MUNI
SCI

Teorie informací a hodnocení statistických modelů

Bakalářská Práce

Tomáš Petit
505485@mail.muni.cz

Přírodovědecká fakulta, Masarykova Univerzita

27-06-2023

Obsah Práce

- Teorie Informace
  - Shannonova Entropie
  - Relativní Entropie
- Akaikeho informační kritétium
  - Metoda maximální věrohodnosti
  - Asymptotické vlastnosti věrohodnosti
  - Střední hodnota logaritmu věrohodnostní funkce
  - Vychýlení logaritmu věrohodnostní funkce
  - AIC
- Modelování pomocí AIC
  - Lineární a polynomiální regrese
  - Histogramy
  - Rovnost dvou diskrétních distribucí
  - Rovnost středních hodnot a rozptylů
  - Mallowsovo C_p
  - Analýza hlavních komponent

Motivace do problematiky

Test podílem věrohodnosti

- Statistický model s prostorem parametrů Θ
- Věrohodnostní funkce $L(\theta)$
- H_0 Nulová vs. H_1 Alternativní hypotéza
  - Parametr $\theta \in \Theta_0 \subseteq \Theta$
  - Parametr $\theta \in \Theta \setminus \Theta_0$

$$\lambda_{LR} = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} \right] \tag{1}$$

$$= -2[l(\theta_0) - l(\hat{\theta})], \tag{2}$$

kde

$$l(\hat{\theta}) = \ln[\sup_{\theta \in \Theta} L(\theta)] \tag{3}$$

$$l(\theta_0) = \text{ maximum za platnosti } H_0 \tag{4}$$

Kullback-Leiblerova divergence

Statistické modely dány pomocí $f(x)$, $g(x)$ spojité náhodné veličiny X.

$$\ln\left(\frac{f(x)}{g(x)}\right) \tag{5}$$

Pak

$$I(f;g) = E_{f(x)}\left[\ln\left(\frac{f(x)}{g(x)}\right)\right] = \int_{-\infty}^{\infty} f(x)\ln\left(\frac{f(x)}{g(x)}\right)dx \tag{6}$$

Divergence $\neq$ metrika !

"Pravda" vypadne

$$E_{f(x)}\left[\ln\left(\frac{f(x)}{g(x)}\right)\right] = E_{f(x)}[\ln(f(x))] - E_{f(x)}[\ln(g(x))] \tag{7}$$

Expected likelihood

$$E_{f(x)}[\ln(g(x))] = \int_{-\infty}^{\infty} f(x) \ln(g(x)) dx \tag{8}$$

$$= \sum_{\alpha=1}^{n} \hat{f}(x_\alpha) \ln(g(x_\alpha)) \tag{9}$$

$$= \frac{1}{n} \sum_{\alpha=1}^{n} \ln(g(x_\alpha)) \tag{10}$$

Tedy

$$n \int_{-\infty}^{\infty} f(x) \ln(g(x)) dx = \sum_{\alpha=1}^{n} \ln(g(x_\alpha)) \tag{11}$$

Informační kritéria

Takeuchiho informační kritérium

$$\mathrm{TIC} = -2\sum_{\alpha=1}^{n} \ln f(X_\alpha|\hat{\theta}) + 2\mathrm{tr}\{J(\theta)I(\theta)^{-1}\} \tag{12}$$

Akaikeho informační kritérium

$$\mathrm{AIC} = -2\sum_{\alpha=1}^{n} \ln f(X_\alpha|\hat{\theta}) + 2p \tag{13}$$

Aplikace

- Hodnocení statistických modelů
  - Lineární regresní modely (polynomiální, ANOVA etc.)
  - PCA
  - Histogramy
  - Ekvivalence množin kategoriálních dat
  - Časové řady
  - Testování rovnosti středních hodnot

Příklad 1

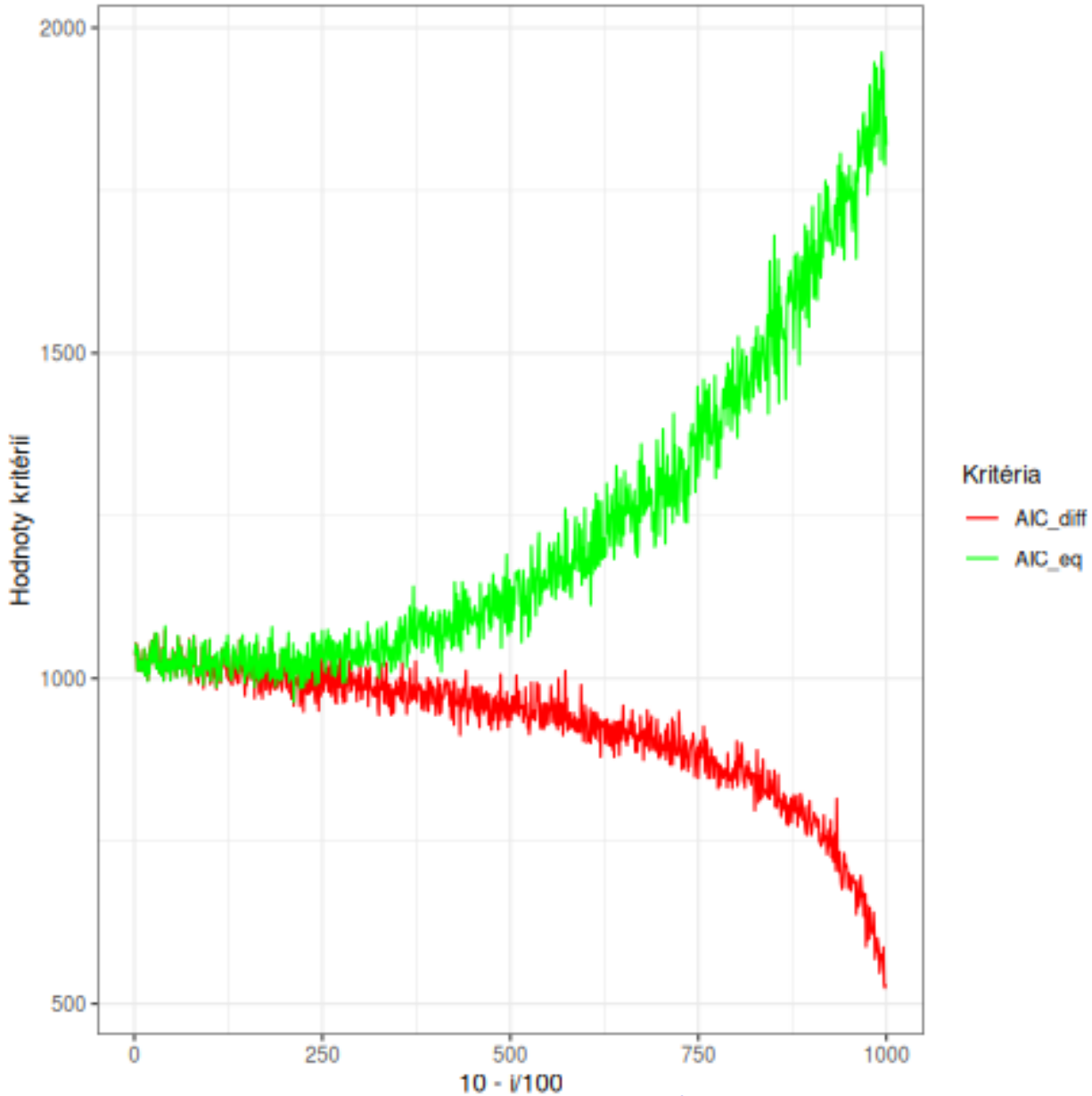
Volně dostupná data **Swiss**

- Model 1 - Fertility $\sim$ Catholic
- Model 2 - Fertility $\sim$ Catholic + Education + Infant.Mortality + Agriculture
- Model 3 - Full model

Model	**AIC**	R^2	R^2_{adj}
1	364.3479	0.215	0.1976
2	325.2408	0.6993	0.6707
3	326.0716	0.7067	0.671

Table: Srovnání AIC, R^2 a R^2_{adj}

Příklad 2


2000
1500
1000
500
Hodnoty kritérií
0
250
500
750
1000
10 - i/100
Kritéria
AIC_diff
AIC_eq

Shrnutí

- Základ v teorii maximální věrohodnosti
- Lze provést bodové odhady i intervaly spolehlivosti
- Lze použít jako základ pro statistiku

Děkuji za pozornost!

MASARYK
UNIVERSITY