M9750 Robustní a neparametrické statistické metody

Statistické modely

- $\mathbf{X} = (X_1, \dots, X_n)'$ je vektor pozorování.
- Parametrický model - $\mathbf{X}$ má sdruženou distribuční funkci $F(\mathbf{x}, \boldsymbol{\theta})$, F známe až na hodnotu neznámého parametru $\boldsymbol{\theta}$.
- Neparametrický model - nepředpokládá žádný specifický tvar rozdělení, neznámý parametr je nekonečněrozměrný.
- Semiparametrický model - model má parametrickou i neparametrickou složku s konečně i nekonečněrozměrným parametrem.

Robustní postupy

Robustní metody

- nejsou citlivé na porušení předpokladů modelu (normalita, odlehlá pozorování, apod.).
- zachovávají si eficienci, pokud předpoklady porušeny nejsou.

Odlehlá pozorování (outliers). Proč je hned neodstraníme z dat?

- Jsme líní. Neodhalíme je.
- Nejsme líní, ale neodhalíme je (typicky ve vyšších dimenzích).
- Máme málo pozorování, nechceme ztratit informaci obsaženou v datech.
- Odstraněním podhodnotíme odhad rozptylu.

Normalita

Poincaré (1912): *Všichni věří v normální rozdělení chyb. Experimentátoři proto, že je pokládají za matematický teorém, a matematikové proto, že je pokládají za experimentální fakt.*

Ověřování normality

Obecně mluvíme o metodách pro ověření shody teoretického a empirického rozdělení = vyhovují data naší představě (našemu modelu)? Dále se budeme zabývat ověřováním předpokladu normality dat.

Grafické metody:

- histogram
- jádrový odhad hustoty
- boxplot
- Q-Q plot
- P-P plot

Statistické testy sloužící k ověřování normality

- regresní testy: Shapirův - Wilkův test
- testy založené na empirických distribučních funkcích: Kolmogorovův - Smirnovův test
- testy dobré shody: Pearsonův χ^2 test

Matematický model:

- $X_1, \ldots, X_n$ je náhodný výběr z rozdělení s distribuční funkcí F.
- H_0: F je distribuční funkce (nějakého) normálního rozdělení.
- H_1: F není distribuční funkce normálního rozdělení.

Shapirův - Wilkův test

- je založen na porovnání dvou odhadů rozptylu σ^2 – výběrového rozptylu S^2 a nejlepšího odhadu získaného metodou nejmenších čtverců za předpokladu normality.
- Testová statistika:
$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$
- $(a_1, \ldots, a_n)^T = \frac{\mathbf{m}^T\mathbf{V}^{-1}}{(\mathbf{m}^T\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}}$,
- $\mathbf{m} = (m_1, \ldots, m_n)^T$, $m_i = \mathbb{E}Y_{(i)}$,
- $\mathbf{V} = (v_{ij})_{i,j=1}^{n}$, $v_{ij} = \mathrm{C}(Y_{(i)}, Y_{(j)})$,
- $Y_1, \ldots, Y_n$ je náhodný výběr z normovaného normálního rozdělení $\mathcal{N}(0,1)$.
- $W \leq 1$ a pro alternativu svědčí malé hodnoty W.
- Rozdělení W za H_0 je tabelováno.
- Test se hodí pro malé rozsahy výběru ($n \leq 50$).

Kolmogorovův - Smirnovův test

- Nulová hypotéza H_0: $F = F^*$, kde F^* je distribuční funkce $\mathcal{N}(\mu, \sigma^2)$ s μ a σ^2 známými.
- Definujme empirickou distribuční funkci

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x\}}.$$

- Testová statistika $D_n = \max_{x \in \mathbb{R}} \{|\widehat{F}_n(x) - F^*(x)|\}$.
- Za platnosti H_0 má $\sqrt{n} D_n$ asymptotické rozdělení stejné jako $\sup_{t \in [0,1]} |B(t)|$, kde $B(t)$ je Brownův most v $\mathcal{C}(0, 1)$.
- Test se dá použít jen, když μ a σ^2 jsou známé.

Lillieforsova modifikace Kolmogorovova - Smirnovova testu

- Nulová hypotéza H_0: F je distribuční funkce (nějakého) normálního rozdělení.
- Upravená testová statistika $D_n^* = \max_{i=1,\dots,n} \left\{ \left| \frac{i}{n} - \Phi\left(\frac{x_{(i)} - \widehat{\mu}}{\widehat{\sigma}}\right) \right| \right\}$.
- Asymptotické rozdělení $\sqrt{n} D_n^*$ je jiné než $\sqrt{n} D_n$.
- Proto se používají upravené kvantily (kritické hodnoty).

Pearsonův χ^2 test dobré shody

- Nulová hypotéza H_0: $F = F^*$, kde F^* je distribuční funkce $\mathcal{N}(\mu, \sigma^2)$ s μ a σ^2 známými.
- Definujme si intervaly $(b_{i-1}, b_i]$, $i = 1, \ldots, k$, kde $b_0 = -\infty, b_k = \infty$.
- Označme Y_i počet pozorování, které padnou do intervalu $(b_{i-1}, b_i]$ pro $i = 1, \ldots, k$.
- Určíme očekávaný počet pozorování, které by měly padnout do intervalu $(b_{i-1}, b_i]$:

$$np_i = np_i(\mu, \sigma^2) = nP(X_1 \in (b_{i-1}, b_i]) = n \int_{b_{i-1}}^{b_i} f(x, \mu, \sigma^2) dx.$$

- Testová statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i(\mu, \sigma^2))^2}{np_i(\mu, \sigma^2)}.$$

- χ^2 má za platnosti nulové hypotézy asymptoticky χ^2 rozdělení s $k - 1$ stupni volnosti.

Modifikace Pearsonova χ^2 testu dobré shody

- Nulová hypotéza H_0: F je distribuční funkce (nějakého) normálního rozdělení.
- Testová statistika

$$\widetilde{\chi}^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i(\widehat{\mu}, \widehat{\sigma}^2))^2}{np_i(\widehat{\mu}, \widehat{\sigma}^2)}.$$

- $\widetilde{\chi}^2$ má za platnosti nulové hypotézy asymptoticky χ^2 rozdělení s $k-3$ stupni volnosti.

Volba tříd:

- $p_i(\widehat{\mu}, \widehat{\sigma}^2) = \frac{1}{k}$.
- Heuristická pravidla pro počet tříd – $k \doteq 2n^{2/5}$, nebo $k \doteq 15(n/100)^{2/5}$.

Matematické nástroje robustnosti

- Model: $X_1, \ldots, X_n$ je náhodný výběr; X_i má rozdělení rozdělení pravděpodobnosti $P = P_{\boldsymbol{\theta}}$, kde $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$.
- Neznámý parametr $\boldsymbol{\theta}$ budeme chápat jako nějaký funkcionál daného rozdělení pravděpodobnosti, tj. $\boldsymbol{\theta} = T(P)$.
- Jeho přirozeným (empirickým) odhadem je funkcionál $T(P_n)$, kde P_n je empirické rozdělení pravděpodobnosti náhodných veličin $X_1, \ldots, X_n$.
- P_n je diskrétní rovnoměrné rozdělení na množině $\{X_1, \ldots, X_n\}$ a jeho příslušná distribuční funkce je empirická distrbuční funkce.

Matematické nástroje robustnosti

Definition

Nechť $\mathcal{P}$ je množina všech pravděpodobnostních rozdělení na $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Nechť $P, Q \in \mathcal{P}$ a $0 \leq t \leq 1$. Rozdělení pravděpodobnosti $P_t(Q) = (1 - t)P + tQ$ nazveme kontaminací P rozdělením Q v poměru t.

Definition

Nechť T je funkcionál na $\mathcal{P}$. Řekneme, že T je diferencovatelný (v Gâteauxově smyslu) podle P ve směru Q, jestliže existuje limita

$$T'_Q(P) = \lim_{t \to 0_+} \frac{T((1-t)P + tQ) - T(P)}{t}.$$

$T'_Q(P)$ se nazývá Gâteauxova derivace T podle P ve směru Q.

Taylorův rozvoj

- Taylorův rozvoj funkcionálu T: $T(Q) = T(P) + T'_Q(P) + o_P(1)$.
- Nyní zvolme $Q = P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, kde δ_x je Diracova míra v x, tj. $\delta_x(x) = 1$, $\delta_x(y) = 0$ jinak.
- $$\begin{aligned} T(P_n) - T(P) &= T'_{P_n}(P) + o_P(1) = \frac{1}{n}\sum_{i=1}^n T'_{\delta_{x_i}}(P) + o_P(1) \\ &= \frac{1}{n}\sum_{i=1}^n T'_{x_i}(P) + o_P(1). \end{aligned}$$
- $\frac{1}{n}\sum_{i=1}^n T'_{x_i}(P)$ je (přibližně) chyba odhadu $T(P_n)$.
- Člen $T'_{x_i}(P)$ je příspěvek X_i k této chybě.

Influenční funkce

Definition

Influenční funkcí funkcionálu T v bodě P nazveme derivaci T podle P ve směru $\delta_x, x \in \mathbb{R}$, tj.

$$IF(x, T, P) = T'_x(P) = \lim_{t \to 0_+} \frac{T((1-t)P + t\delta_x) - T(P)}{t}.$$

- Influenční funkce popisuje efekt kontaminace našeho rozdělení jedním bodem x na na odhad, který hledáme.
- Má-li být odhad robustní, influenční funkce by měla být omezená.

Kvantitativní charakteristiky robustnosti

- Globální citlivost funkcionálu T pro rozdělení pravděpodobnosti P: $\gamma^* = \sup_{x \in \mathbb{R}} |IF(x, T, P)|$.
- Lokální citlivost funkcionálu T pro rozdělení pravděpodobnosti P: $\lambda^* = \sup_{x,y \in \mathbb{R}, x \neq y} \left| \frac{IF(y,T,P) - IF(x,T,P)}{y-x} \right|$.
- Bod selhání ϵ^*.
- A další...

Bod selhání

- Označme $\mathbf{x}^0$ počáteční realizaci náhodného výběru a příslušný funkcionál $T_n(\mathbf{x}^0)$.
- Dále v $\mathbf{x}^0$ nahradíme m jeho složek co nejnepříznivějšími hodnotami (i $+-\infty$), označme jej $\mathbf{x}^{(m)}$ a příslušný funkcionál $T_n(\mathbf{x}^{(m)})$.
- Bod selhání odhadu T_n ve výběru $\mathbf{x}^0$ nazveme číslo $\epsilon_n^*(T_n, \mathbf{x}^0) = \frac{m^*(\mathbf{x}^0)}{n}$, kde $m^*(\mathbf{x}^0)$ je nejmenší celé číslo, pro které $\sup_{\mathbf{x}^{(m)}} |T_n(\mathbf{x}^{(m)}) - T_n(\mathbf{x}^0)| = \infty$.
- Pokud $\epsilon_n^*(T_n, \mathbf{x}^0)$ nezávisí na $\mathbf{x}^0$ definujeme bod selhání odhadu T_n jako $\epsilon^* = \lim_{n\to\infty} \epsilon_n^*(T_n, \mathbf{x}^0)$.

M-odhady jednorozměrného parametru θ

- M-odhad parametru θ je definován jako $\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \rho(X_i, \theta)$, kde ρ je nějaká, vhodně zvolená funkce.
- Existuje-li derivace $\psi(\cdot, \theta) = \frac{d\rho(\cdot,\theta)}{d\theta}$ a je spojitá, pak $\widehat{\theta}$ je (jedním z) řešení rovnice $\sum_{i=1}^{n} \psi(X_i, \theta) = 0$.
- Influenční funkce M-odhadu je

$$IF(x, T, P) = -\frac{\psi(x, T(P))}{\int_{\mathbb{R}} \frac{d\rho(y,\theta)}{d\theta}\Big|_{\theta=T(P)} dP(y)}.$$

- Má-li být M-odhad robustní, měl by mít omezenou funkci ψ.

M-odhady parametru polohy (posunutí)

- Model polohy: X_i mají distribuční funkci $F(x-\theta)$, kde F je symetrická kolem bodu θ.
- Ekvivalentně: $X_i = \theta + \epsilon_i$, kde ϵ_i mají distibuční funkci F, symetrickou kolem 0.
- M-odhad parametru polohy θ je definován jako $\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^n \rho(X_i - \theta)$, kde ρ je nějaká, vhodně zvolená funkce.
- Existuje-li derivace $\psi(y) = \rho'(y)$ a je spojitá, pak $\widehat{\theta}$ je (jedním z) řešení rovnice $\sum_{i=1}^n \psi(X_i - \theta) = 0$.
- Influenční funkce M-odhadu pro parametr polohy je

$$IF(x, T, P) = \frac{\psi(x - T(P))}{\int_{\mathbb{R}} \psi'(y) dP(y)}.$$

M-odhady parametru polohy - volba funkce ψ

- $\rho(x) = x^2$, $\psi(x) = x$ ... výběrový průměr - není robustní.
- $\rho(x) = |x|$, $\psi(x) = \mathrm{sign}(x)$ ... výběrový medián.
- $\psi_H(x) = \begin{cases} x, & |x| \leq k, \\ k \cdot \mathrm{sign}(x), & |x| > k \end{cases}$, kde $k > 0$ je pevně zvolená konstanta ... Huberův odhad.
- $\psi(x) = \frac{2x}{1+x}$ ... věrohodnostní funkce Cauchyho rozdělení.
- $\psi_T(x) = \begin{cases} x\left[1 - \left(\frac{x}{k}\right)^2\right], & |x| \leq k, \\ 0, & |x| > k \end{cases}$, kde $k > 0$ je pevně zvolená konstanta ... Tukeyho biweight.
- $\psi_A(x) = \begin{cases} \sin\left(\frac{x}{k}\right), & |x| \leq k\pi, \\ 0, & |x| > k\pi \end{cases}$, kde $k > 0$ je pevně zvolená konstanta ... Andrewsova sinusová funkce.

M-odhady parametru polohy - volba funkce ψ - pokrač.

- $\psi(x) = \begin{cases} |x| \cdot \mathrm{sign}(x), & |x| \leq a, \\ a \cdot \mathrm{sign}(x), & a < |x| \leq b, \\ a\frac{c-|x|}{c-b} \cdot \mathrm{sign}(x), & b < |x| \leq c, \\ 0, & |x| > c \end{cases}$, kde $0 < a < b < c$ jsou pevně zvolené konstanty ... Hampel.
- $\psi(x) = \begin{cases} x, & |x| \leq k, \\ 0, & |x| > k \end{cases}$, kde $k > 0$ je pevně zvolená konstanta ... skipped mean.
- $\psi(x) = \begin{cases} \mathrm{sign}(x), & |x| \leq k, \\ 0, & |x| > k \end{cases}$, kde $k > 0$ je pevně zvolená konstanta ... skipped median.

L-odhady parametru polohy

- Označme $X_{(1)} \leq \ldots \leq X_{(n)}$ pořádkové statistiky (uspořádaný náhodný výběr) pro náhodný výběr $X_1, \ldots, X_n$.
- L-odhad parametru polohy je definován jako $T_n = \sum_{i=1}^{n} c_i h(X_{(i)})$, kde h je nějaká funkce a c_i jsou vhodné konstanty.
- Pro odvozování teoretických vlastností se předpokládá, že

$$c_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} J(s) ds,$$

kde $J : [0, 1] \to \mathbb{R}$ je nějaká funkce.

L-odhady parametru polohy - příklady

- Výběrový průměr $\overline{X}$.
- Výběrový medián $\widetilde{X}$.
- Střed rozpětí $\frac{X_{(1)}+X_{(n)}}{2}$.
- α - useknutý průměr

$$\overline{X}_{\alpha}^{u} = \frac{1}{n - 2\lfloor n\alpha \rfloor} \sum_{i=\lfloor n\alpha \rfloor+1}^{n-\lfloor n\alpha \rfloor} X_{(i)}.$$

- α - winsorizovaný průměr

$$\overline{X}_{\alpha}^{W} = \frac{1}{n} \left\{ \lfloor n\alpha \rfloor X_{\lfloor n\alpha \rfloor} + \sum_{i=\lfloor n\alpha \rfloor+1}^{n-\lfloor n\alpha \rfloor} X_{(i)} + \lfloor n\alpha \rfloor X_{(n-\lfloor n\alpha \rfloor)} \right\}.$$

R-odhady parametru polohy

- Jsou inverzí pořadových testů o parametru polohy θ.
- Testujme hypotézu $H_0 : \theta = \theta_0$, kde θ_0 je známá hodnota, pomocí pořadového testu.
- Testová statistika je $S_n(\theta_0) = \sum_{i=1}^{n} \operatorname{sign}(X_i - \theta_0) a_n(R_i^+(\theta_0))$.
- $R_i^+(\theta_0)$ je pořadí $|X_i - \theta_0|$ mezi $|X_1 - \theta_0|, \ldots, |X_n - \theta_0|$ a $a_n : \{1, \ldots, n\} \to \mathbb{R}$ je nějaká funkce pořadí.
- Za platnosti nulové hypotézy platí $\mathbb{E} S_n(\theta_0) = 0$.
- To nás vede k tomu, hledat odhad θ jako řešení rovnice $S_n(\theta_0) = 0$.
- $S_n(\theta_0)$ je nerostoucí, schodovitá funkce - řešení nemusí existovat.
- R-odhad parametru θ tedy definujeme jako $T_n = \frac{1}{2}(T_n^+ + T_n^-)$, kde $T_n^+ = \inf\{t : S_n(t) > 0\}$ a $T_n^- = \sup\{t : S_n(t) < 0\}$.

R-odhady parametru polohy - příklady

- $a_n(1) = \ldots = a_n(n) = 1$, pak $T_n = \widetilde{X}$ je výběrový medián (inverze znaménkového testu).
- $a_n(i) = i$, pak $T_n = \operatorname{med}\left\{\frac{X_i+X_j}{2}, 1 \leq i \leq j \leq n\right\}$ je Hodgesův - Lehmannův odhad (inverze Wilcoxonova testu).
- $a_n(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$, pak T_n je inverzí van der Waerdenova testu; musí být počítán numericky.

Odhady parametru polohy ve více dimenzích

- Buď nyní $\boldsymbol{\theta}$ neznámý p-rozměrný parametr polohy. Označme jeho odhad $\widehat{\boldsymbol{\theta}}$.
- Uvažujme kvadratickou ztrátovou funkci $L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$.
- Riziko odhadu $\widehat{\boldsymbol{\theta}}$ definujeme jako $R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$.

Definition

Buďte $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$ dva odhady parametru $\boldsymbol{\theta}$. Řekneme, že $\widehat{\boldsymbol{\theta}}_1$ dominuje $\widehat{\boldsymbol{\theta}}_2$, jestliže $R(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}) \leq R(\widehat{\boldsymbol{\theta}}_2, \boldsymbol{\theta})$ pro všechny hodnoty $\boldsymbol{\theta}$ a existuje $\boldsymbol{\theta}_0$ tak, že $R(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_0) < R(\widehat{\boldsymbol{\theta}}_2, \boldsymbol{\theta}_0)$.

Definition

Odhad $\widehat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ je přípustný (admissible), jestliže neexistuje žádný jiný odhad parametru $\boldsymbol{\theta}$, který by jej dominoval.

Jamesův - Steinův odhad

- Nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je náhodný výběr z p-rozměrného normálního rozdělení $\mathcal{N}_p(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_p)$, kde $\sigma^2 > 0$ je známé.
- Je-li $p \geq 3$, pak výběrový průměr $\overline{\mathbf{X}}$ není přípustný odhad parametru $\boldsymbol{\theta}$.
- Dominuje jej mj. Jamesův - Steinův odhad $\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{n\|\overline{\mathbf{X}}\|^2}\right)\overline{\mathbf{X}}$.
- Tento odhad také není přípustný, dominuje jej mj. positive rule Jamesův - Steinův odhad $\widehat{\boldsymbol{\theta}}^{+}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{n\|\overline{\mathbf{X}}\|^2}\right)_{+}\overline{\mathbf{X}}$.