Selekcja zmiennych w wysokowymiarowym modelu liniowym

Konrad Furmańczyk

Wydział Zastosowań Informatyki i Matematyki SGGW

Zakopane 9.09.2016

Model liniowy

$Y = X\beta + \epsilon$, (ϵ_i) i.i.d., $E(\epsilon_1) = 0$, $Var(\epsilon_1) = \sigma^2$,
ϵ_1 subgaussowski ze stałą $\sigma > 0$ tzn. $E(exp(u\epsilon_1)) \leq exp(\sigma^2 u^2/2)$
dla dowolnego $u \in R$
X -deterministyczna macierz wymiaru n x p; $\beta \in R^p$;
$p = p(n) >> n$;

Będziemy rozpatrywać tylko modele rzadkie (skończenie wiele β_i rożnych od zera) oraz macierz X będzie spełniać pewne warunki regularności.

Least Absolute Shrinkage and Screening Operator (LASSO)

LASSO

$\widehat{\beta_L} = argmin_{\beta \in R^p}\left(\frac{1}{2n} \parallel Y - X\beta \parallel_2^2 + \lambda_n \parallel \beta \parallel_1\right)$

dla pewnego $\lambda_n > 0$

Progowe LASSO (TL)

$\widehat{\beta_{TL,j}} = \widehat{\beta_{L,j}} \mathbf{1}\left\{\left|\widehat{\beta_{L,j}}\right| \geq \delta_n\right\}$ dla $j = 1, ..., p$

dla pewnego progu $\delta_n > 0$;

Model rzadki

$I_0 = \{j : \beta_j \neq 0\}$,
$I_1 = \{j : \beta_j = 0\}$
$|I_0| = p_0 < n$
p_0 -stałe, niezależne od n

Dwustopniowa procedura selekcji TLSD

W pierwszym kroku (TL)
wybieramy zbiór zmiennych $S_1 = \left\{1 \leq j \leq p : \left|\widehat{\beta_{L,j}}\right| \geq \delta_n\right\}$

W drugim kroku stosujemy procedurę stepdown (SD) multitestowania

(h_0) $\quad$ H$_i$: $\beta_i = 0$ vs. H_i': $\beta_i \neq 0$ dla $i \in S_1$

Procedura stepdown

Określamy p-value dla testowania hipotezy H_i wobec H_i' jako $\pi_i = 2\left(1 - \Phi(|t_i|)\right)$ dla $i \in S_1$, gdzie Φ dystrybuanta rozkładu $N(0,1)$.

Statystyka testowa $t_i = \widehat{\beta_{ols,i}} / se\left(\widehat{\beta_{ols,i}}\right)$, gdzie

$$\widehat{\beta_{ols}} = \left(X_{S_1}' X_{S1}\right)^{-1} X_{S_1}' Y$$

$$se\left(\widehat{\beta_{ols,i}}\right) = \begin{cases} \sigma\sqrt{m_{i,i}} & \text{gdy } \sigma \text{ znane} \\ S\sqrt{m_{i,i}} & \text{gdy } \sigma \text{ nieznane} \end{cases}$$

S -pewny zgodny estymator σ oraz $\left(X_{S_1}' X_{S1}\right)^{-1} = (m_{i,j})_{i,j \in S_1}$

Procedura stepdown

Niech $s_1 = |S_1|$
Uporządkowanie $\pi_{(1)} \leq ... \leq \pi_{(s_1)}$ wyznacza "kolejnośc" hipotez zerowych $H_{(1)} \leq ... \leq H_{(s_1)}$

$\alpha_1 \leq ... \leq \alpha_{s1}$ -pewne stale (mogą zależyć od n)

Gdy $\pi_{(1)} > \alpha_1$ nie odrzucamy żadnej z hipotez H_i;
(h_1) *w.p.p. gdy* $\pi_{(1)} \leq \alpha_1, ..., \pi_{(r)} \leq \alpha_r$ to odrzucamy $H_{(1)}, ..., H_{(r)}$, gdzie r jest największą liczbą spełniającą (h_1).

Przykłady procedur stepdown

Przykłady procedur SD

a) *Holm* $\alpha_j = \frac{q_n}{s_1+1-j}$

b) *UHolm* $\alpha_j = \frac{([\gamma j]+1)q_n}{s_1+[\gamma j]+1-j}$ dla $0 \leq \gamma \leq 1$

c) *BH* $\alpha_j = \frac{jq_n}{s_1}$

d) *Bonferroni* $\alpha_j = \frac{q_n}{s_1}$ dla pewnego $q_n \to 0$

Warunki dla α_i dla procedur stepdown

(A_1) $\quad \alpha_{s1} \to 0$ przy $n \to \infty$

(A_2) $\quad (1/n)\log(1/\alpha_1) \to 0$ przy $n \to \infty$

Uwaga. Dla $q_n = 1/(n\log(n))$ zachodzą warunki $(A_1) - (A_2)$.

Warunki regularności

warunki regularności dla macierzy X

(B_1) $\| x_j \|_2 / \sqrt{n} = 1$ dla $j = 1, ..., p$, gdzie x_j -j-ta kolumna macierzy X

(B_2) *niech* $v_{I1} = \{v_i : i \in I_1\}$, $v_{I0} = \{v_i : i \in I_0\}$;
$C(I_0; 3) = \{v \in R^p : \| v_{I1} \|_1 \leq 3 \| v_{I0} \|_1\}$;
dla pewnego $\gamma > 0$ zachodzi $(1/n) v' X' X v \geq \gamma \| v \|_2^2$ dla każdego $v \in C(I_0; 3)$

Warunki regularności

warunki regularności modelu i estymacji

(B_3) $\min_{j \in I_0} |\beta_j| \geq C_1 \lambda_n$ dla pewnej stałej $C_1 > 0$

(B_4) $C_2 \lambda_n \leq \delta_n \leq \lambda_n (C_1 - 3/\gamma)$ dla pewnej stałej $C_2 > 0$

(B_5) $\lambda_n = C_\lambda \sqrt{\log(p)/n}$ dla $C_\lambda = 2\sigma$.

warunek zgodności estymatora S

(C) $S \to^P \sigma$ przy $n \to \infty$

Zgodność procedury selekcji

Twierdzenie o zgodności

Procedura selekcji TLSD spełniająca warunki:
$(A_1) - (A_2), (B_1) - (B_5)$ gdy σ jest znane oraz dodatkowo warunek (C) gdy σ jest nieznana, jest zgodną procedurą selekcji.

Zgodność procedury

Procedura selekcji jest zgodna gdy $P\left(\widehat{l} = l_0\right) \to 1$ przy $n \to \infty$,

gdzie $\widehat{l}$ oznacza liczbę wybranych istotnych zmiennych.

Fakty pomocnicze

Lemat 1

Niech zachodzą warunki $(B_2), (B_4), (B_5)$.
Wtedy dla dowolnego $r > 0$ z prawdopodobieństwem co najmniej $1 - 2\exp(-\frac{1}{2}(r-2)\log(p))$ zachodzi $|S_1| \leq p_0\left(1 + C_3/\gamma^2\right)$ dla pewnej stałej $C_3 > 0$.

Lemat 2

Niech zachodzą warunki $(B_2), (B_3), (B_5)$.
Wtedy dla dowolnego $r > 0$ z prawdopodobieństwem co najmniej $1 - 2\exp(-\frac{1}{2}(r-2)\log(p))$ zachodzi $I_0 \subset S_1$.

Szkic dowodu

Procedura selekcji jest zgodna gdy $P(V \geq 1) \to 0$ oraz $P(R \neq p_0) \to 0$ przy $n \to \infty$, gdzie V -fałszywie odrzucone zmienne, R -liczba odrzuconych zmiennych w wyniku zastosowania procedury TLSD. Korzystając z Lematu 2 wystarczy pokazać, że $P\left(\widetilde{V} \geq 1\right) \to 0$ oraz $P\left(\widetilde{R} \neq p_0\right) \to 0$, gdzie $\widetilde{V}$ -fałszywie odrzucone zmienne, $\widetilde{R}$ -liczba odrzuconych zmiennych w wyniku testowania $(h_0) - (h_1)$.

Szkic dowodu

Na mocy Lematu 1 wystarczy pokazać, że (*p. Furmanczyk*, 2016)

(i) $\mathrm{P}(\pi_i \leq \alpha_{s1}) \to 0$ dla $i \in I_1$

(ii) $\max_{j \in I_0} (1 - F_j(\alpha_1)) \in 0$ przy $n \to \infty$, gdzie F_j -dystrybuanta p-value π_j dla $j \in I_0$.

Warunki $(i) - (ii)$
są implikowane przez warunki $(A_1), (A_2), (C)$.

Symulacje

Macierz X stała dla wszystkich symulacji generowana z $N_p(0, Id)$

Symulowane modele

$(M1)$ $Y = \sum_{j=1}^{p_0} X_j + \epsilon$

$(M2)$ $Y = 0.3X_1 + X_2 + 0.4X_3 + 1.5X_4 + \epsilon$

$(M3)$ $Y = 0.3X_1 + 0.3X_2 + X_3 + X_4 + 0.4X_5 + 0.4X_6 + 1.5X_7 + 1.5X_8 + \epsilon,$

gdzie $\epsilon \sim N(0, 1)$

Symulacje

W symulacjach przyjęto $\lambda_n = \delta_n = \sqrt{\log(p)/n}$. Wykonano 1000 replikacji MC selekcji z modeli $(M1) - (M3)$
używając procedur TLSD(Holm, UHolm, Bonf., BH) oraz dla porównania TL, SCAD.

SCAD (Smoothly Clipped Absolute Deviation - Fan and Li (2001))

$$\widehat{\beta} = argmin_{\beta \in R^p} \parallel Y - X\beta \parallel_2^2 /n + \sum_{i=1}^{p} J_\lambda (|\beta_i|)$$

SCAD

Kara dla $a = 3.7$

$$J_\lambda(\theta) = \begin{cases} \lambda|\theta| & \text{dla } |\theta| \leq \lambda \\ -\left(\theta^2 - 2a\lambda|\theta| + \lambda^2\right)/\left(2(a-1)\right) & \text{dla } \lambda < |\theta| \leq a\lambda \\ (a+1)\lambda^2/2 & \text{dla } |\theta| > a\lambda \end{cases}$$

dla LASSO mamy karę

$$J_\lambda(\theta) = \lambda|\theta|$$

Wyniki symulacji dla M1

	p=500 p_0=5	p=500 p_0=10	p=500 p_0=20	p=1000 p_0=20	p=2000 p_0=5	p=2000 p_0=10	p=2000 p_0=20
Bonf	992	991	980	980	998	996	982
Holm	992	991	966	969	998	996	971
UHolm_0.01	992	991	966	969	998	996	971
UHolm_0.1	992	991	966	969	998	996	969
UHolm_0.5	992	991	966	969	998	996	969
UHolm_0.9	992	991	966	969	998	996	969
BH	992	991	966	969	998	996	969
SCAD	1000	1000	1000	1000	1000	1000	1000
TL	994	992	981	984	1000	1000	989

Tabela 1. Częstość wybranych prawdziwych modeli z 1000 symulacji dla $n = 200$.

Wyniki symulacji dla M2

	p=500 n=100	p=500 n=200	p=1000 n=100	p=1000 n=200	p=2000 n=100	p=2000 n=200
Bonf	1000	995	1000	997	1000	998
Holm	1000	995	1000	997	1000	998
UHolm_0.01	1000	995	1000	997	1000	998
UHolm_0.1	1000	995	1000	997	1000	998
UHolm_0.5	1000	995	1000	997	1000	998
UHolm_0.9	1000	995	1000	997	1000	998
BH	1000	995	1000	997	1000	998
SCAD	584	931	521	902	450	879
TL	40	342	20	235	18	148

Tabela 2. Częstość wybranych prawdziwych modeli z 1000 symulacji

Wyniki symulacji dla M3

	p=500 n=100	p=500 n=200	p=1000 n=100	p=1000 n=200	p=2000 n=100	p=2000 n=200
Bonf	995	999	994	997	992	997
Holm	995	999	994	997	992	997
UHolm_0.01	995	999	994	997	992	997
UHolm_0.1	995	999	994	997	992	997
UHolm_0.5	995	999	994	997	992	997
UHolm_0.9	995	999	994	997	992	997
BH	995	999	994	997	992	997
SCAD	154	880	95	821	37	778
TL	71	313	36	205	30	118

Tabela 3. Częstość wybranych prawdziwych modeli z 1000 symulacji

Literatura

- Benjamini, Y., Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Infer.* **82**, 163-170.
- Bunea, F., Wegkamp, M.H., Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *J. Statist. Plann. Infer.* **136**, 12, 4349-4364.
- Fan, J.and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.
- Friedman, J., Hastie, T., Simon, N. and Tibshirani, R. (2015), glmnet: Lasso and elastic-net regularized generalized linear models. R package version 2.0.
- Furmańczyk, K. (2014). Selection in parametric models via some stepdown procedures. *Appl. Math. (Warsaw)* **41**, 81-92.

Literatura

- Furmańczyk, K. (2015). On some stepdown procedures with application to consistent variable selection in linear regression. *Statistics* **49**, 614-628.
- Furmańczyk, K. (2016). Variable selection using stepdown procedures in high-dimensional linear models. *Appl. Math. (Warsaw)* DOI: 10.4064/am2286-6-2016
- Furmańczyk, K. (2016). Stepdown procedures with thresholded Lasso for high dimensional parametric model selection. preprint
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65-70.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.