Hĺbka funkcionálnych dát

STANISLAV NAGY

s.nagy@volny.cz
KPMS UK, Praha

ABSTRAKT

Dôležitým neparametrickým nástrojom štatistickej analýzy mnohorozmerných dát je štatistická hĺbka. Na funkcionálne a špeciálne nekonečnerozmerné funkcionálne dáta sa však doposiaľ koncept hĺbky nepodarilo úspešne zovšeobecniť. V prípade konečnerozmerných funkcionálnych dát sa ponúka možnosť využiť izomorfizmus takéhoto priestoru s euklidovským priestorom. Iným prístupom k problematike hĺbky nekonečnerozmerných dát je koncept pásových hĺbok. V príspevku ilustrujeme rôzne prístupy na jednoduchých príkladoch.

ŠTATISTICKÁ HĹBKOVÁ FUNKCIA

Nech $\mathcal{F}$ je trieda pravdepodobnostných rozdelení na priestore $\mathbb{R}^d$. ***Štatistická hĺbková funkcia*** na $\mathbb{R}^d$ ako je definovaná v [1] je obmedzené a nezáporné zobrazenie $D(.;.) : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$, splňujúce podmienky:

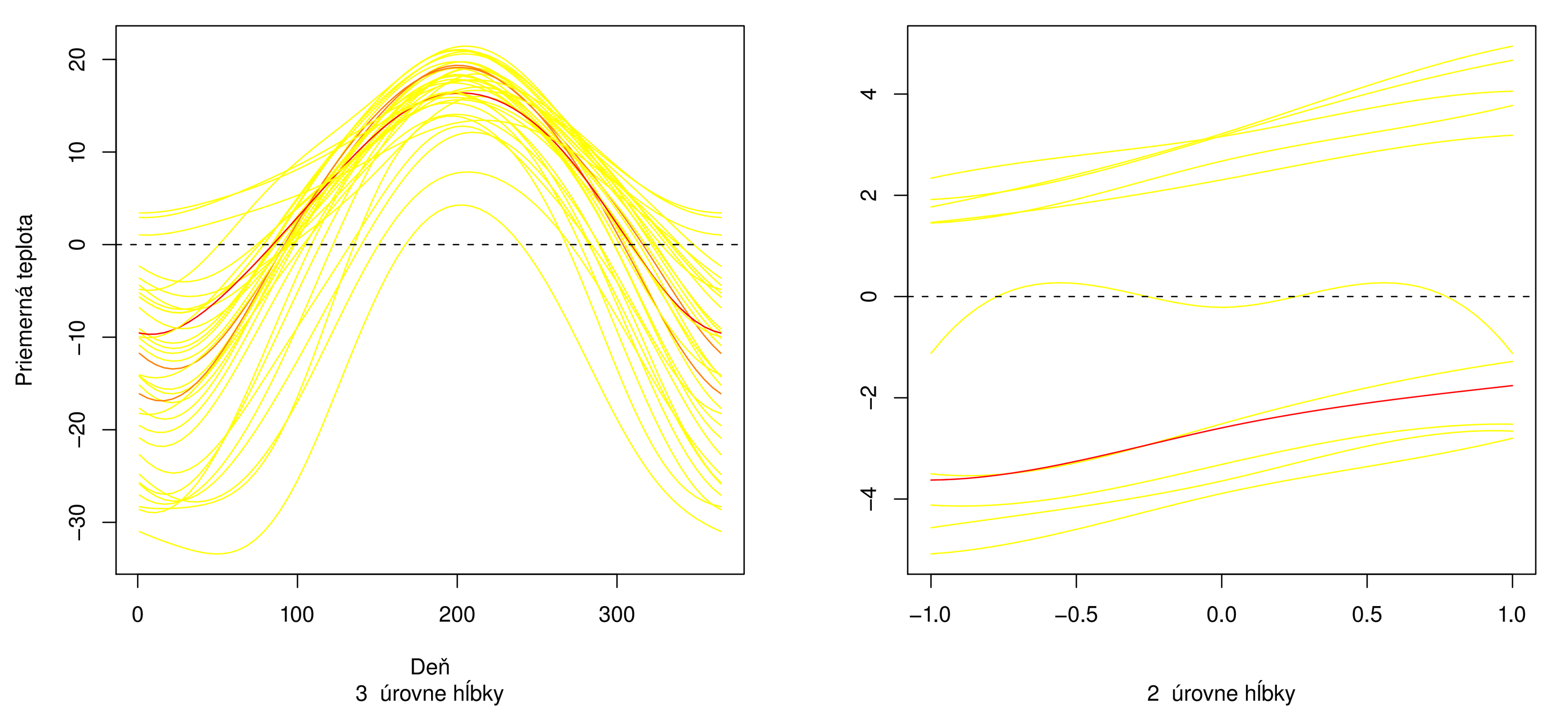
1. invariancia voči afinnej transformácii rozdelenia
2. maximalita v strede symetrie pre symetrické rozdelenia
3. relatívna monotónia vzhľadom k najhlbšiemu bodu
4. nulovosť limity v každom smere

Nech I je kompaktný interval. V prípade, že priestor uvažovaných funkcií $\mathcal{L} \subset C(I)$ je d-rozmerný s L^2-ortonormálnou bázou $\{\varphi_i(t)\}_{i=1}^d$, platí, že pre každú funkciu $f(t) = \sum_{i=1}^d c_i \varphi_i(t) \in \mathcal{L}$ je funkcionál

$$D_{\mathcal{L}}(f) := D\left((c_1, \ldots, c_d)^T\right)$$

štatistickou hĺbkovou funkciou na priestore $\mathcal{L}$.

Ilustrujme na reálnych dátach ročného priebehu teploty v Kanade a na simulovaných dátach taktúto indukovanú Tukeyho hĺbku. Funkcie náhodného výberu sú s rastúcou hĺbkou vykresľované tmavšou farbou.

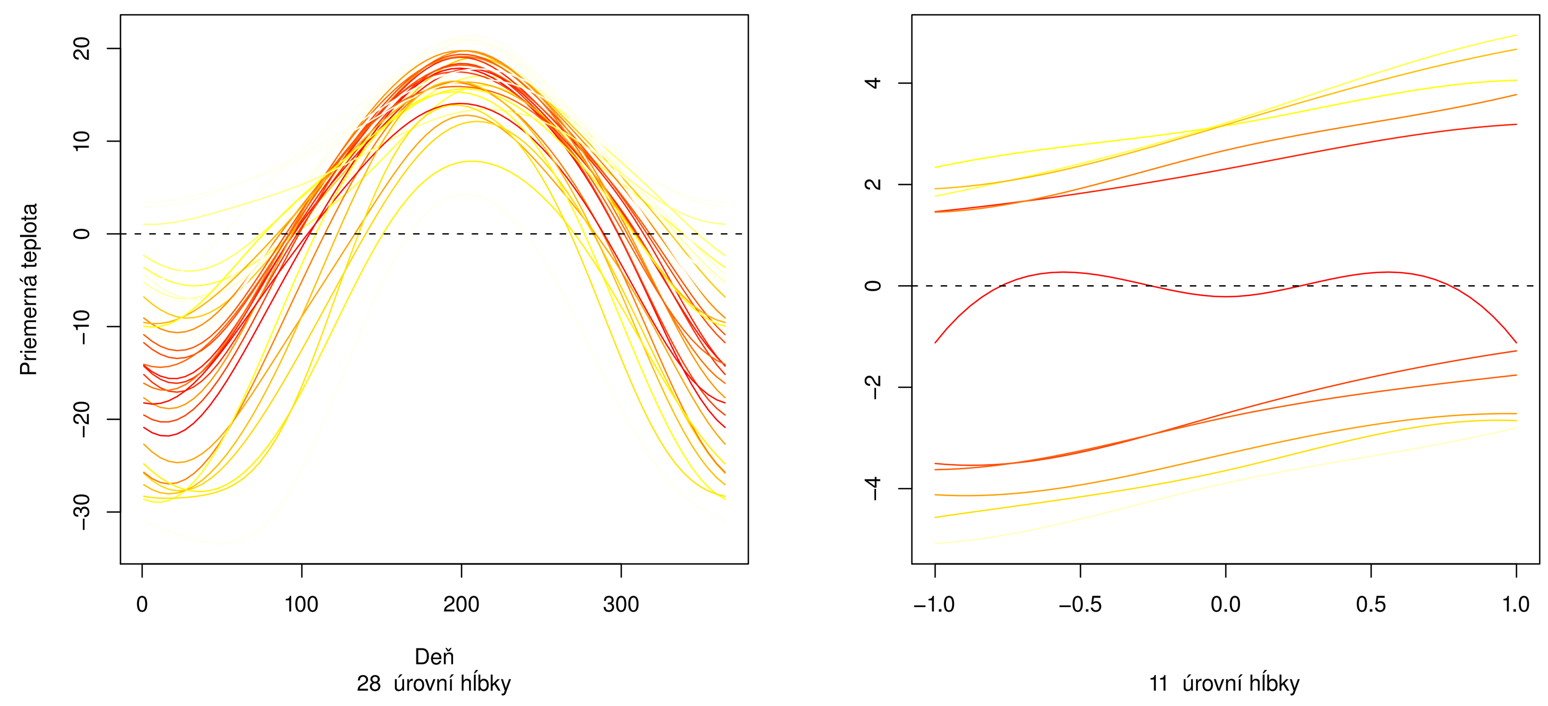


PÁSOVÁ HĹBKA

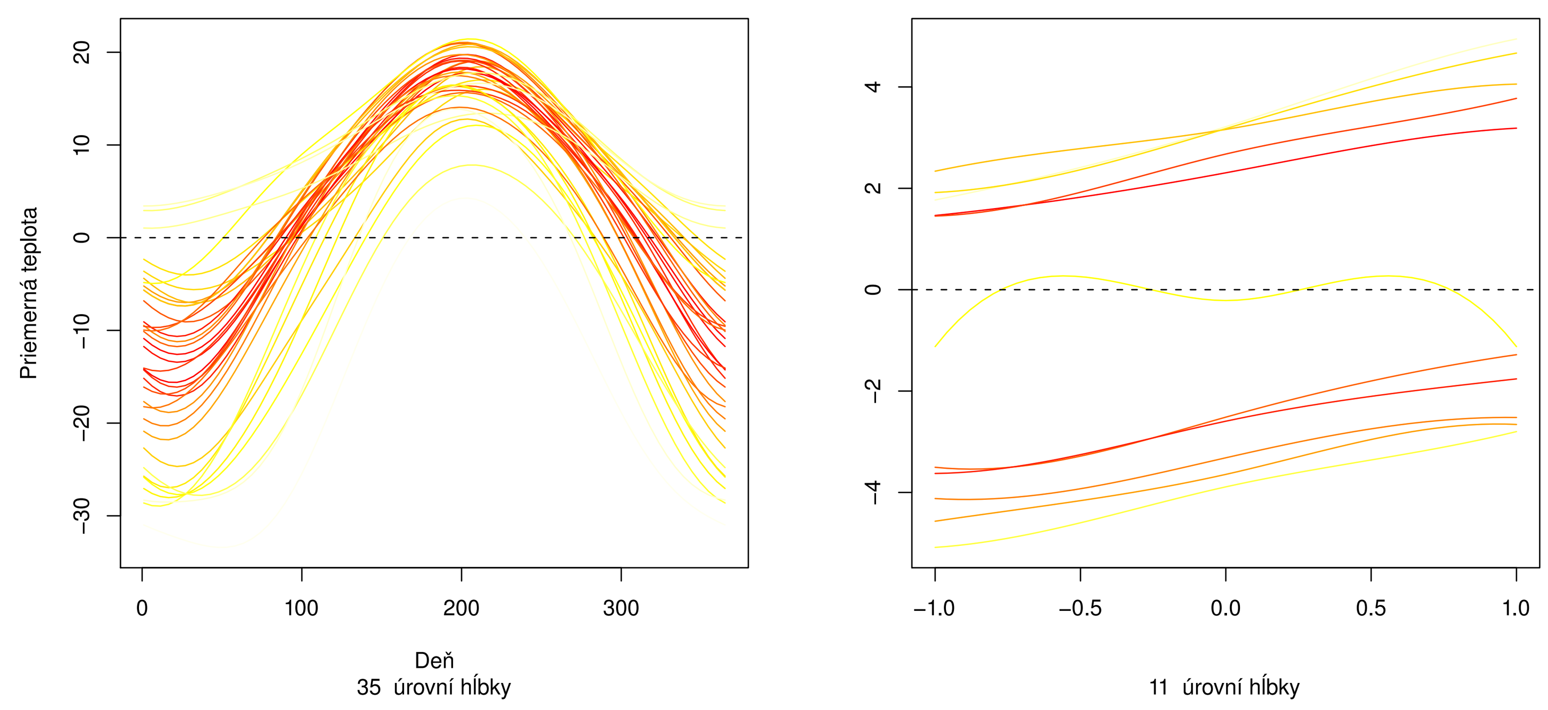
Iným prístupom k určovaniu hĺbky funkcionálnych dát je použitie rôznych verzií ***pásovej hĺbky*** definovanej v [2] alebo [3]. Táto je založená na princípe vyhodnocovania toho, nakoľko graf funkcie $f \in C(I)$, ktorej hĺbku zisťujeme, leží v ***páse*** tvorenom grafmi $2 \leq j \leq J$ náhodných funkcií z náhodného výberu z rozdelenia $\mathcal{F}$ o rozsahu $n > J$, kde J je predom volená konštanta, často $J = 3$. Pás tvorený grafmi funkcií $\{f_1, \ldots, f_j\} \subset C(I)$ je pritom definovaný ako

$$\left\{(x,y) \in \mathbb{R}^2 : (x \in I) \,\&\, \left(\min_{i=1,\ldots,j} f_i(x) \leq y \leq \max_{i=1,\ldots,j} f_i(x)\right)\right\}.$$

Voľbou indikátoru za mieru náležania dostávame štandardnú pásovú hĺbku, ktorej aplikácia na dáta je ilustrovaná na nasledujúcom obrázku, voľbou miery náležania ako Lebesgueovej miery takých bodov nezávisle premennej, že graf f leží v páse dostávame ***generalizovanú pásovú hĺbku***. Vlastnosti takýchto pásových hĺbok sú študované v [3].



Mediánová, teda najhlbšia funkcia, by však mala nielen ležať „uprostred" funkcií náhodného výberu, mala by mať aj podobný tvar ako ostatné funkcie. Za predpokladu hladkosti funkcií by teda zovšeobecnením mohla byť ***K-pásová hĺbka***, ktorú pre pevné $K \in \mathbb{N}$ definujeme ako vážený geometrický priemer generalizovaných pásových hĺbok funkcie $\frac{\partial^k f}{\partial x^k}$ voči deriváciám náhodnému výberu $\left\{\frac{\partial^k f_i}{\partial x^k}\right\}_{i=1}^n$ pre $0 \leq k \leq K$. Ukážka aplikácie takejto hĺbky na dáta je na poslednom obrázku.



ZÁVER

Koncept indukovaných hĺbok konečnerozmerných funkcií zrejme nie je dobrým riešením, z dôvodu prílišnej redukuje informácie obsiahnutej vo funkciách. Naopak pásová hĺbka dobre odlišuje **pozorovania odľahlé v polohe**, vhodná K-pásová aj **pozorovania odľahlé v tvare**. Vlastnosti takto definovaných funkcionálov však ešte neboli systematicky skúmané.

Literatúra.

[**1**] Zuo Y. a Serfling R. (2000). *General notions of statistical depth function.* Ann. Stat. **28**, 461–482.

[**2**] López-Pintado S. a Romo J. (2004). *Depth-based inference for functional data.* Computational Statistics and Data Analysis **51**, 4957–4968.

[**3**] López-Pintado S. a Romo J. (2009). *On the concept of depth for functional data.* Journal of the American Statistical Association **104**, 718–734.