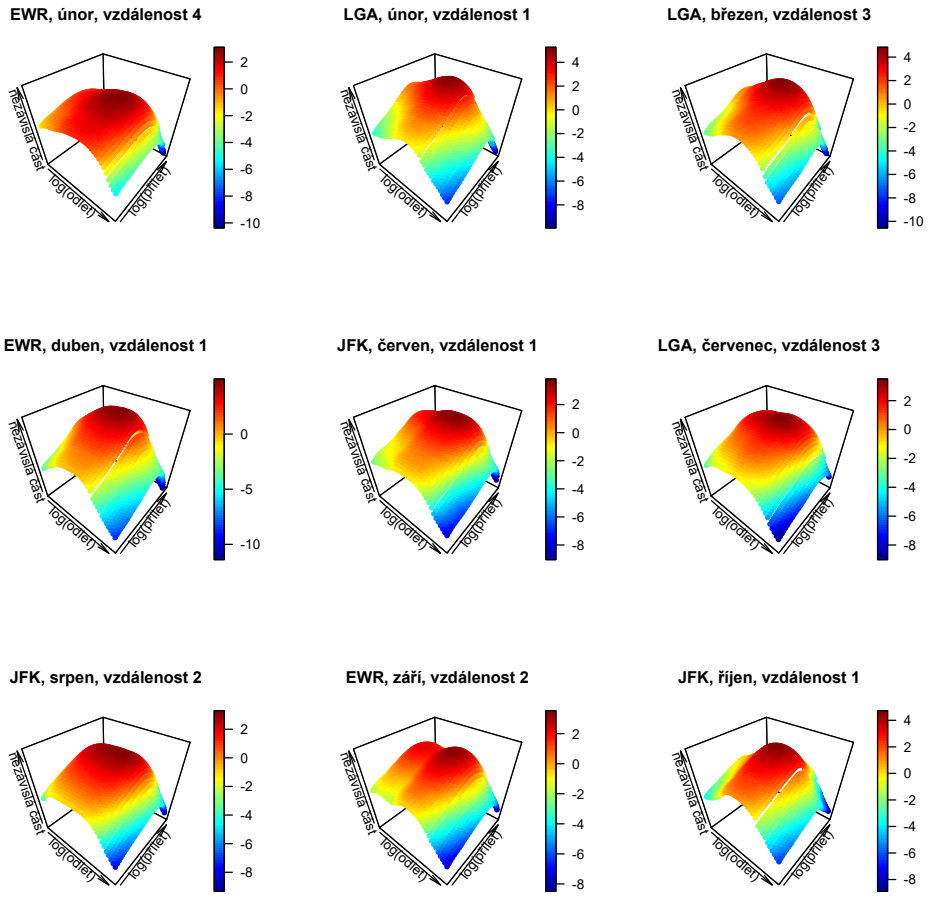
Nyní už se zaměříme na ortogonální rozklad odhadnutých dvourozměrných hustot. Veškeré výpočty budeme provádět v clr prostoru L_0^2, budeme tedy pracovat s clr hustotami, a i když zde a v následujících kapitolách této práce budeme hovořit o hustotách, budeme mít na mysli jejich clr transformace, pokud nebude uvedeno jinak.

Nezávislou část hustoty určíme jako součet clr marginálních hustot a interakční část jako rozdíl původní clr hustoty a její nezávislé části, jak bylo popsáno v kapitole 1.3. Pro diskretizované hustoty provedeme výpočty takto:

$$\begin{aligned}
\operatorname{clr}(\widehat{f}_{\mathrm{ind},jk}) &= \operatorname{clr}(\widehat{f}_{X,g,j}) + \operatorname{clr}(\widehat{f}_{Y,g,k}),\\
\operatorname{clr}(\widehat{f}_{\mathrm{int},jk}) &= \operatorname{clr}(\widehat{f}_{jk}) - \operatorname{clr}(\widehat{f}_{\mathrm{ind},jk}) = \operatorname{clr}(\widehat{f}_{jk}) - \operatorname{clr}(\widehat{f}_{X,g,j}) - \operatorname{clr}(\widehat{f}_{Y,g,k}),
\end{aligned}$$

kde $j, k = 1, \ldots, 40$.

Obě složky rozkladu si opět zobrazíme pro vybrané skupiny letů. Na obrázku 9 vidíme nezávislé části hustot ($\mathrm{clr}(\widehat{f}_{\mathrm{ind}})$); stejně jako u clr hustot (obrázek 6) jsou si tyto také velmi podobné, avšak při bližším pohledu na ně jsme schopni mezi nimi najít určité rozdíly.

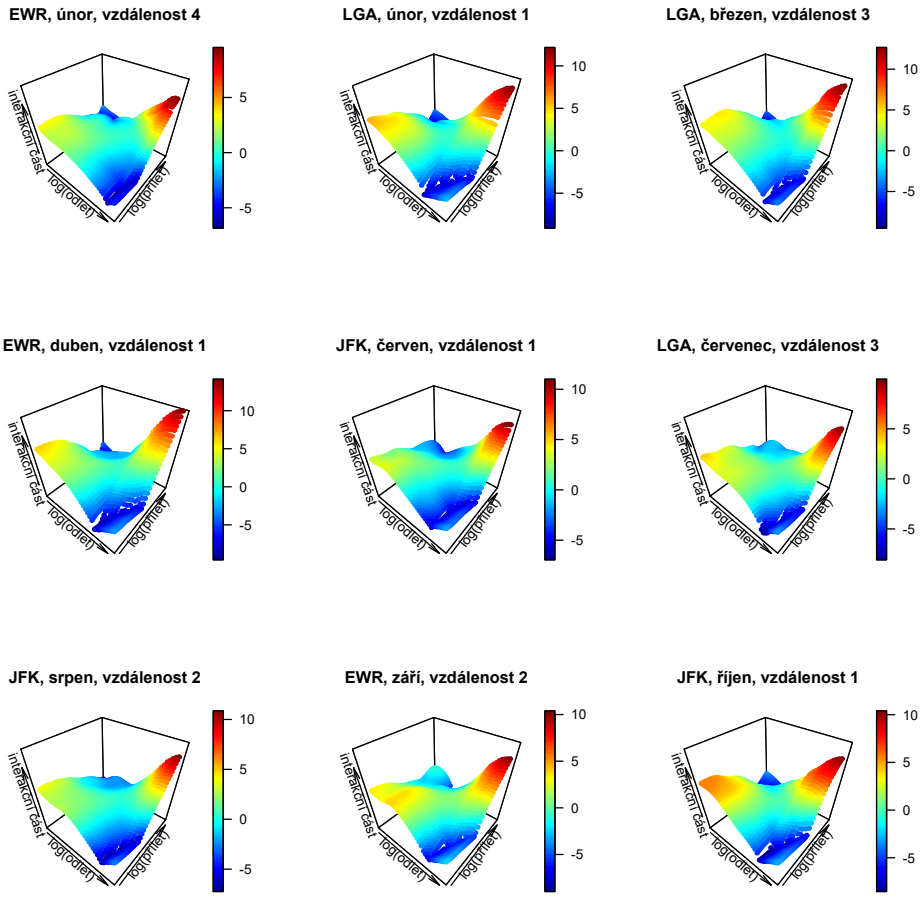


Obrázek 9: Diskretizované nezávislé části hustot vybraných skupin letů.

Je také dobře vidět, že nezávislé části hustot jsou složeny z clr (geometrických) marginálních hustot pro jednotlivá zpoždění. Například u hustoty pro lety z letiště Newark v září do destinací ve 2. vzdálenostní kategorii si můžeme všimnout, že se zde projevilo zvlnění pozorované u geometrické

marginální hustoty pro zpoždění při odletu (obrázek 7).

Obrázek 10 je pak věnován interakčním částem hustot ($\mathrm{clr}(\widehat{f}_{\mathrm{int}})$), i ty mají ve všech zobrazených skupinách podobnou strukturu.



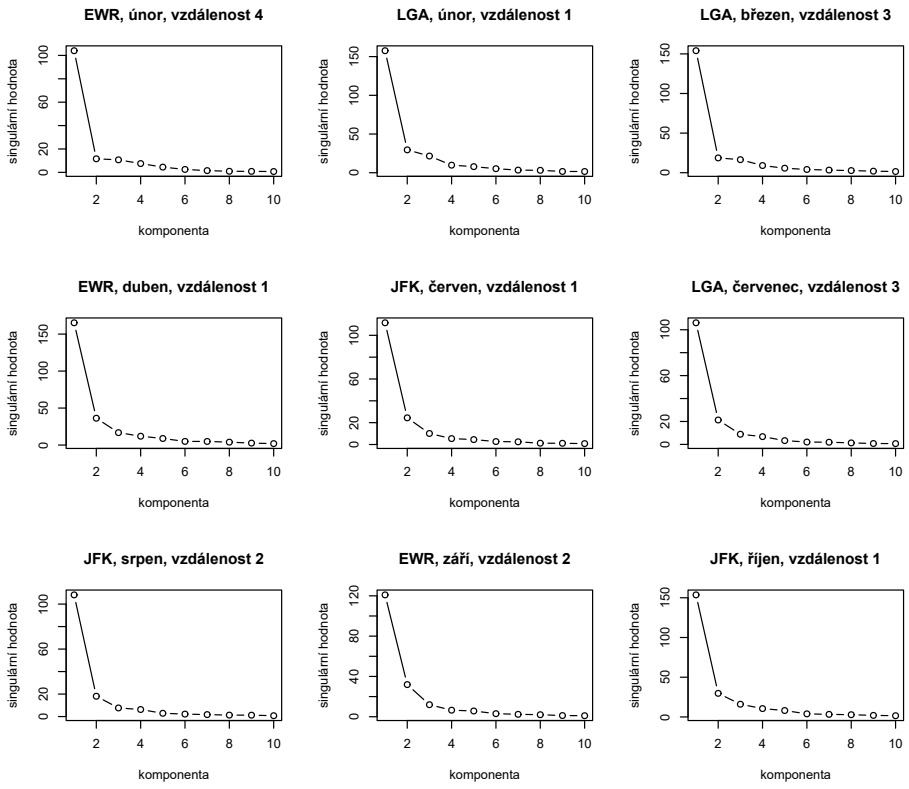
Obrázek 10: Diskretizované interakční části hustot vybraných skupin letů.

Z jednotlivých grafů můžeme snadno vyčíst, že existuje velká závislost mezi velkými zpožděními, tzn. je-li let významně opožděn již při odletu, potom lze očekávat velké zpoždění i při příletu do cílové destinace (viz velké hodnoty interakčních částí hustot). Závislost pozorujeme také mezi malými zpožděními, tj. je-li let vypraven přibližně v plánovaném čase, nebývá ani na pří-

letu hodně opožděn. U letů z letiště Newark v září do destinací ve 2. vzdálenostní kategorii můžeme zaznamenat malou odlišnost od ostatních zobrazených skupin – je-li zpoždění při odletu v řádu několika minut (přibližně 20–30 minut), pak se dá očekávat, že se toto zpoždění během letu nezvýší, nebo se dokonce sníží.

Dále provedeme singulární rozklad interakční hustoty, jemuž byla věnována kapitola 1.4. Pro diskretizované interakční hustoty vypadá rozklad takto:

$$\mathrm{clr}(\widehat{f}_{\mathrm{int},jk}) = \sum_{i=1}^{\infty} \gamma_i^{1/2} \varphi_{i,j} \psi_{i,k}, \quad j,k = 1,\ldots,40.$$

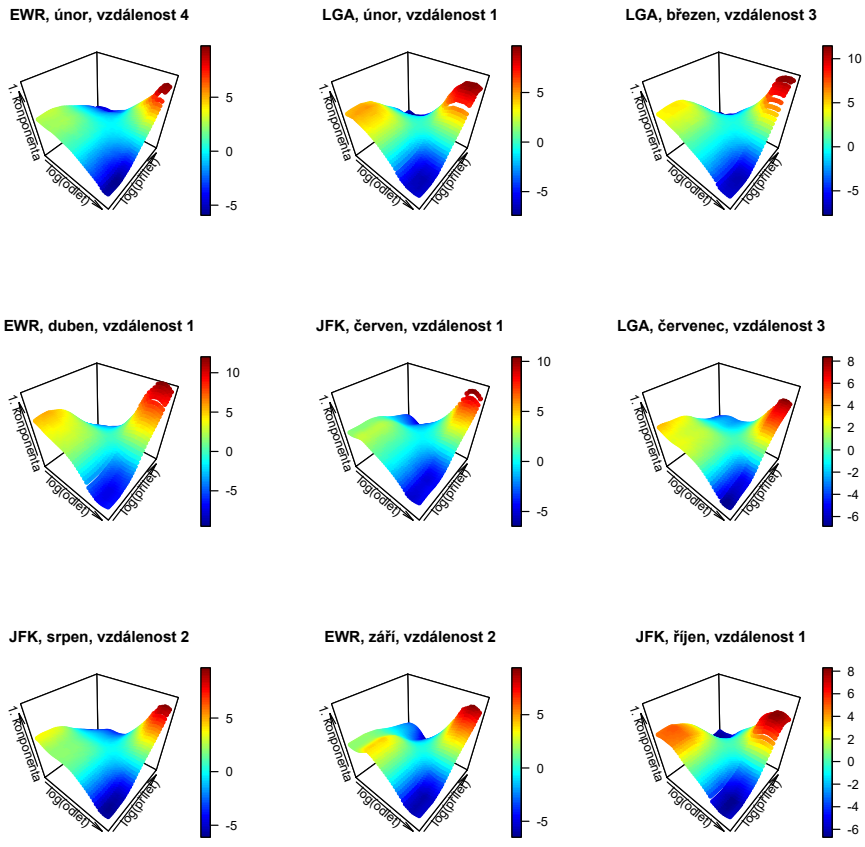


Obrázek 11: Prvních deset singulárních hodnot rozkladů interakčních částí hustot pro vybrané skupiny letů.

Podíváme-li se na singulární hodnoty rozkladů hustot pro jednotlivé skupiny, vidíme, že první singulární hodnota je vždy výrazně vyšší než ostatní hodnoty (obrázek 11), to znamená, že pomocí prvních komponent rozkladu můžeme získat dobrou aproximaci interakční části hustoty, tj.

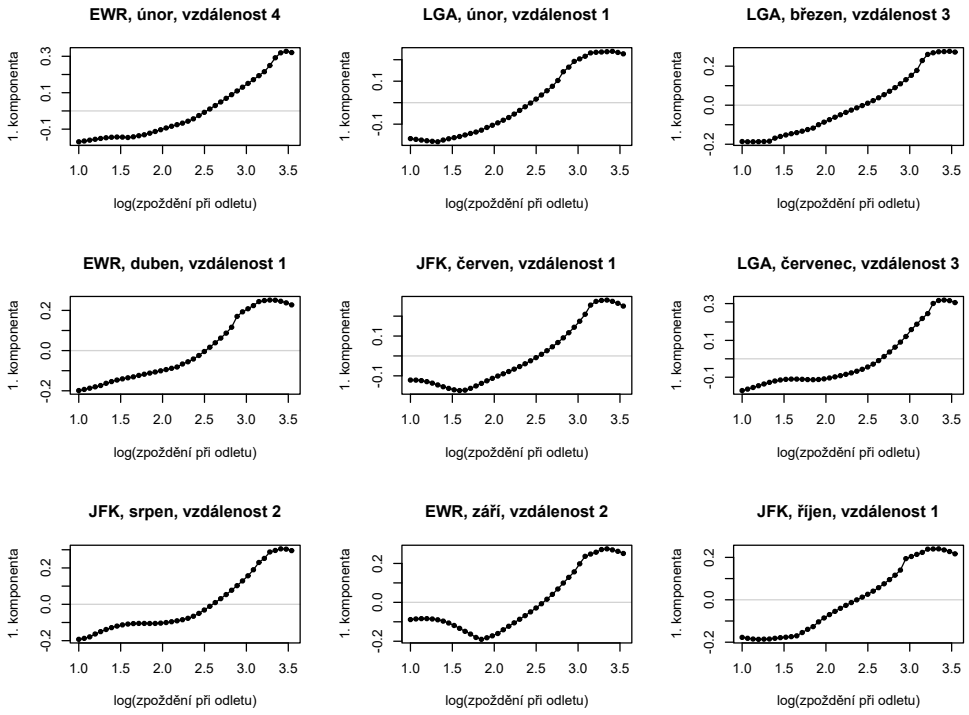
$$\mathrm{clr}(\widehat{f}_{\mathrm{int},jk}) \approx \gamma_1^{1/2} \varphi_{1,j} \psi_{1,k}, \quad j,k = 1,\ldots,40.$$

Na obrázku 12 vidíme, že aproximace interakčních částí hustot jsou skutečně velmi podobné původním interakčním hustotám (obrázek 10).

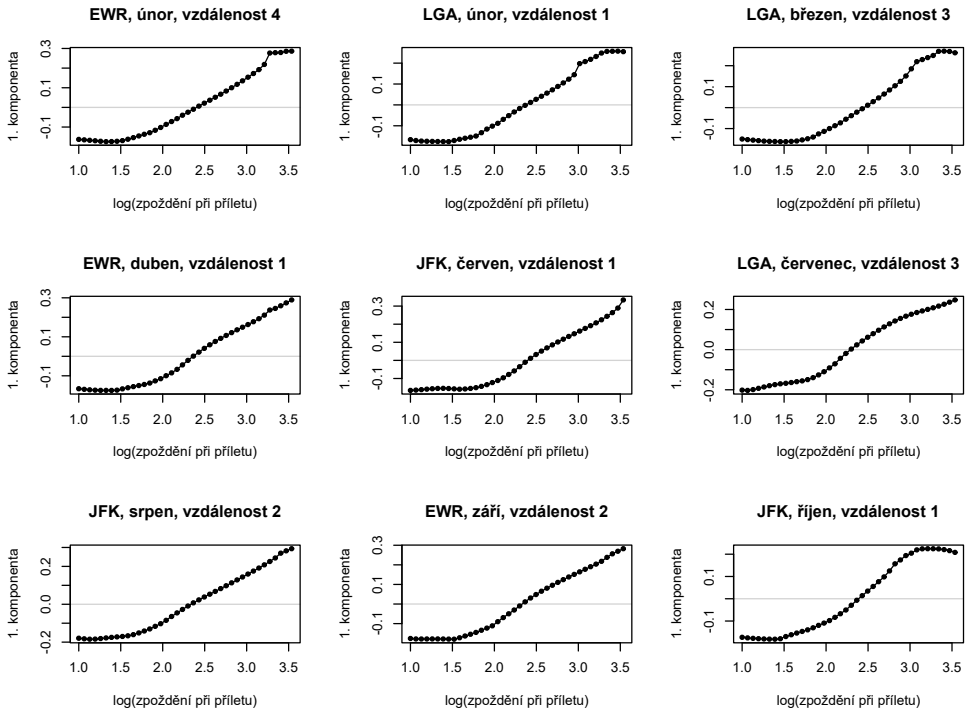


Obrázek 12: Aproximace interakčních částí hustot pomocí prvních komponent jejich singulárních rozkladů.

Můžeme si také zobrazit jednotlivé první komponenty (φ_1, ψ_1). Na obrázku 13 vidíme první komponenty odpovídající zlogaritmovanému zpoždění při odletu, na obrázku 14 pak ty odpovídající zlogaritmovanému zpoždění při příletu. Opět je možné vidět, že komponenty jsou si velmi podobné. U zpoždění při odletu si nejvýraznější odlišnosti můžeme všimnout u letů z letiště Newark v září do destinací ve 2. vzdálenostní kategorii, u zpoždění při příletu však takovou odlišnost nepozorujeme.



Obrázek 13: První komponenty singulárních rozkladů interakčních částí hustot odpovídající zlogaritmovanému zpoždění při odletu.



Obrázek 14: První komponenty singulárních rozkladů interakčních částí hustot odpovídající zlogaritmovanému zpoždění při příletu.

3.4. Možnosti analýzy nezávislých částí hustot

V této kapitole se zaměříme na nezávislé části hustot (jejich clr transformace), které jsou součtem clr (geometrických) marginálních hustot. Naším cílem bude porovnat FPCA popsanou v kapitole 2.3 a dvourozměrnou SFPCA z kapitoly 2.2, zaměříme se hlavně na porovnání skórů, které získáme těmito dvěma přístupy.

Začneme s FPCA, kterou aplikujeme na nezávislé části (diskretizovaných) hustot. Pro každou skupinu letů si vytvoříme řádkový vektor hodnot nezávislé části hustoty v bodech zvolené mřížky 40x40, tj.

$$\widetilde{\boldsymbol{f}}_s = \Big(\mathrm{clr}(\widehat{f}_{\mathrm{ind},1,1}), \mathrm{clr}(\widehat{f}_{\mathrm{ind},2,1}), \ldots, \mathrm{clr}(\widehat{f}_{\mathrm{ind},40,40})\Big)_s ,$$

kde $s = 1, 2 \ldots, 132$ označuje jednotlivé skupiny letů. Vektory $\widetilde{\boldsymbol{f}}_s$ pak budou tvořit řádky matice $\widetilde{\mathbf{F}}_{\text{ind}}$ typu 132 x 1600.

Hlavní komponenty budeme stejně jako v kapitole 2.3 značit $\phi_l, l = 1, 2 \ldots,$ a jejich hodnoty v bodech mřížky budeme zapisovat do sloupcových vektorů

$$\boldsymbol{\phi}_j = (\phi_{1,1}, \phi_{2,1}, \ldots, \phi_{40,40})'_j, \quad j = 1, 2, \ldots,$$

které pak budou tvořit sloupce matice $\boldsymbol{\Phi}$. Jsme však schopni najít pouze 132 hlavních komponent, neboť máme 132 skupin letů (pozorování), to ale vůbec nevadí, protože nás budou zajímat jen skóry odpovídající prvním dvěma komponentám.

Jelikož je u metody hlavních komponent zvykem pracovat s centrovanými daty, budeme centrovat i naše hustoty, tj. pro každý sloupec matice $\widetilde{\mathbf{F}}_{\text{ind}}$ spočítáme aritmetický průměr hodnot v něm, ten pak od těchto hodnot odečteme a výslednou matici označíme $\mathbf{F}_{\text{ind}}$.

Hlavní komponenty a skóry získáme pomocí singulárního rozkladu matice $\mathbf{F}_{\text{ind}}$, tj. $\mathbf{F}_{\text{ind}} = \mathbf{U}_{\text{ind}}\mathbf{D}_{\text{ind}}\mathbf{V}'_{\text{ind}}$, kde matice $\mathbf{U}_{\text{ind}}, \mathbf{V}_{\text{ind}}$ mají ortonormální sloupce, $\mathbf{U}_{\text{ind}}$ je řádu 132, $\mathbf{V}_{\text{ind}}$ je typu 1600 x 132 a $\mathbf{D}_{\text{ind}}$ je diagonální matice řádu 132, na jejíž hlavní diagonále se nacházejí singulární hodnoty. Ve sloupcích matice $\mathbf{V}_{\text{ind}}$ jsou hodnoty hlavních komponent v bodech mřížky, proto tedy platí $\boldsymbol{\Phi} = \mathbf{V}_{\text{ind}}$. Skóry jednotlivých nezávislých částí hustot nalezneme v řádcích matice $\mathbf{Z}_{\text{ind}} = \mathbf{U}_{\text{ind}}\mathbf{D}_{\text{ind}}$, tzn. řádek $\boldsymbol{z}_s^{BD} = (z_{s,1}^{BD}, \ldots, z_{s,132}^{BD})$, $s = 1, \ldots, 132$; v řádcích matice $\mathbf{U}_{\text{ind}}$ jsou pak normované skóry.

V kapitole 2 sice FPCA a dvourozměrná SFPCA působily docela odlišně, nicméně jejich praktické použití pro diskretizované hustoty je velmi podobné, postup se liší v podstatě jen tím, jak vypadají výše uvedené matice. Clr marginální hustoty (clr transformace geometrických marginálních hustot) si

pro každou skupinu letů zapíšeme do dvou řádkových vektorů

$$\widetilde{\boldsymbol{f}}_{s,X} = (\mathrm{clr}(\widehat{f}_{X,g,1}), \ldots, \mathrm{clr}(\widehat{f}_{X,g,40}))_s, \quad s = 1, \ldots, 132,$$
$$\widetilde{\boldsymbol{f}}_{s,Y} = (\mathrm{clr}(\widehat{f}_{Y,g,1}), \ldots, \mathrm{clr}(\widehat{f}_{Y,g,40}))_s, \quad s = 1, \ldots, 132,$$

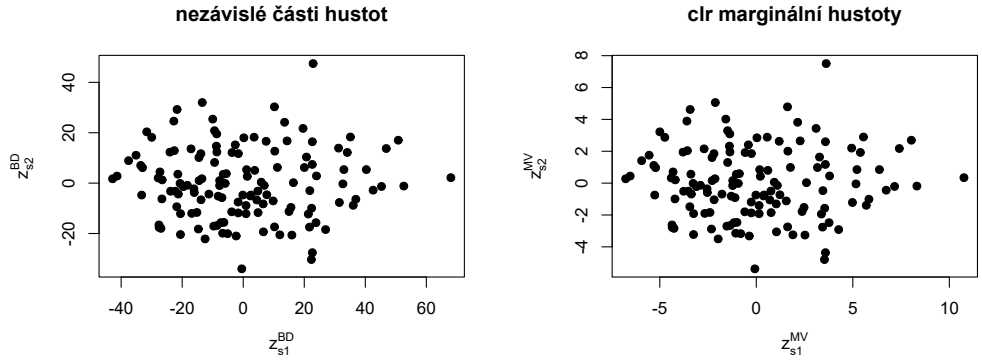
tyto vektory pak složíme do jednoho, tj. $(\widetilde{\boldsymbol{f}}_{s,X}, \widetilde{\boldsymbol{f}}_{s,Y})$, a ty pak budou tvořit řádky matice $\widetilde{\boldsymbol{F}}_g$ typu 132 x 80. Stejně jako u nezávislých částí hustot budeme centrovat, a získáme tak matici $\boldsymbol{F}_g$.

Provedeme singulární rozklad matice $\boldsymbol{F}_g$, tj. $\mathbf{F}_g = \mathbf{U}_g\mathbf{D}_g\mathbf{V}'_g$, zde je matice $\mathbf{U}_g$ typu 132 x 80, $\mathbf{V}_g$ je řádu 80 a diagonální matice $\mathbf{D}_g$ je řádu 80. I nyní se ve sloupcích matice $\mathbf{V}_g$ budou nacházet hodnoty hlavních komponent ve zvolených bodech, prvních 40 hodnot každého sloupce bude odpovídat komponentě pro první marginální hustotu (označené X – zpoždění při odletu) a zbylých 40 hodnot bude odpovídat komponentě druhé marginální hustoty (Y – zpoždění při příletu), tedy

$$\mathbf{v}_j = (\mathbf{v}_{X,j}, \mathbf{v}_{Y,j})' = (v_{1,j}, v_{2,j}, \ldots, v_{80,j})', \quad j = 1, \ldots, 80.$$

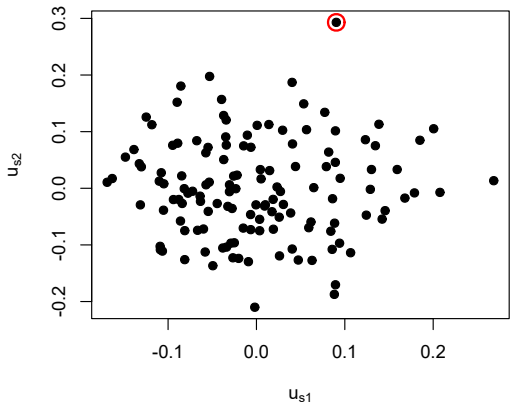
Vektory $\mathbf{v}_{X,j}, \mathbf{v}_{Y,j}$ pak budou odpovídat hodnotám složek komponent $\boldsymbol{\xi}_j$ popsaných v kapitole 2.2. Skóry pro jednotlivé dvojice marginálních hustot opět nalezneme v řádcích matice $\mathbf{Z}_g = \mathbf{U}_g\mathbf{D}_g$, tzn. $\boldsymbol{z}_s^{MV} = (z_{s,1}^{MV}, \ldots, z_{s,80}^{MV})$, $s = 1, \ldots, 132$, a normované skóry pak v matici $\mathbf{U}_g$.

Nyní si můžeme zobrazit skóry odpovídající prvním dvěma hlavním komponentám pro nezávislou část hustoty i pro clr marginální hustoty. Když se podíváme na obrázek 15, všimneme si shody mezi zobrazenými skóry, ty se skutečně liší pouze měřítkem.



Obrázek 15: Skóry odpovídající prvním dvěma hlavním komponentám nezávislých částí hustot (vlevo) a clr marginálních hustot (vpravo) – liší se pouze měřítkem.

Podíváme-li se na normované skóry, zjistíme, že jsou naprosto totožné (obrázek 16).



Obrázek 16: Normované skóry odpovídající prvním dvěma hlavním komponentám nezávislých částí hustot a clr marginálních hustot, červeně je označeno pozorování, které považujeme za odlehlé.

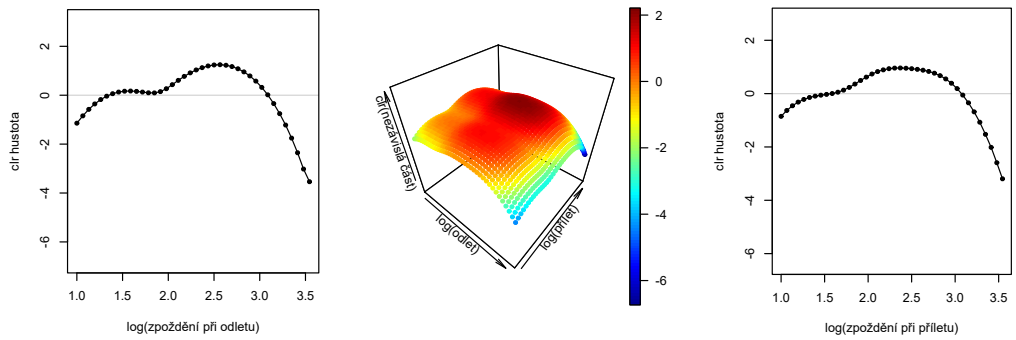
Tato shoda mezi skóry je daná tím, že nezávislá část hustoty (její clr transformace) je součtem clr marginálních hustot, hodnoty nezávislé části hustoty tedy získáme jako součty všech možných dvojic hodnot clr marginálních hustot, nezávislá část hustoty tak bude z pohledu metody hlavních

komponent obsahovat naprosto totožnou informaci jako jednotlivé marginální hustoty (matice $\mathbf{F}_{\text{ind}}$ a $\mathbf{F}_g$ se sice budou lišit svými rozměry, ale budou obsahovat v podstatě stejnou informaci). S využitím předchozího značení můžeme tento vztah zapsat maticově pro jednotlivé skupiny letů, tj.

$$\widetilde{\boldsymbol{f}}_s = \text{vec}\left[\text{diag}\left(\widetilde{\boldsymbol{f}}_{s,X}\right)\mathbf{J} + \mathbf{J}\,\text{diag}\left(\widetilde{\boldsymbol{f}}_{s,Y}\right)\right],$$

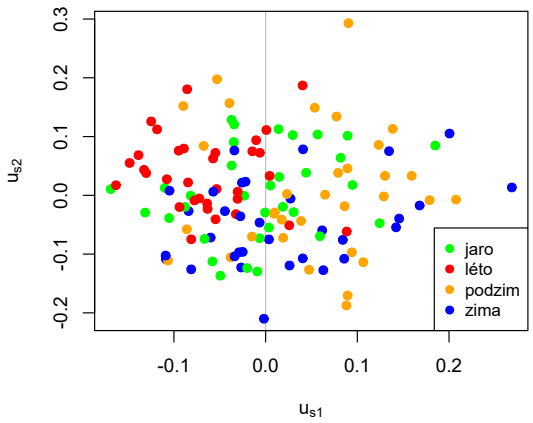
kde $s = 1, \ldots, 132$ značí skupinu letů a $\mathbf{J}$ je jedničková matice řádu 40, tedy matice obsahující samé jedničky.

Na obrázku 16 je také červeně označeno pozorování, které můžeme považovat za odlehlé, to znamená, že se jeho nezávislá část hustoty, respektive geometrické marginální hustoty, nějakým způsobem liší od hustot odpovídajícím ostatním skupinám letů. Zde se jedná o lety z letiště Newark v září do destinací ve 4. vzdálenostní kategorii; podíváme-li se na nezávislou část hustoty (její clr transformaci) pro tyto lety, vidíme, že je plošší než ty odpovídající ostatním skupinám. Totéž bude platit i pro clr marginální hustoty. Na obrázku 17 si tyto plošší hustoty můžeme prohlédnout a případně je porovnat s hustotami na obrázcích 7, 9, 8 v kapitole 3.3.



Obrázek 17: Clr marginální hustoty (vlevo a vpravo) a nezávislá část hustoty (uprostřed) odpovídající odlehlému pozorování.

Dále nás také zajímá interpretace jednotlivých komponent. Podíváme-li se, jakým skupinám letů odpovídají zobrazené skóry, může nás napadnout, že by první hlavní komponenta mohla souviset se sezónností. Na obrázku 18 vidíme, že hodnoty skórů (u_{s1}) odpovídajících skupinám letů v letních měsících (červen, červenec, srpen – červená) jsou převážně záporné, naopak hodnoty skórů pro skupiny letů v podzimních měsících (září, říjen, listopad – oranžová) jsou spíše kladné. Vyšší kladné hodnoty skórů pozorujeme také u některých skupin letů v zimních měsících (prosinec, leden, únor – modrá).

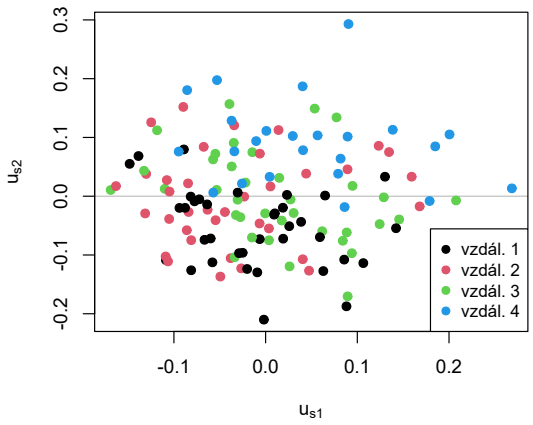


Obrázek 18: Normované skóry odpovídající prvním dvěma hlavním komponentám nezávislých částí hustot (respektive clr marginálních hustot) s barvami podle ročního období.

Pokud bychom si zobrazili ve vhodném pořadí, tj. podle hodnot skórů odpovídajících první komponentě, nezávislé části hustot a geometrické marginální hustoty (pracujeme sice s clr hustotami, ale ty můžeme snadno transformovat pomocí inverzní clr transformace a zobrazit), viděli bychom, že lety ve skupinách, kterým odpovídají nižší hodnoty skórů, mají tendenci k mírně menšímu zpoždění při odletu než ty, jimž odpovídají vyšší hodnoty skórů. Pro zpoždění při příletu je to pak právě naopak. Vzhledem k tomu, že většině skupin letů v letních měsících odpovídají malé hodnoty skórů, mohli

bychom usoudit, že zpoždění letů, jimž odpovídají malé hodnoty skórů, mohou být spojena s velkým leteckým provozem a případně se změnami počasí. Silný letecký provoz je příčinou čekání na povolení k odletu, a to způsobuje zpoždění při odletu, navíc také letadla musí letět tak, aby si vzájemně nekřížila dráhu letu, což může ovlivnit délku letu, a tím se může zvýšit zpoždění při příletu. Zpoždění se během letu může zvýšit také vlivem změn počasí, například v létě jsou lety ovlivňovány výskytem bouřek. Vyšší hodnoty skórů jsou spojeny většinou s lety v podzimních a zimních měsících, kdy letecký provoz bývá menší, a nemá tak velký vliv na zpoždění. V těchto obdobích jsou však často zhoršené povětrnostní podmínky, tj. špatná viditelnost, sníh, a lety tak mohou nabírat velká zpoždění již při odletu (počasí neumožňuje včasný odlet letadla), a pokud pak během letu nenastanou další komplikace, zpoždění během letu se výrazně nezvýší.

Podobně se můžeme zamyslet nad interpretací druhé hlavní komponenty, ta by podle obrázku 19 mohla souvislost se vzdálenostními kategoriemi.



Obrázek 19: Normované skóry odpovídající prvním dvěma hlavním komponentám nezávislých částí hustot (respektive clr marginálních hustot) s barvami podle vzdálenostních kategorií.

Vidíme, že hodnoty skórů (u_{s2}) odpovídající skupinám letů do destinací

v 1. vzdálenostní kategorii (černá) jsou většinou záporné, kdežto u skupin letů do destinací ve 4. vzdálenostní kategorii (modrá) převažují kladné hodnoty skórů. Když bychom si opět ve vhodném pořadí zobrazili nezávislé části hustot a geometrické marginální hustoty (nebo jejich clr transformace), mohli bychom si všimnout, že s rostoucími hodnotami skórů odpovídajících druhé hlavní komponentě jsou zobrazené hustoty plošší (méně špičaté). Tento efekt také odpovídá tomu, co jsme viděli u odlehlého pozorování popsaného výše – vyšší kladná hodnota skóru odpovídajícího druhé hlavní komponentě, ploché clr hustoty, skupina letů do destinací ve 4. vzdálenostní kategorii. Platí, že čím jsou hustoty plošší, tím je rozdělení zpoždění rovnoměrnější. Kromě rozdílů ve špičatosti bychom mohli pozorovat, že s rostoucími hodnotami skórů (tedy s rostoucími vzdálenostmi letů) mírně roste i tendence k většímu zpoždění. To může například souviset s tím, že dálkové lety mohou být komplikovanější, co se týče jejich celkové organizace.

Ještě by nás mohlo napadnou, zda by bylo možné pomocí skórů rozlišit jednotlivá letiště, ze kterých byly lety vypraveny. Kdybychom si však skóry obarvili podle jednotlivých letišť, viděli bychom, že takovéto rozlišení letišť není možné, proto zde obrázek ani neukazujeme. Při porovnání s barvami podle vzdálenostních kategorií by se projevilo to, že z letiště LaGuardia nebyly vypraveny žádné lety do destinací ve 4. vzdálenostní kategorii, což jsme však byli schopni vyčíst již přímo z dat.

3.5. Analýza interakčních částí hustot

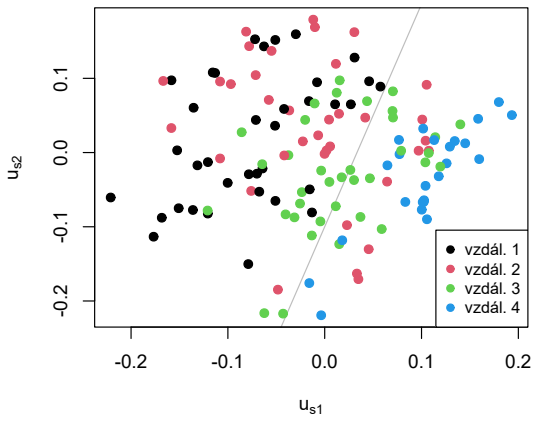
V poslední kapitole této práce se podíváme na dvourozměrnou FPCA pro interakční části hustot (jejich clr transformace).

Budeme postupovat úplně stejně jako v případě nezávislých částí hustot. Pro každou skupinu letů si vytvoříme řádkový vektor hodnot diskretizované

interakční části hustoty v bodech zvolené mřížky a tyto vektory budou tvořit řádky matice $\widetilde{\mathbf{F}}_{\text{int}}$. Po centrování pak dostaneme matici $\mathbf{F}_{\text{int}}$ a provedeme její singulární rozklad, tj. $\mathbf{F}_{\text{int}} = \mathbf{U}_{\text{int}}\mathbf{D}_{\text{int}}\mathbf{V}'_{\text{int}}$.

I v tomto případě si zobrazíme (normované) skóry a zamyslíme se nad interpretací prvních dvou hlavních komponent. Opět se zaměříme na rozlišení letišť, vzdálenostních kategorií a sezónnosti.

První hlavní komponenta by mohla souviset se vzdálenostními kategoriemi. Na obrázku 20 vidíme, že od sebe dokážeme oddělit skóry odpovídající skupinám letů do destinací v 1. a 4. vzdálenostní kategorii – hodnoty skórů (u_{s1}) odpovídajících letům do destinací v 1. vzdálenostní kategorii jsou většinou záporné, kdežto ty pro 4. vzdálenostní kategorii jsou až na výjimky kladné.

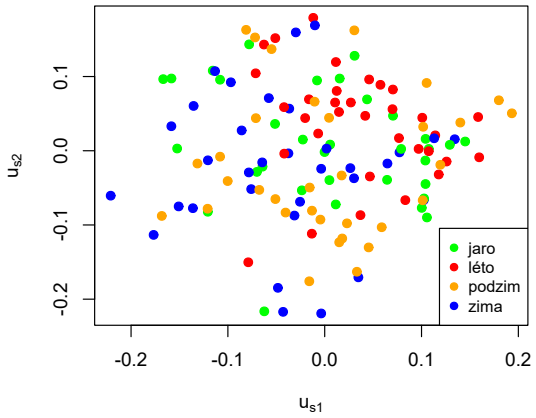


Obrázek 20: Normované skóry odpovídající prvním dvěma hlavním komponentám interakčních částí hustot s barvami podle vzdálenostních kategorií.

Stejně jako u nezávislých částí hustot bychom i zde mohli u clr transformovaných interakčních částí hustot pozorovat jejich zplošťování se zvyšujícími se hodnotami skórů odpovídajících první hlavní komponentě. To znamená, že u kratších letů je závislost mezi zpožděními větší než u dlouhých letů,

u kterých je více prostoru ke změně zpoždění během letu.

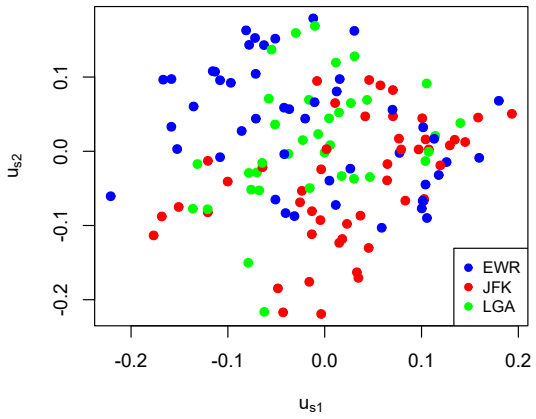
U sezónnosti bude situace trochu komplikovanější, neboť zde nevidíme (obrázek 21) žádné specifické vlastnosti u hodnot skórů pro podzimní a zimní lety. U skórů pro letní lety vidíme, že jejich hodnoty jsou převážně kladné nebo záporné blízké nule, a to jak u první, tak u druhé komponenty. A u skórů pro jarní lety si všimneme, že skóry odpovídající druhé hlavní komponentě nenabývají velkých záporných hodnot (až na jednu výjimku). Právě proto bychom mohli říct, že by druhá hlavní komponenta mohla částečně souviset se sezónností. Kdybychom si v tomto případě zobrazili interakční části hustot (jejich clr transformace), nepozorovali bychom žádné jejich výrazné změny v souvislosti s měnícími se hodnotami skórů.



Obrázek 21: Normované skóry odpovídající prvním dvěma hlavním komponentám interakčních částí hustot s barvami podle ročních období.

Stejně jako u nezávislých částí hustot ani nyní se nám nepodaří rozlišit jednotlivá letiště pomocí skórů odpovídajících prvním dvěma hlavním komponentám (obrázek 22). Mohli bychom sice říct, že hodnoty skórů pro letiště LaGuardia vzhledem k první komponentě jsou převážně v intervalu kolem nuly. Skóry pro letiště Newark nenabývají velkých záporných hodnot vzhle-

dem k druhé komponentě a skóry pro letiště JFK nenabývají velkých kladných hodnot vzhledem k této komponentě. Žádné větší odlišení letišť však není možné.



Obrázek 22: Normované skóry odpovídající prvním dvěma hlavním komponentám interakčních částí hustot s barvami podle letišť.

I když se nám úplně nepodařilo rozlišit skupiny letů pomocí skórů podle letišť, vzdálenostních kategorií a sezónnosti, můžeme na základě hodnot skórů obecně říct, že vztah (závislost) mezi zpožděním při odletu a při příletu se u jednotlivých skupin letů nějakým způsobem liší. Platí také, že čím jsou skóry dále od nuly (počátku souřadnicového systému), tím více se struktura závislosti mezi zpožděními u daných skupin letů liší od průměrné (typické) závislostní struktury.

Závěr

V této diplomové práci jsem se věnovala hustotám rozdělení pravděpodobnosti, a to hlavně těm dvourozměrným. To, že jsem s nimi pracovala jako s prvky Bayesova prostoru, mi umožnilo provést jejich ortogonální rozklad na nezávislou a interakční část, a každé z nich jsem se potom mohla věnovat zvlášť. Všechny teoretické postupy jsem se snažila ilustrovat na datovém souboru týkajícím se letů z newyorských letišť.

Při analýze nezávislých částí hustot, které jsou složeny z geometrických marginálních hustot, jsem se zaměřila hlavně na funkcionální metodu hlavních komponent a její modifikace vhodné pro analýzu hustot. Snažila jsem se zjistit, zda existuje nějaký vztah mezi skóry nezávislých částí hustot a skóry geometrických marginálních hustot, respektive mezi skóry jejich clr transformací. V diskrétním případě se podařilo ukázat, že vztah mezi skóry existuje, a potvrzují to i praktické výsledky. Nicméně je potřeba se tomuto vztahu ještě více věnovat, a tak se zde nabízí příležitost k dalšímu výzkumu.

U interakčních částí hustot jsem se zaměřila na jejich singulární rozklad, který dosud nebyl nijak teoreticky popsán, a tak jsem zpočátku musela vycházet pouze z několika intuitivních úvah, které se mi nakonec podařilo teoreticky popsat, i zde však jistě zůstává prostor pro další zkoumání tohoto rozkladu.

Tato práce mi umožnila vyzkoušet si nové způsoby analýzy dat a mohla jsem se také věnovat teoretickým úvahám. Myslím si, že se mi podařilo obohatit teorii Bayesových prostorů o nové poznatky a možnosti analýzy hustot. A i když mi občas zabralo mnoho času některé věci pochopit, zpracovávání tohoto tématu mě bavilo a našla jsem zde také několik dílčích témat, kterým bych se mohla věnovat v rámci svého dalšího výzkumu.

Literatura

[1] Filzmoser, P., Hron, K., Menafoglio, A.: *Logratio Approach to Distributional Modeling.* Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas–Agnan, Cham: Springer International Publishing (2021) 451–470.

[2] Genest, C., Hron, K., Nešlehová, J. G.: *Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula–based representation.* Journal of Multivariate Analysis 198 (2023) 105228.

[3] Gervini, D.: *The functional singular value decomposition for bivariate stochastic processes.* (2012) arXiv:1211.7336v1.

[4] Hron, K., Machalová, J., Menafoglio, A.: *Bivariate densities in Bayes spaces: orthogonal decomposition and spline representation.* Statistical Papers 64 (2023) 1629–1667.

[5] Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: *Simplical principal component analysis for density functions in Bayes spaces.* Computational Statistics and Data Analysis 94 (2016) 330–350.

[6] Talská, R., Menafoglio, A., Hron, K., Egozcue, J. J., Palarea–Albaladejo, J.: *Weighting the domain of probability densities in functional data analysis.* Stat 9.1 (2020) e283.

[7] Zhou, L., Pan, H.: *Principal Component Analysis of Two–Dimensional Functional Data.* Journal of Computational and Graphical Statistics 23.3 (2014) 779–801.