# KOMPOZIČNÍ REGRESE S FUNKCIONÁLNÍ ZÁVISLE PROMĚNNOU

R. Talská, A. Menafoglio, K. Hron, J. Machalová, E. Fišerová

Katedra matematické analýzy a aplikací matematiky,
Přírodovědecká fakulta, Univerzita Palackého v Olomouci

**11. - 16. 9. 2016**

# Obsah prezentace

- Motivace a cíl příspěvku
- Hustoty z pohledu Bayesových prostorů
- Funkcionální regresní model v $L^2$ prostoru
- Funkcionální regresní model v Bayesově prostoru
- Shrnutí a výhled

# Motivace

- (Funkcionální) regresní analýza je široce užívána pro modelování **lineárního vztahu** mezi (funkcionální) **vysvětlovanou proměnnou** a **množinou** (skalárních) **proměnných**.
- **Otázka**: Je relevantní použít funkcionální modely navržené v $L^2$ prostoru pro hustoty rozdělení pravděpodobností v roli vysvětlované proměnné?
- **Problém**: $L^2$ prostor nezachycuje geometrické vlastnosti hustot.
  **Řešení**: Reprezentace hustot pomocí Bayesových prostorů.
- **Cíl**: Vytvořit funkcionální regresní model s využitím metodiky Bayesových prostorů.

# Geometrie Bayesova prostoru

- Hustoty jsou borelovsky měřitelné funkce (nezáporné, obvykle s jednotkovým integrálem), které obsahují informaci o relativních příspěvcích borelovských podmnožin nosiče hustot na celkovou míru tohoto nosiče.
- Bayesovy prostory $\mathcal{B}^2$ obsahují ekvivalentní třídy hustot s nosičem na $I = [a, b]$ (jejichž druhá mocnina logaritmu je integrovatelná).
- $\mathcal{B}^2$ má strukturu **separabilního Hilbertova prostoru**.
- **Centrovaná log-podílová transformace** (clr) = izometrický izomorfismus mezi $\mathcal{B}^2$ a $L^2$. Pro $f \in \mathcal{B}^2(I)$ máme:

$$f_c(t) = \mathrm{clr}\,[f(t)] = \ln f(t) - \frac{1}{\eta}\int_I \ln f(t)\,dt, \quad t \in I.$$

# Funkcionální regresní model v $L^2$ prostoru

Předpokládejme vysvětlovanou proměnnou $y(t)$ v $L^2$ a $p = r + 1$ nezávislých skalárních proměnných $x_0, \ldots, x_r$.

$\mathbf{y}(t) = (y_1(t), \ldots, y_N(t))$ ... vektor pozorovaných funkcí
$\mathbf{X} = \left[(x_{ij})\right]_{N \times p}$ ... matice plánu

Regrese *s absolutním členem*: první sloupec $\mathbf{X}$ je tvořen 1.

**Funkcionální lineární model** je ve tvaru

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{r} x_{ij}\beta_j(t) + \varepsilon_i(t), \quad i = 1, \ldots, N, \qquad (1)$$

kde $\beta_j(t), j = 0, \ldots, r$ jsou neznámé funkcionální regresní parametry v $L^2$ a $\varepsilon_i(t), i = 1, \ldots, N$ jsou funkcionální náhodné chyby v $L^2$.

Model (1) maticově: $\mathbf{y}(t) = \mathbf{X}\beta(t) + \varepsilon(t)$.

# Funkcionální regresní model v $L^2$ prostoru

**Odhady** $\hat{\beta}_j(t), j = 0, \ldots, r$ získáme minimalizací následujících kriterií:

1. Reziduální součet čtverců (RSČ):

$$\text{RSČ}(\beta) = \int_I [\mathbf{y}(t) - \mathbf{X}\beta(t)]' [\mathbf{y}(t) - \mathbf{X}\beta(t)] \, dt, \qquad (2)$$

2. Penalizované RSČ (PENRSČ):

$$\text{PENRSČ}(\beta) = \int_I [\mathbf{y}(t) - \mathbf{X}\beta(t)]' [\mathbf{y}(t) - \mathbf{X}\beta(t)] \, dt + \lambda \int_I [L\beta(s)]' [L\beta(s)] \, ds. \qquad (3)$$

# Funkcionální regresní model v $\mathcal{B}^2$ prostoru

Předpokládejme vysvětlovanou proměnnou $y(t)$ v $\mathcal{B}^2$ a $p = r + 1$ nezávislých skalárních proměnných $x_0, \ldots, x_r$.

$\mathbf{y}(t) = (y_1(t), \ldots, y_N(t))$ ... vektor pozorovaných hustot
$\mathbf{X} = \left[(x_{ij})\right]_{N \times p}$ ... matice plánu

Regrese *s absolutním členem*: první sloupec **X** je tvořen 1.

**Funkcionální lineární model** je ve tvaru

$$y_i(t) = \beta_0(t) \oplus \bigoplus_{j=1}^{r} \left[x_{ij} \odot \beta_j(t)\right] \oplus \varepsilon_i(t), \quad i = 1, \ldots, N, \qquad (4)$$

kde $\beta_j(t), j = 0, \ldots, r$ jsou neznámé funkcionální regresní parametry v $\mathcal{B}^2$ a $\varepsilon_i(t), i = 1, \ldots, N$ jsou funkcionální chyby nebo rezidua v $\mathcal{B}^2$.

# Funkcionální regresní model v $\mathcal{B}^2$ prostoru

**Odhady** $\hat{\beta}_j(t), j = 0, \ldots, r$ získáme použitím

$$\mathrm{RSČ}(\beta) = \sum_{i=1}^{N} \|\varepsilon_i(t)\|_{\mathcal{B}}^2 = \sum_{i=1}^{N} \left\| \bigoplus_{j=0}^{r} [x_{ij} \odot \beta_j(t)] \ominus y_i(t) \right\|_{\mathcal{B}}^2 \tag{5}$$

Kritérium (5) v $L^2$ prostoru:

$$\mathrm{RSČ}(\beta) = \sum_{i=1}^{N} \|\mathrm{clr}(\varepsilon_i(t))\|_2^2 = \sum_{i=1}^{N} \left\| \sum_{j=0}^{r} [x_{ij} \cdot \mathrm{clr}(\beta_j(t))] - \mathrm{clr}(y_i(t)) \right\|_2^2$$

s B-splajnovou reprezentací s omezující podmínkou na nulový integrál:

$$\mathrm{clr}(y_i(t)) = \sum_{k}^{K} c_{ik}\varphi_k(t), \qquad \int_I \sum_{k}^{K} c_{ik}\varphi_k(t)\, dt = 0, \tag{6}$$

$$\mathrm{clr}(\beta_j(t)) = \sum_{k}^{K} b_{jk}\varphi_k(t) \qquad \int_I \sum_{k}^{K} b_{jk}\varphi_k(t)\, dt = 0. \tag{7}$$

# B-splajnová reprezentace clr transformovaných hustot

**Úloha**: nalézt **splajn s nulových integrálem** na $I = [a, b]$: $\int_a^b s_k(x)\,\mathrm{d}x = 0$.

$\rightarrow$ podmínka je *přenesena* na podmínku pro B-splajnové koeficienty

**Věta (Nutná a postačující podmínka na b)**

*Pro splajn* $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$, $s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x)$,

*podmínka* $\int_a^b s_k(x)\,dx = 0$ *je splněna tehdy a jen tehdy, když*

$$\sum_{i=-k}^{g} b_i \left(\lambda_{i+k+1} - \lambda_i\right) = 0.$$

# Regrese s B-splajnovými koeficienty

**Vysvětlovaná** $y_i(t)$: splajny s koeficienty
$\boldsymbol{Y}_{(i)} = (Y_{i,1}, \ldots, Y_{i,g+k+1})'$ (řádky $\underline{\boldsymbol{Y}}_{N\times(g+k+1)}$)

**Mnohorozměrný lineární regresní model** je dán vztahem

$$\underline{\boldsymbol{Y}}_{(N\times(g+k+1))} = \boldsymbol{X}_{(N\times p)}\boldsymbol{B}_{(p\times(g+k+1))} + \underline{\varepsilon}_{(N\times(g+k+1))} \qquad (8)$$

Ekvivalentní zápis (8):

$$(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{g+k+1}) = \boldsymbol{X}(\beta_1, \beta_2, \ldots, \beta_{g+k+1}) + (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{g+k+1})$$

$\boldsymbol{Y}_{(i)} = (Y_{1,i}, \ldots, Y_{g+k+1,i})$, $i = 1, \ldots, N$ ... nezávislé mnohorozměrné proměnné se stejnou neznámou maticí $\boldsymbol{\Sigma}$

$\boldsymbol{X}$ ... matice plánu s plnou sloupcovou hodností

$\underline{\varepsilon}$ ... matice náhodných chyb

# Regrese v praxi

- Hustoty $y_i(t), t \in I = [a, b]$ jsou reprezentovány pomocí kompozičních vektorů $\mathbf{W}_i = (W_{i1}, \ldots, W_{iD})'$.
- Vyjádříme $\mathbf{W}_i$ pomocí $\boldsymbol{Z}_{(i)}$ v $\boldsymbol{R}^D$ aplikováním clr transformace:

$$Z_{ij} = \ln \frac{W_{ij}}{\sqrt[D]{\prod_{j=1}^{D} W_{ij}}}, j = 1, \ldots, D \rightarrow \underline{\boldsymbol{Z}}_{N \times D} = (\boldsymbol{Z}_{(1)}, \ldots, \boldsymbol{Z}_{(N)})'$$

- Clr hustoty získáme vyhlazením clr dat $\boldsymbol{Z}_{(i)}$:

$$s_i(x) = \sum_{j=1}^{g+k+1} Y_{ij} B_j^{k+1}(x),$$

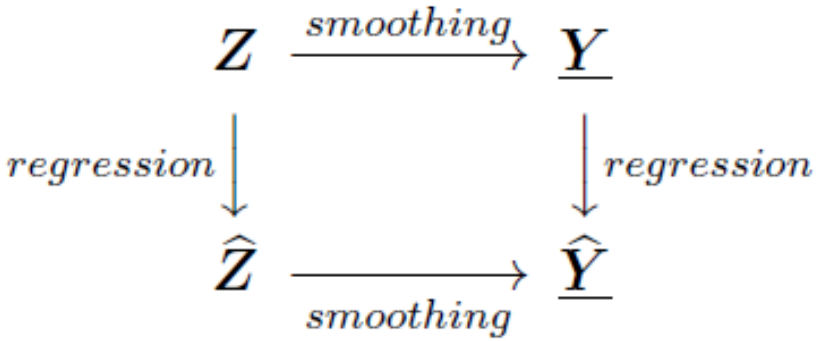
kde $\boldsymbol{Y}_{(i)} = (Y_{i,1}, \ldots, Y_{i,g+k+1})' = \boldsymbol{V}\boldsymbol{Z}_{(i)}, \quad i = 1, \ldots, N,$
nebo maticově: $\underline{\boldsymbol{Y}}_{N \times (g+k+1)} = \underline{\boldsymbol{Z}}_{N \times D} \boldsymbol{V}'_{D \times (g+k+1)}$

# Regrese v praxi

- Predikované hodnoty:

$$\underline{\widehat{\boldsymbol{Y}}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\underline{\boldsymbol{Y}}, \quad \underline{\widehat{\boldsymbol{Z}}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\underline{\boldsymbol{Z}}$$

- Lze ukázat, že pokud predikujeme $\underline{\boldsymbol{Y}}$ (B-splajnových koeficientů), stejný výsledek získáme vyhlazením predikovaných clr dat $\underline{\boldsymbol{Z}}$:




$\rightarrow$ kritérium PENRSČ je zahrnuto v RSČ kritériu díky použití vyhlazovacích splajnů.

# Shrnutí a výhled

- Použití přístupu Bayesových prostorů je nutnost v případě, že se hustota vyskytuje v roli vysvětlované proměnné ve funkcionálních lineárních regresních modelech.
- Odhad regresních funkcí (hustot) je možné převést na úlohu odhadu B-splajnových koeficientů clr transformovaných hustot.
  (Funkcionální model se převede na mnohorozměrný model s B-splajnovými koeficienty v roli vysvělované proměnné).
- Hladkost regresních odhadů (hustot) $\hat{\beta}_j(t)$ je automaticky kontrolována vstupní B-splajnovou reprezentací.
- Výhled: aplikace teoretických poznatků na reálná data.

# Reference

van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn, V., *Bayes Hilbert spaces*. Australian & New Zeland Journal of Statistics 56(2), pp. 171-194, 2014.

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, pp. 1175-1182, 2006.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P., *Simplicial principal component analysis for density functions in Bayes spaces*. Computational Statistics and Data Analysis 94, pp. 330-350, 2016.

Machalová, J., Hron, K., Monti, G.S., *Preprocessing of centred logratio transformed density functions using smoothing splines*. Journal of Applied Statistics, 2015, http://dx.doi.org/10.1080/02664763.2015.1103706.

Ramsay, J.O., Silverman, B.W., *Functional data analysis*. Springer, New York, 2005.