Analiza dużych zbiorów danych raport nr 4

Dominik Mika

13 października 2021

Część teoretyczna

Zadanie 1

W tym zadaniu wygenerujemy macierz ortonormalną $X_{n\times p}$, gdzie $n = 500$ i $p = 500$. Następnie stworzymy model liniowy

$$Y = X\beta + \epsilon,$$

gdzie $\epsilon \sim N(0, I_{500\times 500})$, a wektor β wyraża się wzorami:

1. $\beta_1 = \cdots = \beta_k = 4$ i $\beta_{k+1}, \cdots, \beta_{500} = 0$,
2. $\beta_1 = \cdots = \beta_k = \sqrt{\frac{80}{k}}$ i $\beta_{k+1}, \cdots, \beta_{500} = 0$.

Dla k odpowiednio równego 5, 20 i 100. Zauważmy, że dla $k = 5$ obie postacie β dają nam ten sam wektor. Stąd będziemy rozważać 5 wersji tego wektora. W tej i dalszej części jako β_1 i β_2 będę oznaczał jako typy wektora β.

Dla takiego modelu wykonamy następujące obliczenia:

i)

Dla kazdego $i \in \{1, 2, \cdots, p\}$ wyliczymy teoretycznie obciążenie, wariancję i błąd średniokwadratowy dla estymatora $\hat{\beta}_i$ uzyskanego za pomocą regresji grzbietowej z parametrem wygładzającym γ.

Wiemy, że estymator β zadaje się następującym wzorem:

$$\hat{\beta} = (XX' + \gamma I)^{-1}X'Y = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

Oznaczmy $Z = X'\epsilon \sim N(0, \sigma^2 I)$.

MSE

W celu obliczenia MSE obliczmy najpierw

$$\mathbb{E}(\hat{\beta}_i - \beta_i)^2 = \mathbb{E}\left(\frac{1}{1+\gamma}\beta_i + \frac{1}{1+\gamma}Z_i - \beta_i\right)^2 = \frac{1}{(1+\gamma)^2}\mathbb{E}(-\beta_i\gamma + Z_i)^2 = \frac{\gamma^2\beta_i^2 + \sigma^2}{(1+\gamma)^2}.$$

Stąd

$$MSE = \mathbb{E}\|\hat{\beta} - \beta\|^2 = \frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2}.$$

Obciążenie

$$\mathbb{E}(\hat{\beta}_i) - \beta_i = \frac{1}{1+\gamma}\beta_i - \beta_i = \frac{-\gamma}{1+\gamma}\beta_i$$

Wariancja

$$\mathbb{E}(\hat{\beta}_i)^2 = \frac{1}{(1+\gamma)^2}\mathbb{E}(\beta_i^2 + Z_i^2) = \frac{\beta_i^2 + \sigma^2}{(1+\gamma)^2}$$

$$\mathrm{Var}(\hat{\beta}_i) = \mathbb{E}(\hat{\beta}_i)^2 - (\mathbb{E}\hat{\beta}_i)^2 = \frac{\beta_i^2 + \sigma^2}{(1+\gamma)^2} - \frac{\beta_i^2}{(1+\gamma)^2} = \frac{\sigma^2}{(1+\gamma)^2}$$

ii)

Dla kazdego z powyzszych przypadków wyznaczymy parametry γ i λ umożliwiające osiągnięcie minimalnej wartości MSE.

Regresja grzbietowa

W podpunkcie **i)** wyliczyliśmy MSE, dlatego teraz będziemy minimalizować je po γ.

Oznaczmy $||\beta||^2 = c$

Nasza funkcja wygląda następująco:

$$MSE = f(\gamma) = \frac{\gamma^2 c + p\sigma^2}{(1+\gamma)^2}$$

Najpierw obliczymy pochodną i znajdziemy wartość krytyczną.

$$f'(\gamma) = \frac{2\gamma c - 2p\sigma^2}{(1+\gamma)^3}$$

$$f'(\gamma) = 0 \iff \gamma = \frac{p\sigma^2}{c}$$

Teraz sprawdzimy, czy jest to minimum.

$$f''(\frac{p\sigma^2}{c}) = \frac{2(p\sigma^2 - c)}{(1 + \frac{p\sigma^2}{c})^4}$$

Co jest zawsze dodatnie, czyli $\frac{p\sigma^2}{||\beta||^2}$ to wartość γ minimalizujaca MSE.

LASSO

Dla LASSO wyniki otrzymamy przy pomocy symulacji komputerowej.

iii), iv)

W poniższej tabeli znajdują się optymalne wartości γ i MSE regresji grzbietowej i metodzie LS w oparciu o teoretyczne wyniki oraz optymalne λ i MSE dla metody LASSO otrzymane za pomocą symulacji komputerowej.

	γ	ridge MSE	λ	LASSO MSE	LS MSE
$k = 5, \beta_{1,2}$	6.2500	68.9655	0.087	30.4361	500
$k = 20, \beta_1$	1.5625	195.1220	0.070	84.6507	500
$k = 100, \beta_1$	0.3125	380.9524	0.029	396.0581	500
$k = 20, \beta_2$	6.2500	68.9655	0.094	57.1170	500
$k = 100, \beta_2$	6.2500	68.9655	0.102	76.7784	500

v)

W tym podpunkcie chcemy policzyć średnią liczbę fałszywych odkryć i moc identyfikacji istotnych zmiennych dla LASSO z "optymalną" wartością parametru λ. Wiemy, że metoda LASSO przyjmuje zmienną X_j za istotną, gdy $|\beta_j^{LS}| \geq \lambda$ oraz, że $\beta_j^{LS} \sim N(c, 1)$. Stąd możemy wyznaczyć te wartości następująco:

1. Średnia liczba fałszywych odkryć:

$$(p - k)\mathbb{P}(|\beta_j^{LS}| \geq \lambda | \beta_j = 0) = 2(p - k)\Phi^{-1}(-\lambda)$$

2. Moc identyfikacji istotnej zmiennej:

$$\mathbb{P}(|\beta_j^{LS}| \geq \lambda | \beta_j = c) = \mathbb{P}(\beta_j^{LS} \geq \lambda \vee \beta_j^{LS} \leq -\lambda | \beta_j = c) = 1 - \Phi^{-1}(\lambda - c) + \Phi^{-1}(-\lambda - c)$$

Część symulacyjna

W tej części będziemy dokonywać obliczeń na modelu opisanym w zadaniu pierwszym. Zdefiniujemy jeszcze tylko macierz planu X.

Zadanie 2

W tym zadaniu wygenerujemy macierz $X_{500\times500}$ tak, że jej elementy są niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma = 1/\sqrt{n})$. Następnie dla wszystkich modeli liniowych rozważanych z zadaniu 1 wykonamy następujące działania:

a) Wyestymujemy β

   i) za pomocą LASSO i regresji grzebietowej z parametrami wybranymi za pomocą walidacji krzyżowej

   ii) za pomocą adaptacyjnego LASSO w dwóch wersjach

   - LASSO z walidacją krzyżową z wagami $w_i = \frac{1}{\hat{\beta}_i}$, gdzie $\hat{\beta}_i$ jest uzyskane za pomocą regresji grzbietowej z parametrem wybranym przy użyciu walidacji krzyżowej
   - adaptacyjne LASSO II, które polega na tym, że w pierwszym kroku wyznaczamy $\hat{\beta}$ korzystając z LASSO z kros-walidacją. Następnie w drugim kroku estymujemy σ. Dalej wyznaczamy wagi $w_i = \frac{\hat{\sigma}}{|\hat{\beta}_i|}$ (zmienne dla których $\hat{\beta}_i = 0$ usuwamy z modelu). Na pozostałych zmiennych stosujemy ważone LASSO z parametrem wygładzającym $\lambda = \hat{\sigma}\Phi^{-1}(1 - \frac{0.2}{2p})$

b) Porównamy $SE = ||\hat{\beta} - \beta||^2$ dla tych czterech metod.

c) Następnie zastosujemy technikę knockoff'ów dla LASSO i regresji grzbietowej tak aby kontrolować FDR na poziomie 0.2 i porównamy FDR i moc tych procedur z FDR i mocą dla zwykłego LASSO z kros-walidacją, a także dla obu wersji adaptacyjnego LASSO.

d) Na koniec powtórzymy punkty a)-c) 100 razy i porównamy MSE, FDR i moc dla powyższych metod.

b) Porównanie SE

Poniżej znajduje się tabela z wartościami SE dla różnych przypadków:

	SE			
	cv.RIDGE	cv.LASSO	ad.LASSO1	ad.LASSO2
$k = 5, \beta_{1,2}$	76.698	31.319	39.849	111.71
$k = 20, \beta_1$	306.24	138.88	128.99	308.92
$k = 100, \beta_1$	1521.2	409.35	485.94	573.71
$k = 20, \beta_2$	75.494	57.777	64.535	160.96
$k = 100, \beta_2$	75.658	80	93.142	80

Widzimy, że wartość SE mocno zależy od rozważanego wektora β. Największe wartości SE obserwujemy, gdy wektor beta jest pierwszej postaci i $k = 100$. Natomiast porównując je pośród różnych metod estymacji widzimy, że zdecydowanie największe wartości dostajemy dla regresji grzbietowej, a najmniejsze dla LASSO z kros-walidacją ewentualnie dla adaptacyjnego LASSO nr 1.

b) Porównanie mocy i FDR

Poniżej znajdują się tabele z wartościami FDR i mocy dla różnych przypadków:

FDR					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.84	0.89	0.84	0.66	0.71
$k = 20, \beta_1$	0.81	0.7	0.81	0.21	0.17
$k = 100, \beta_1$	0.59	0.43	0.58	0.04	0.04
$k = 20, \beta_2$	0.79	0.73	0.79	0.53	0.43
$k = 100, \beta_2$	0	0.59	0	0	0

Moc					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	1	1	1	0.7	0.8
$k = 20, \beta_1$	0.95	1	0.95	0.35	0.25
$k = 100, \beta_1$	0.99	0.93	0.99	0.22	0.24
$k = 20, \beta_2$	0.65	0.65	0.65	0.4	0.2
$k = 100, \beta_2$	0	0.24	0	0	0

Pierwsze co możemy zauważyć to to, że w obu tabelach dostaliśmy raczej duże wartości za wyjątekiem knockoff'ów. Jedynie w ostatnim przypadku bety wartości są bliskie zera za wyjątkiem pierwszego adaptacyjnego LASSO. Mniejszymi wartościami wyróżniają się estymatory uzyskane przy pomocy knockoff'ów. Na tej podstawie ciężko nam jest wybrać najlepszy estymator, bo wszystkie wydają się być dość słabe. Możliwe, że najlepszym wyborem byłoby pierwsze adaptacyjne LASSO, ale możemy wybrać inne ze względu na ich pewne własności.

Spróbujmy spojrzeć jeszcze na uśrednione wyniki.

d) Uśrednione wyniki dla 100 powtórzeń

SE				
	cv.RIDGE	cv.LASSO	ad.LASSO1	ad.LASSO2
$k = 5, \beta_{1,2}$	76.965	36.844	35.939	106.62
$k = 20, \beta_1$	306.49	105.03	95.949	204.72
$k = 100, \beta_1$	1526.1	512.79	628.61	637.95
$k = 20, \beta_2$	76.209	71.104	78.72	92.284
$k = 100, \beta_2$	76.075	82.019	94.312	83.228

Dla uśrednionych SE możemy obserwować te same własności co poprzednio. Znów najmnniejsze wartości dostajemy dla LASSO z kros-walidacją. Wyniki dla drugiego adaptacyjnego LASSO nieco

się ustabilizowały, ale pierwsze adaptacyjne LASSO nadal daje nam lepsze wyniki. Najgorzej z tych metod wypada regresja grzbietowa.

Poniżej znajdują się tabele z wartościami FDR i mocy dla różnych przypadków:

FDR					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.74	0.57	0.21	0.073	0.093
$k = 20, \beta_1$	0.76	0.57	0.76	0.14	0.14
$k = 100, \beta_1$	0.6	0.43	0.6	0.17	0.17
$k = 20, \beta_2$	0.62	0.65	0.056	0.091	0.091
$k = 100, \beta_2$	0.34	0.53	0.0077	0.048	0.048

Moc					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.91	0.89	0.26	0.23	0.13
$k = 20, \beta_1$	0.98	0.95	0.98	0.31	0.41
$k = 100, \beta_1$	0.97	0.85	0.97	0.35	0.25
$k = 20, \beta_2$	0.4	0.42	0.041	0.022	0.032
$k = 100, \beta_2$	0.064	0.13	0.0009	0.004	0.005

Wyniki znów wydają się być dość podobne. Gdyby zależało nam na wyborze metody dobrze wykrywającej istotne zmienne to byłoby to LASSO z kros-walidacją lub pierwsze adaptacyjne LASSO. Jednak gdybyśmy obawiali się fałszywych odkryć możliwe, że wybralibyśmy metodę knockoff'ów.

Zadanie 3

W ostatnim zadaniu powtórzymy zadanie drugie, gdy $X_i \sim N(0, \frac{1}{n}\Sigma)$, gdzie $\Sigma_{ii} = 1$ i $\Sigma_{ij} = 0.5$.

Wyniki dla jednego powtórzenia prezentują się następująco:

SE				
	cv.RIDGE	cv.LASSO	ad.LASSO1	ad.LASSO2
$k = 5, \beta_{1,2}$	78.452	49.896	48.579	52.281
$k = 20, \beta_1$	307.13	179.76	213.58	179.15
$k = 100, \beta_1$	1305.3	588.32	650.53	553.16
$k = 20, \beta_2$	76.573	71.077	90.854	71.142
$k = 100, \beta_2$	65.56	131.54	127.9	239.24

Widzimy, że wartości SE zachowują się dość podobnie do tych z poprzedniego zadania. Znów najlepsze wyniki dostajemy dla LASSO z kros-walidacją.

FDR					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.88	0.6	0.88	0	0
$k = 20, \beta_1$	0.75	0.6	0.74	0.375	0.083
$k = 100, \beta_1$	0.49	0.47	0.48	0.09	0.098
$k = 20, \beta_2$	0.73	0.52	0.73	0	0
$k = 100, \beta_2$	0.65	0.52	0.66	0	0

Moc					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.8	0.8	0.8	0	0
$k = 20, \beta_1$	0.85	0.9	0.85	0.5	0.55
$k = 100, \beta_1$	0.97	0.92	0.95	0.1	0.46
$k = 20, \beta_2$	0.65	0.5	0.65	0	0
$k = 100, \beta_2$	0.37	0.35	0.34	0	0

Znów widzimy, że dostajemy duże wartości dla mocy i FDR. Są one nawet większe niż poprzednio. Widzimy, że dość odporna na korelację wydaję się metoda knockoff'ów, co wydaje się być zrozumiałe, ponieważ jest ona skonstruowana w sposób, który uwzględnia korelacje. Dla pojedyńczego powtórzenia znów prawdopodobnie wybralibyśmy pierwsze adaptacyjne LASSO, ale to oczywiście zależy od specyfikacji naszych danych. Dlatego spójrzmy jeszcze na uśrednione wyniki.

Uśrednione wyniki znajdują się w tabelach poniżej:

SE				
	cv.RIDGE	cv.LASSO	ad.LASSO1	ad.LASSO2
$k = 5, \beta_{1,2}$	78.476	53.686	52.139	99.647
$k = 20, \beta_1$	306.68	156.01	164.33	218.14
$k = 100, \beta_1$	1314.1	657.68	749.74	692.21
$k = 20, \beta_2$	76.524	93.549	96.622	142.73
$k = 100, \beta_2$	65.476	151.29	161.62	198.98

Uśrednione wyniki potwierdzają nasze poprzednie obserwacje. W ogólności znów najlepszym estymatorem jest LASSO z kros-walidacją. Jednak regresja grzbietowa w porównaniu do drugiego zadania wypada lepiej na tle innych metod. Dla ostatniego przypadku beta daje nam nawet najlepsze wyniki.

FDR					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.85	0.56	0.84	0.099	0.122
$k = 20, \beta_1$	0.74	0.62	0.74	0.2	0.258
$k = 100, \beta_1$	0.52	0.49	0.49	0.322	0.19
$k = 20, \beta_2$	0.81	0.72	0.81	0.074	0.061
$k = 100, \beta_2$	0.66	0.59	0.65	0.199	0.044

Moc					
	cv.LASSO	ad.LASSO1	ad.LASSO2	knock.RIDGE	knock.LASSO
$k = 5, \beta_{1,2}$	0.81	0.63	0.81	0.132	0.13
$k = 20, \beta_1$	0.93	0.85	0.92	0.33	0.45
$k = 100, \beta_1$	0.93	0.9	0.91	0.1	0.62
$k = 20, \beta_2$	0.5	0.36	0.49	0.04	0.023
$k = 100, \beta_2$	0.36	0.29	0.34	0.031	0.006

Po raz kolejny obserwujemy, że uśrednione wyniki raczej pokrywają się z tymi dla jednego powtórzenia. Znów otrzymujemy raczej duże wartości oprócz knockoff'ów. Widzimy, że korelacja między zmiennymi najbardziej wpłynęła znacznie na nasze wyniki. Szczególnie dla drugiego adaptacyjnego LASSO. Jeżeli dla tych przypadków mielibyśmy wybierać najlepsze metody to prawdopodobnie znów byłoby to LASSO z kros-walidacją lub pierwsze adaptacyjne LASSO. Jednak to znów zależy od naszych potrzeb. Gdybyśmy bali się fałszywych odkryć najlepszym wyborem byłyby knockoff'y.

Podsumowanie

Patrząc na całokształt naszych wyników obserwowaliśmy, że często różniły się one dla różnych przypadków postaci wektora β co nas nie dziwi, ponieważ każda z metod jest skonstruowana w inny sposób. Porównując wyniki chcielibyśmy oczywiści wybrać najlepszą metodę dla tak skonstruowanego modelu. Ogólnie wyróżniały się dwie metody. Mianowicie LASSO z kros-walidacją i pierwsze adaptacyjne LASSO dla najlepszej predykcji. Oczywiście w zależności od naszych potrzeb możemy wybrać inną metodę, ale te wydawały się być najbardziej uniwersalne w naszym przypadku. Natomiast do wykrywania istotnych zmiennych z małym FDR najlepsze wydają się być knockoff'y.