ROZHODOVACÍ PROCESY A KLASIFIKACE

Garant: **Igor Vajda**

Řešitelé: **Pavel Boček, Lucie Fajfrová, Tomáš Hobza, Tomáš Marek, Jan Nielsen, Petr Tichavský, Karel Vrbenský**

A: Pokročilé metody statistické analýzy dat

B: Optimalizace pomocí divergencí

Cíl: Nabídka vlastních statistických metod a optimalizačních postupů přijatých k publikaci ve významných zahraničních matematických časopisech.
Teoretický výzkum (publikace v časopisech, 2005-2009)
Algoritmizace (výzkumné zprávy, 2006-2008)
Programování (2005-2008)
Experimentování (publikace, výzkumné zprávy, 2006-2009)
Uživatelská finalizace programů (2009)

Statistické metody a postupy:

1. Odhadování stochastického modelu P datového zdroje
2. Testování jednoduché hypotézy $\mathcal{H} : P$
3. Testování složité hypotézy $\mathcal{H} : P \in \mathcal{P}$
4. Klasifikace do tříd daných modely $P_1, P_2, \ldots P_m$

Vše na základě empirie $\hat{P}$ získané z dat daného zdroje.

Odhadování: $P = \text{argmin}_{\mathcal{P}} D(P, \hat{P})$
Testování jednoduché: $D(P, \hat{P}) \geq D_{kritick\acute{e},\alpha}$
Testování složité: $\min_{\mathcal{P}} D(P, \hat{P}) \geq D_{kritick\acute{e},\alpha}$
Klasifikace: $\text{argmin}_i D(P_i, \hat{P})$
$D(P, \hat{P})$: divergence modelu P a reality $\hat{P}$.

$D(P, \hat{P}) \equiv$ Kullback $\Rightarrow$ MLE, LRT, GLRT
Rozmach divergencí - zdroj starostí

EOTEST (Empirická optimalizace testu): maximalizace citlivosti

$$\frac{1}{m} \sum_{i=1}^{m} \left(\frac{\pi(P_i) - \pi(P)}{d(P_i, P)} \right)^2$$

$P_1, \ldots, p_m$ alternativní hypotézy
$\pi(Q)$ relativní četnost zamítnutí simulovaných při zdroji Q
EOEST (Empirická optimalizace estimátoru)

Program A1: DIEFEM (Diskrétní eficientní metody)

P. Boček, T. Marek

- D. Morales, L. Pardo, I. Vajda (2003): Asymptotic laws for dispersity statistics in product multinomial models. *Journal of Multivariate Analysis* 85, 335-360
- D. Morales, L. Pardo, I. Vajda (2005): Efficient estimation in continuous models based on finitely quantized observations. *Comunications in Statistics - Theory and Methods*
- *Metrika* (2003) **57**, 1-27

A_{i1}	A_{i2}	$\dots$	A_{ir_i}

$\mathcal{X}$

$\hat{p}^i = (\hat{p}_{i1}, \dots, \hat{p}_{ir_i})$ relativní četnosti z n_i realizací
$q^i = (q_{i1}, \dots, q_{ir_i})$ hypotéza $\mathcal{H}_i$
testujeme $\mathcal{H} = (\mathcal{H}_1, \dots, \mathcal{H}_k)$

Statistika

$$T_k = \sum_{i=1}^{k} \frac{n_i}{n} D(\hat{p}_i, q_i)$$

$$n = \sum_{i=1}^{k} n_i, \qquad r = \sum_{i=1}^{k} r_i$$

$$\frac{nT_k}{c_D} \xrightarrow{\mathcal{L}} \chi^2_{r-k} \quad \text{když} \quad \min\{n_1, \ldots, n_k\} \to \infty$$

$$\frac{nT_k}{c_D\sqrt{r}} - \sqrt{r} \xrightarrow{\mathcal{L}} N(0,2) \quad \text{když} \quad \min\{r_1, \ldots, r_k\} \to \infty$$

Odhadujeme $\theta \in \Theta \subset R^k$ na základě $D(\mathbf{p}_\theta, \hat{\mathbf{p}}_n)$ kde

$$
\begin{aligned}
\mathbf{p}_\theta &= (p_{\theta_i} \equiv P_\theta(A_{ni}) : 1 \le i \le r_n) \\
\hat{\mathbf{p}}_n &= (p_{ni} \equiv P_n(A_{ni}) : 1 \le i \le r_n) \\
P_n &= \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \quad \frac{r_n}{n} \to 0 \quad \text{když } n \to \infty \\
A_{ni} &= [X_{n:i+1}, X_{n,i}) \\
\hat{\mathbf{p}}_n &= \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right), \quad r_n = n
\end{aligned}
$$

Program A2: ROME (Robustní metody)

T. Hobza

- F. Liese, I. Vajda (2003, 2004): A general asymptotic theory of M-estimators, I, II. *Mathematical Methods of Statistics* **12**, 454-477; **13**, 82-95
- T. Hobza, L. Pardo, I. Vajda (2004): Median estimators of parameters of logistic regression in models with disrete or continuous response. *Výzkumná zpráva ÚTIA* č. 2124, prosinec 2004.

$$Y_i \sim F_{\pi(\mathbf{x}_i'\beta_0)}(y), \qquad 1 \ll i \ll n,$$

$$F_\pi(y), \pi \in (0,1) \quad \text{distribuční funkce na} \quad R$$

$$\begin{array}{ll} \mathbf{x}_i & \text{známé regresory v} \quad R^k \\ \beta_0 & \text{neznámý parametr v} \quad R^k \end{array}$$

Při $\pi(t) = e^t/(1+e^t)$ jde o model logistické regrese, jinak jde o zobecněný lineární nebo obecněji pseudolineární model. Jde o robustní M-odhady parametru β_0 a robustní testy hypotéz $\mathcal{H} : \beta_0$ resp. $\mathcal{H} : \beta_0 \in B_0 \subset R^k$

Program A3: ANASIG (Analýza vícerozměrných signálů)
P. Tichavský, J. Nielsen

- P. Tichavský, Z. Koldovský, and E. Oja, "Performance Analysis of the FastICA Algorithm and Cramér-Rao Bounds for Linear Independent Component Analysis, submitted to *IEEE Trans. on Signal Processing*, partially presented at ICASSP'2005 Philadelphia, submitted to IEEE SSP Workshop Bordeaux 2005
- Z. Koldovský and P. Tichavský, "Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound". Submitted to *IEEE Statistical Signal Processing Workshop*, Bordeaux 2005.

Program A.3.1. Analýza nezávislých komponent (ICA)
Aplikace: Odstraňování artefaktů v EEG a MEG datech.
Spolupráce: Doc. Krajča, fakultní nemocnice Bulovka,
Doc. Stančák, 3. fakultní nemocnice Univerzity Karlovy

Program A.3.2. Slepá separace konvolutorních směsí
Aplikace: zpracování řečového signálu, odšumování, coctail party problem.
Spolupráce: FEL ČVUT, Prof. Sovka,
TU Liberec, Prof. Nouza

Lineární model pro analýzu nezávislých komponent

Lineární ICA model: $\mathbf{X}=\mathbf{AS}$

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_d \end{bmatrix} = \begin{bmatrix} s_1(1) & \dots & s_1(N) \\ \vdots & & \vdots \\ s_d(1) & \dots & s_d(N) \end{bmatrix}$$

- $s_i(t)$ jsou nezávislé realizace náhodné veličiny s distribuční funkcí $F_i(x) = P(s_i(t) < x)$ a hustotou $f_i(x)$.
- $\mathbf{A}$ - neznámá regulární matice o rozmìrech $d \times d$
- $\mathbf{X}$ - měřená (pozorovaná) data

Separace konvolutorních směsí: $\mathbf{X}(t) = \mathbf{A}(t) \circ \mathbf{S}(t)$

Lineární model pro analýzu nezávislých komponent

Lineární ICA model: $\mathbf{X} = \mathbf{A}\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_d \end{bmatrix} = \begin{bmatrix} s_1(1) & \dots & s_1(N) \\ \vdots & & \vdots \\ s_d(1) & \dots & s_d(N) \end{bmatrix}$$

- $s_i(t)$ jsou nezávislé realizace náhodné veličiny s distribuční funkcí $F_i(x) = P(s_i(t) < x)$ a hustotou $f_i(x)$.
- $\mathbf{A}$ - neznámá regulární matice o rozmìrech $d \times d$
- $\mathbf{X}$ - měřená (pozorovaná) data

Separace konvolutorních směsí: $\mathbf{X}(t) = \mathbf{A}(t) \circ \mathbf{S}(t)$

Lineární model pro analýzu nezávislých komponent

Lineární ICA model: $\mathbf{X}=\mathbf{AS}$

$$\mathbf{S}=\begin{bmatrix}\mathbf{s}_1\\ \vdots\\ \mathbf{s}_d\end{bmatrix}=\begin{bmatrix}s_1(1) & \dots & s_1(N)\\ \vdots & & \vdots\\ s_d(1) & \dots & s_d(N)\end{bmatrix}$$

- $s_i(t)$ jsou nezávislé realizace náhodné veličiny s distribuční funkcí $F_i(x)=P(s_i(t)<x)$ a hustotou $f_i(x)$.
- **A** - neznámá regulární matice o rozmìrech $d \times d$
- **X** - měřená (pozorovaná) data

Separace konvolutorních směsí: $\mathbf{X}(t)=\mathbf{A}(t)\circ\mathbf{S}(t)$

Lineární model pro analýzu nezávislých komponent

Lineární ICA model: $\mathbf{X}=\mathbf{AS}$

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_d \end{bmatrix} = \begin{bmatrix} s_1(1) & \dots & s_1(N) \\ \vdots & & \vdots \\ s_d(1) & \dots & s_d(N) \end{bmatrix}$$

- $s_i(t)$ jsou nezávislé realizace náhodné veličiny s distribuční funkcí $F_i(x) = P(s_i(t) < x)$ a hustotou $f_i(x)$.
- **A** - neznámá regulární matice o rozmìrech $d \times d$
- **X** - měřená (pozorovaná) data

Separace konvolutorních směsí: $\mathbf{X}(t) = \mathbf{A}(t) \circ \mathbf{S}(t)$

Program B1: EXPROPO (Exponenciální procesy a pole)
L. Fajfrová

- D. Morales, L. Pardo, I. Vajda (2000): Renyi statistics in directed families of exponential experiments. *Statistics* **34**, 151-174.
- D. Morales, L. Pardo, I. Vajda (2004): Renyi statistics for testing composite hypotheses in general exponential models. *Statistics* **38**, 133-147.

$$
\begin{aligned}
f_\theta(x) &= \exp\{\theta' T(x) + C(\theta)\}, \quad T : \mathcal{X} \to R^k, \theta \in R^k \\
C(\theta) &= \ln \int e^{\theta' T(\mathbf{x})} d\mu(x) \\
R_\alpha(\theta, \hat\theta) &= \frac{C(\alpha\theta + (1-\alpha)\hat\theta) - \alpha\, C(\theta) - (1-\alpha) C(\hat\theta)}{\alpha(1-\alpha)}
\end{aligned}
$$

Renyiho divergence řádu $\alpha \in R - \{0, 1\}$
Exponenciální sekvence $\mathbf{x} = (x_1, \ldots, x_n) \in R^n$
Lévyho náhodné procesy a náhodná pole
$\mathbf{x} = (x_t : t \in [0, \infty)), \quad$ nebo $\quad t \in [0, \infty)^2, x_t \in R^k$

$(\delta, \Sigma, \nu)-$ charakteristický triplet

$$C(\theta) = \delta'\theta + \tfrac{1}{2}\,\theta'\Sigma\theta + \int_{R^k}[e^{x'\theta} - 1 - \tau(x)'\theta]d\nu(x)$$

$$\tau(x) = \left(\tfrac{x_1}{1+x_1^2}, \ldots, \tfrac{x_k}{1+x_k^2}\right)$$

Odhadování parametru θ_0

Rozhodování: testy hypotéz, identifikace modelů

EODIT pro Rényiho statistiky $R_\alpha(\theta, \hat{\theta})$

Program B2: MIDIA (Minimální divergenční adaptace)
K. Vrbenský, T. Marek

- I. Vajda, E.C. van der Meulen (2005): On minimum divergence adaptation of discrete bivariate distributions to given marginals. *IEEE Trans. on Information Theory* **51**, 313-320.

$$
\begin{aligned}
\tilde{P} &= \operatorname{argmin}_{P \in \mathcal{P}_{a,b}} D_\phi(P, Q) \\
\mathcal{P}_{a,b} &= \{\tilde{P} : a, b \,\text{marginály}\, \tilde{P}\} \\
D_\phi(P, Q) &= \sum_{i,j} Q_{ij} \phi\left(\frac{P_{ij}}{Q_{ij}}\right)
\end{aligned}
$$

Program B3: DIBARI (Divergence a Bayesovská rizika)

I. Vajda, T. Hobza