prof. dr hab. Jan Mielniczuk
Instytut Podstaw Informatyki PAN
i Wydział MiNI PW

20.IX.2019

Recenzja rozprawy doktorskiej mgr Michała Kosa
Identyfikacja istotnych predyktorów w dużych bazach danych. Własności teoretyczne i zastosowania praktyczne.

Przedstawiana recenzja dotyczy rozprawy doktorskiej w przewodzie doktorskim mgr M. Kosa otwartym w Instytucie Matematycznym Wydziału Matematyki i Informatyki UWr i napisanej pod kierunkiem prof. M. Bogdan.

Zakres tematyczny rozprawy Przedmiotem rozprawy są metody selekcji istotnych (aktywnych) predyktorów w problemach regresyjnych, a bardziej konkretnie, kontroli intensywności fałszywych odrzuceń FDR i mocy procedury SLOPE, zaproponowanej w pracy Bogdan et al (2015). Procedura SLOPE dokonuje wyboru zmiennych aktywnych spośród p potencjalnych predyktorów na podstawie rozwiązania problemu wypukłego $\inf_b\{l(b) + \sum_{i=1}^p \lambda_i |b|_{(i)}\}$, gdzie $l(b)$ jest pewną wypukłą funkcją straty, $\lambda_1 \geqslant ... \geqslant \lambda_p > 0$ jest ciągiem ustalonych parametrów, a $|b|_{(1)} \geqslant .. \geqslant |b|_{(p)}$ uporządkowanym nierosnąco ciągiem wartości bezwzględnych współrzędnych wektora $b \in R^p$. Główne wyniki teoretyczne pracy dotyczace tych zagadnień dotyczą trzech sytuacji: ciągu modelu liniowych z kwadratową funkcją straty z ustaloną liczbą p (twierdzenie 3.12), ciągu rzadkich modeli liniowych gdy $p \to \infty$ i jednocześnie spełnione są warunki rzadkości dla liczby k aktywnych predyktorów: $k/p \to 0$ i $k^2 \log p/n \to 0$ (twierdzenie 3.30) i ciągu modeli logistycznych z logistyczną funkcją straty dla ustalonego p (twierdzenie 3.50). Twierdzenia zachodzą dla sytuacji planu losowego, gdy predyktory są niezależnymi zmiennymi losowymi o rozkładzie $N(0, 1/n)$, gdzie n jest licznością próby, a ciąg λ_i jest wybrany jako ciąg odpowiednio przeskalowanych kwantyli rozkładu normalnego rzędu $1 - qi/2p$, gdzie $0 < q < 1$ jest ograniczeniem, na którym chcemy kontrolować FDR. W pierwszym przypadku autor uzyskał oszacowanie $\text{limsup}\, FDR \leqslant q(p-k)/p$ oraz ograniczenie dolne na moc Π postaci $\text{liminf}\, \Pi \geqslant (1-q/2p)^k$, w drugim twierdzenie o zbieżności tych wielkości odpowiednio do 0 i 1, a w trzecim przypadku wynik analogiczny do wyniku twierdzenia 3.12 dla FDR. Istotnym elementem pracy

są również dwa wyniki dotyczące własności SLOPE dla ogólnej funkcji straty l (Lematy 3.1 i 3.2). W częsci praktycznej praca zawiera badania symulacyjne dotyczące zachowania się procedury selekcji SLOPE.

Ocena pracy. Uwagi ogólne. Procedura SLOPE oparta jest na podobnej metodologii co procedury Holma, Benjaminiego-Hochberga i Benjaminiego-Yakuteliego w których, w kontekście regresyjnym, ranguje się predyktory według swojej użyteczności i kryteria uznania predyktor za aktywny ustala się tym bardziej restrykcyjnie im wyższą rangę użyteczności ma dany predyktor. Metodologia ta jest istotnie różna od użytej w innej metody selekcji predyktorów opartej na adaptacyjnej metodzie LASSO, gdzie kara dla predyktora jest *zmniejszona*, jeśli wstępny esymator odpowiadającego mu współczynnika wskazuje na jego użyteczność. Ostatnie prace (Su i Candès (2016) i Bellec et al (2018)) pokazują, ze procedura SLOPE, która została wprowadzona jako procedura selekcji kontrolująca FDR ma również bardzo dobre własności estymacyjne i predykcyjne otrzymywane bez konieczności wykorzystania parametrów bazujących na nieznanych parametrach modelu, typu parametr k lub minimalna wartość aktywnych współczynników. Tematyka rozprawy dotyczy więc nowoczesnej i intensywnie badanej tematyki.

Praca doktorska mgr M. Kosa prezentuje wysoki poziom techniczny i wymagała od autora dużej wytrwałości i konsekwencji w przeprowadzeniu dowodów. Na podkreślenie zasługuje fakt, że rozważany plan eksperymentu w modelu regresyjnym jest losowy, co stwarza inne i często poważniejsze problemy techniczne niż w przypadku gdy zakłada się, że predyktory są deterministyczne. Autor rozwiązuje istotny problem dotyczący nowej i obecnie intensywnie badanej tematyki - jest to niewątpliwy sukces doktoranta. Za najbardziej istotne w pracy uważam twierdzenia 3.12 i 3.30 dotyczące modelu liniowego oraz lematy 3.1 i 3.2 podające dla ogólnej funkcji straty charakteryzację faktu, że zbiór $\hat{t} = \{i : \hat{\beta}_{SLOPE,i} \neq 0\} = r$ oraz $\hat{\beta}_{SLOPE,i} \neq 0$ przy zachodzeniu $\hat{t} = r$. Fakty te stanowią podstawę do oszacowania FDR w głównych twierdzeniach pracy.

Praca napisana jest bardzo starannie, z dużą dbałością o szczegóły techniczne. Jednakże nie udało się autorowi uniknąć pewnych problemów technicznych, które powodują, zę twierdzenia 3.47 i 3.50 wymagają wzmocnienia założeń dla zachowania poprawności tezy (por. uwagi wymienione poniżej). Ze względu na stopień skomplikowania dowodów, autor podaje przed dowodami formalnymi -co pomocne- ich zarys uzupełniony o dyskusję głównych kroków dowodowych.

Badania symulacyjne wskazują na stosunkowo dobrą kontrolę FDR przez procedurę SLOPE: problemy pojawiają się przy większym k ($k \geqslant 50$) i $n \leqslant 1000$, czyli dla trudniejszych modeli. Trzeba również tu zauważyć, że losowe plany rozpatrzone w pracy są bliskie planom ortogonalnym: dodatkowe trudności w selekcji pojawiają się przy niebadanych w pracy istotnych zależnościach między grupami predyktorów.

Poniższe problemy uważam za istotne:

1. Dowód twierdzenia 3.47 zawiera istotną lukę związaną z faktem, że (por. str. 79) musimy udowodnić, że $P(\inf_{b \in S(r_\eta, b^0)} h(b) > 0) > 1 - \eta$, a nie $P(h(b) > 0) > 1 - \eta$ dla każdego $b \in S(r_\eta, b^0)$. Teza twierdzenia jest prawdziwa przy mocniejszym założeniu $E||X_i||_2^3 < \infty$ (poprawa dowodu dokonana przez doktoranta). Dla planu normalnego rozpatrywanego pracy założenie to implikuje ograniczoność p (liczby potencjalnych predyktorów). Podobnego wzmocnienia założeń trzeba dokonać w twierdzeniu 3.50 (poprawa dowodu dokonana przez doktoranta), chyba że założy się, ze dane generowane są z ustalonego modelu logistycznego, którego współczynniki nie zmieniają się z wraz z n.

2. W drugiej części tw. 3.12 i tw. 3.30 oszacowanie z góry na moc jest uzyskane przez szacowanie prawdopodobieństwa $P_{supset} = P(t \subseteq \hat{t})$ wyboru nadzbioru zbioru zmiennych aktywnych $t = \{i : \beta_i^0 \neq 0\}$ i wykorzystaniu zależności $P_{supset} \leqslant \Pi$. Zarówno $P_{supset} \to 1$ jak i $FDR \to 0$ są implikowane przez zgodność selekcji $P(\hat{t} = t)$. W pracy przydałby sie komentarz na temat zależności między tymi miarami.

3. W części numerycznej SLOPE jest porównywany tylko z nie najsilniejszymi konkurentami: z procedurą Benjaminiego-Hochberga, która w modelu liniowym słabo kontroluje FDR, jak i podstawową procedurą Lasso (będącą przypadkiem szczególnym SLOPE). Ciekawsze byłoby porównanie SLOPE z metodami selekcji opartymi na innej metodologii, np. na wspomnianym powyżej adaptacyjnym Lasso.

4. W pracy wyczuwalna jest pewna nierównowaga: z jednej strony, co bardzo pozytywne, autor bardzo szczegółowo przedstawia dowody, natomiast staje sie bardzo lapidarny, wręcz bourbakowski, gdy chodzi o intuicje i metodologię samej procedury SLOPE, jej własności oraz procedur pokrewnych (np. część związana z Lematami 3.1 i 3.2). Oczywiście praca jest obszerna, ale powinna znaleźć się w niej referencja do pracy Bunea i inni (JSPI, 2006) w kontekście procedury BH dla modelu liniowego. Również praca Bellec et al (AS, 2018) cytowana jest tylko w kontekście doboru parametrów λ_i, choć jest ona istotna również z powodu rozpatrzenia w przypadku SLOPE bardziej ogólnych planów deterministycznych

niż plany ortogonalne. Powstaje w związku z tym naturalne pytanie, czy pewne wyniki z pracy nie dałyby sie uzyskać przez warunkowanie macierzą planu X i wykorzystanie wyników dla planu deterministycznego.

5. W doktoracie pominięty jest i nieopatrzony żadnym komentarzem problem estymacji odchylenia standardowego σ (co dla $p > n$ nie jest oczywistym problemem) i wpływu tej estymacji na wyniki w pracy.

6. Plan eksperymentu, w którym predyktory są o niezależnymi zmiennymi normalne o tym samym rozkładzie jest często rozpatrywany jako przykład planu losowego. Nie zmienia to faktu, że taki plan on mocno ograniczający, gdyż jest losowym odpowiednikiem planu ortogonalnego, dla którego problem selekcji zmiennych bardzo istotnie sie upraszcza. W pracy przydałby się komentarz, lub badanie symulacyjne, jak inny rozkład brzegowy zmiennych (np inny niż normalny rozkład subgaussowski), a przede wszystkim zależność między predyktorami wpływa na jakość metody.

Uwagi drobne

1. Pierwsza nierówność na str. 14 jest na tyle nieoczywista, że nosi nazwę tw. Hardy'go-Polyi-Littlewood'a;

2. str.12. W przypadku gdy $p > n$ zbiór S nie musi być wyznaczony jednoznacznie. Brakuje dyskusji, kiedy tak jest;

3. (3.10): korzysta się bez komentarza z ciągłości wartości własnych;

4. Lemat 3.20: H_r został zdefiniowany jako podzbiór R^p, a nie rozszerzonej przestrzeni $\bar{R}^p$, gdzie $\bar{R} = R \cup \{\infty\}$. Formalnie lemat nie jest prawdziwy;

5. (3.12): $\delta_n = s_n/n^{1/2}$ przy założeniach twierdzenia jest dowolnym ciągiem zbieżnym do 0 i tak byłoby prościej go oznaczyć;

6. Równość $||\hat{b}_{S^*} - b^0_{S^*}|| = ||\hat{b} - b^0||$ nie jest wykorzystana w przejściu (3.18) do (3.20). Wystarczy w tym miejscu oczywista nierówność $||\hat{b}_{S^*} - b^0_{S^*}|| \leqslant ||\hat{b} - b^0||$;

7. (3.33): $Supp(b^0) \subset S^*$ z definicji. Sformułowanie (3.33) i następne w dowodzie są nadmiarowe;

8. (3.26): korzysta się z definicji zbioru Ω_r, która pojawia się kilka stron dalej;

9. (3.41) korzysta się z zależności $\lambda_{k^*} \sim (\log p/k^*q)^{1/2}$, która pojawia sie explicite kilka stron dalej;

10. str. 59: pominięte σ w kilku wzorach dla λ_r;

11. dowód lematu 3.2: warto tu zauważyć, że technika małej perturbacji zastosowana w

tym dowodzie jest zbliżona do dowodów własności dla Lasso (istnienie rozwiązania o nośniku liczności nie większej niż n;
12. Literówki i niezręczności językowe: 31_5: powinno być $\sqrt{n}$ zamiast 1 w dwóch miejscach; 82^2: brak transpozycji, definicja u pod (3.61): $||b - b^0||$ zamiast $||b - b^0||^2$; 'str. 11: powinno być 'is positive regression dependent' zamiast ' has positive regression dependence'; 57^8 'shall' a nie 'will' (!).

Dodatkowe założenie prowadzące do uzdrowienia tw. 3.47 i 3.50, o których mowa powyżej, jest dosyć ograniczające, gdyż implikuje konieczność założenia skończoności liczby potencjalnych predyktorów, wydaje się, że poprawienie twierdzenia 3.47 można uzyskać stosując metody z pracy Fan i inni (2014). Natomiast, jak pisałem powyżej, twierdzenia 3.12 i 3.30 oraz eleganckie lematy 3.1 i 3.2 uważam za istotne osiągnięcie pracy.

Konkluzja Jestem przekonany, że mimo niedociągnięć rozprawa doktorska mgr M. Kosa spełnia *z nadmiarem* warunki stawiane rozprawom doktorskim w dziedzinie nauk ścisłych i przyrodniczych, dyscyplina matematyka i wnioskuję o dopuszczenie go do dalszych etapów przewodu doktorskiego.

Jan Mielniczuk