Wrocław, 10 września 2019 r.

dr hab. Maciej Wilczyński
Wydział Matematyki
Politechnika Wrocławska

Recenzja rozprawy doktorskiej magistra Michała Kosa zatytułowanej „Identyfikacja istotnych predyktorów w dużych bazach danych. Własności teoretyczne i zastosowania praktyczne"

1. Opis problematyki i wyników rozprawy. Rozprawa doktorska magistra Michała Kosa, napisana pod kierunkiem dr hab. Małgorzaty Bogdan, liczy 102 strony i składa się ze wstępu, trzech rozdziałów, krótkiego podsumowania oraz bibliografii obejmującej 20 pozycji. W dysertacji przedstawione zostały wyniki badań dotyczących zagadnień związanych z wykorzystaniem procedury *SLOPE* w problemach wielokrotnego testowania.

W wielu ważnych zagadnieniach wnioskowania statystycznego pojawia się problem, który w literaturze anglojęzycznej określany jest mianem *high dimensional data*. Ten termin oznacza, że liczba badanych cech, albo liczba estymowanych parametrów, jest znacznie większa od rozmiaru analizowanej próby, tzn. $p \gg n$. Taka sytuacja występuje na przykład wtedy, gdy konstruuje się model regresji liniowej, by za jego pomocą wyznaczyć geny mające wpływ na pewną cechę ilościową. Potencjalnych regresorów (genów) są wówczas tysiące, a liczba obserwacji (pacjentów) rzadko kiedy przekracza sto. Klasyczne podejście, polegające na obliczeniu estymatora najmniejszych kwadratów, nie jest dobrym rozwiązaniem. Estymator zachowuje się niestabilnie i nie jest wyznaczony jednoznacznie. Na dodatek, większość jego współrzędnych jest niezerowa, co utrudnia prawidłową interpretację modelu, gdyż zawiera on zbyt dużo zmiennych objaśniających. Podobny problem pojawia się także i wtedy, gdy za pomocą regresji logistycznej tworzy się model opisujący wpływ wielu zmiennych objaśniających na dychotomiczną zmienną objaśnianą. W ostatnich dwudziestu lat zaproponowano wiele nowych metod wnioskowania statystycznego, które można z powodzeniem wykorzystywać także i wtedy, gdy $p \gg n$. Wśród nich bardzo istotną rolę odgrywają *metody regularyzacji*. Najpopularniejszą z nich jest *Lasso* [Tibshirani (1996)]. Na uwagę zasługują także *Elastic net* [Zou and Hastie (2005)] i *Dantzig selector* [Candes and Tao (2007)]. Niezwykle ważną modyfikacją *Lasso* jest *SLOPE* [Bogdan, Berg, Sabatti, Su and Candes (2015)].

W pierwszych dwóch rozdziałach rozprawy autor przedstawił najważniejsze pojęcia dotyczące tematyki dysertacji i podał kilka znanych wyników, wykorzystywanych w kolejnych częściach pracy. Omówił między innymi problemy pojawiające się przy analizie *high dimensional data* i pokrótce opisał metodę *Lasso*. Przypomniał podstawowe pojęcia związane z zagadnieniem wielokrotnego testowania, na przykład *family-wise error rate (FWER)* i *false discovery rate (FDR)*. Następnie przytoczył twierdzenia opisujące własności kilku procedur wielokrotnego testowania, w szczególności procedury Benjaminiego-Hochberga. Zdefiniował także metodę *SLOPE*, opisał jej własności i zacytował dwa ważne rezultaty dotyczące wykorzystania tej metody we wnioskowaniu o nieznanym wektorze $\boldsymbol{\beta}$ w modelu regresji liniowej

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

w którym macierz eksperymentu $\boldsymbol{X}$ ma wymiar $n \times p$, a wektor błędów $\boldsymbol{\varepsilon} \stackrel{D}{=} N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$.

Najważniejsza część rozprawy jest zawarta w Rozdziale 3. Autor opisał w nim uzyskane przez siebie wyniki teoretyczne dotyczące wykorzystania procedury *SLOPE* do kontroli *FDR* w modelach regresji liniowej oraz logistycznej.

W pracy opublikowanej w 2015 roku Bogdan, Berg, Sabatti, Su i Candes zaproponowali, by w modelu regresji liniowej (1) estymować nieznany wektor $\boldsymbol{\beta}$ za pomocą reguły decyzyjnej $\widehat{\boldsymbol{\beta}}_{SLOPE} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$ zdefiniowanej wzorem

$$\widehat{\boldsymbol{\beta}}_{SLOPE} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[l(\boldsymbol{\beta}) + \sum_{i=1}^{p} \lambda_i |\beta|_{(i)} \right], \tag{2}$$

gdzie $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ jest ustalonym ciągiem liczbowym, $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \ldots \geq |\beta|_{(p)}$ to uporządkowane nierosnąco moduły współrzędnych wektora $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, a $l(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$. Zaproponowali także, by w problemie wielokrotnego testowania

$$H_{0,j} : \beta_j = 0 \text{ vs } H_{1,j} : \beta_j \neq 0, \quad j = 1, 2, \ldots, p, \tag{3}$$

użyć procedury *SLOPE*, która odrzuca hipotezę $H_{0,j} \iff \widehat{\beta}_j \neq 0$ i udowodnili, że przy odpowiednim doborze wag $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ ta procedura kontroluje *FDR*, gdy $n \geq p$, a macierz $\boldsymbol{X}$ jest deterministyczna i spełnia warunek $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$.

W pracy pochodzącej z 2016 roku Candes i Su rozpatrzyli sytuację, w której $p \gg n$, a macierz $\boldsymbol{X}$ jest losowa i niezależna od $\boldsymbol{\varepsilon}$, przy czym jej elementy są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, 1/n)$. Pokazali między innymi, że gdy $p \to \infty$, $n \to \infty$ i $\dfrac{\lg p}{n} \to \infty$, to estymator $\widehat{\boldsymbol{\beta}}_{SLOPE}$ rzadkiego wektora $\boldsymbol{\beta}$, skonstruowany w oparciu o procedurę *SLOPE*, jest asymptotycznie minimaksowy przy funkcji straty $L(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}) := \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ i odpowiednio dobranych wagach $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$.

W pierwszej części Rozdziału 3. doktorant rozważył problem wielokrotnego testowania (3) w znacznie ogólniejszej wersji, w której symbol l, pojawiający się we wzorze (2), oznacza **dowolną wypukłą i różniczkowalną** funkcję z $\mathbb{R}^p$ w $\mathbb{R}$. Przy tych założeniach autor udowodnił ważne Twierdzenie 3.4, w którego tezie pojawia się formuła opisująca w jaki sposób *FDR* zależy od $\nabla\left(l(\widehat{\boldsymbol{\beta}}_{SLOPE})\right)$, czyli od gradientu funkcji l obliczonego w punkcie $\widehat{\boldsymbol{\beta}}_{SLOPE}$.

Następnie doktorant zajął się modelem regresji liniowej (1), w którym p jest ustalone i $n \to \infty$, przy czym dla każdego $n \geq 1$ macierz eksperymentu $\boldsymbol{X}$ jest **losowa i niezależna** od $\boldsymbol{\varepsilon}$, a jej elementy są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, 1/n)$. Przy tych założeniach autor najpierw udowodnił Twierdzenie 3.10 mówiące o tym, że ciąg $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{SLOPE} - \boldsymbol{\beta}\right)$ jest ograniczony według prawdopodobieństwa. Następnie wykorzystał ten rezultat oraz wspomnianą powyżej formułę do udowodnienia Twierdzenia 3.12, z którego wynika, że w problemie wielokrotnego testowania (3) procedura *SLOPE* asymptotycznie kontroluje *FDR*. To oznacza, że dla każdego $q \in (0, 1)$ można tak dobrać nieujemne wagi $\lambda_1(q) \geq \ldots \geq \lambda_p(q)$, by zachodziła nierówność

$$\limsup_{n \to \infty} FDR \leq \frac{q(p-k)}{p},$$

gdzie k oznacza liczbę niezerowych współrzędnych wektora $\boldsymbol{\beta}$. Pokazał także, że jeśli moduł każdej z niezerowych współrzędnych wektora $\boldsymbol{\beta}$ jest co najmniej równy $2\lambda_1(q)$, to moc Π tej

procedury spełnia warunek

$$\liminf_{n\to\infty} \Pi \geq \left(1 - \frac{q}{2p}\right)^k.$$

W kolejnej części Rozdziału 3. doktorant zajął się modelem regresji liniowej (1), który różni się od poprzedniego tym, że p też jest rozbieżne do ∞ i to szybciej niż n, a liczba niezerowych współrzędnych wektora $\boldsymbol{\beta}$ spełnia warunek $k(n) = o\left(\sqrt{\frac{n}{\log p}}\right)$. Ponadto, niezerowe sygnały β_j są dostatecznie silne, tzn. istnieje liczba $\delta > 0$, taka że $\min_{\beta_j \neq 0} |\beta_j| \geq 2\sigma(1+\delta)\sqrt{2\log p}$. Główne wyniki, odpowiadające tym założeniom, autor zawarł w Twierdzeniu 3.30. Wynika z niego, że w problemie wielokrotnego testowania (3) procedura *SLOPE*, z odpowiednio dobranymi wagami $\lambda_1, \ldots, \lambda_p$, gwarantuje, że $FDR \to 0$ i $\Pi \to 1$.

W ostatniej części Rozdziału 3. doktorant rozpatrzył model regresji logistycznej, opisujący wpływ p zmiennych objaśniających $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ na dychotomiczną zmienną objaśnianą Y, przyjmującą wartości -1 albo 1 z prawdopodobieństwami

$$P(Y = 1 \,|\, \boldsymbol{x}) = 1 - P(Y = -1 \,|\, \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'\boldsymbol{\beta})},$$

zależnymi od jest nieznanego wektora $\boldsymbol{\beta} \in \mathbb{R}^p$. W tym modelu p jest ustalone, a rozmiar próby $n \to \infty$, przy czym dla każdego $n \geq 1$ macierz eksperymentu $\boldsymbol{X}$ jest **losowa**, a jej elementy są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, 1/n)$. Najpierw, dla każdego $n \geq 1$, Autor wyznaczył postać funkcji $l(\boldsymbol{\beta})$, pojawiającej się we wzorze (2) na estymator $\widehat{\boldsymbol{\beta}}_{SLOPE}$. Następnie, przy pewnych dodatkowych założeniach, udowodnił Twierdzenie 3.47 mówiące o tym, że ciąg $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{SLOPE} - \boldsymbol{\beta}\right)$ jest ograniczony według prawdopodobieństwa. Na koniec, wykorzystując ten fakt, udowodnił Twierdzenie 3.50, z którego wynika, że w problemie wielokrotnego testowania (3) procedura *SLOPE* asymptotycznie kontroluje *FDR*, gdy sygnały $\boldsymbol{\beta}$ słabną ze wzrostem n.

Własności procedury *SLOPE*, wynikające z Twierdzeń 3.12, 3.30 i 3.50 doktorant zilustrował obszernymi i dokładnie omówionymi badaniami symulacyjnymi, których wyniki również można znaleźć w Rozdziale 3.

2. Ocena rozprawy. Ważniejsze rezultaty przedstawione w pracy doktorskiej pana Michała Kosa zostały zawarte w dwóch artykułach napisanych wspólnie z promotorką. Jeden z tych artykułów, zatytułowany "On the asymptotic properties of SLOPE", został już przyjęty do publikacji w czasopiśmie Sankhya A.

Rozprawa na pewno jest wartościowa, ale nie czyta się jej łatwo, bo złożoność problematyki rozważanej w pracy spowodowała, że dowody najważniejszych twierdzeń są długie i skomplikowane. Przy ich przeprowadzaniu doktorant wykazał się sprawnością rachunkową, a także pomysłowością i dogłębną znajomością literatury. Mam jednak drobne zastrzeżenia dotyczące argumentów stosowanych przez doktoranta w dowodach niektórych faktów. Nie wpływają one na moją opinię o dysertacji. Oto niektóre z tych zastrzeżeń.

1. **Strona 31.** Skąd wynika równość

$$\begin{aligned}\sqrt{n}\max\left(|(\sigma_{min}(\boldsymbol{X}))^2 - 1|, |(\sigma_{max}(\boldsymbol{X}))^2 - 1|\right) &= \\ = \max\left(|(\sigma_{min}(\sqrt{n}\boldsymbol{X}) - 1)(\sigma_{min}(\boldsymbol{X}) + 1)|, |(\sigma_{max}(\sqrt{n}\boldsymbol{X}) - 1)(\sigma_{max}(\boldsymbol{X}) + 1)|\right)&?\end{aligned}$$

2. **Strona 32.** W dowodzie Lematu 3.17 autor wykorzystuje Twierdzenie Vershynina do uzasadnienia nierówności

$$\Pr(\|\boldsymbol{\varepsilon}/\sigma\|_2 \leq \sqrt{n} + 0.5s_n) \geq 1 - \mathrm{e}^{\frac{-s_n^2\sigma^2}{8}},$$

w której wektor losowy $\boldsymbol{\varepsilon}/\sigma \stackrel{D}{=} N(\mathbf{0}, \boldsymbol{I}_n)$.

   (a) Dlaczego po prawej stronie równości pojawia się σ, skoro rozkład zmiennej losowej $\|\boldsymbol{\varepsilon}/\sigma\|_2$ nie zależy od tego parametru?

   (b) Czy po prawej stronie równości, zamiast $\sqrt{n}$, nie powinna znaleźć się liczba $\mathbb{E}(\|\boldsymbol{\varepsilon}/\sigma\|_2)$?

3. **Strona 34.** W dowodzie Lematu 3.22 przypuszczalnie jest błąd, gdyż stwierdzenie *Let us assume* $|B|_{(i)} = |B_j|$ *and* $j < i$ *then:*

$$\big||A_i| - |B|_{(i)}\big| = ||A_i| - |B_j|| \leq ||A_j| - |B_j||$$

może być nieprawdziwe (np. dla $A_1 = 4, A_2 = 3$, $B_1 = 4, B_2 = 5$ oraz $i = 2$).

Z pewnością nieprawdziwa jest równość $||A_i| - |B_i|| = |A_i - B_i|$.

Redakcja pracy jest staranna i rzetelna. Każdy z dowodów najważniejszych twierdzeń jest poprzedzony krótkim wstępem, wyjaśniającym kolejne kroki rozumowania. Dzięki temu wyniki uyskane przez doktoranta są prezentowane w sposób czytelny. W rozprawie pojawia się wprawdzie niewielka liczba przejęzyczeń, ale łatwo można je wychwycić i poprawić. Kilka przykładów takich uchybień.

1. **Strona 29.** Autor odwołuje się do Lematu 3.15 zamiast do Corollary 3.15.

2. **Strona 32.** W dowodzie Lematu 3.17 uzasadnieniem jednego z wniosków jest nierówność 2.2, a powinno być Twierdzenie 2.2.

3. **Strona 38.** Pojawia się nigdzie niezdefiniowany symbol $\widetilde{\lambda}_i$.

Rozprawa doktorska magistra Michała Kosa zawiera oryginalne wyniki autora, które są nowe i interesujące. Wnoszą one istotny wkład do rozwoju badań nad własnościami procedury *SLOPE* - jednej z najnowszych metod wykorzystywanych we wnioskowaniu statystycznym. W swojej dysertacji doktorant rozwiązał kilka nietrywialnych problemów, stosując przy tym różne, często zaawansowane techniki probabilistyczne i statystyczne. Uzyskane przez niego rezultaty są nie tylko ważne pod względem teoretycznym, ale także mogą mieć szerokie zastosowanie praktyczne, na przykład w genetyce, w medycynie, w radiologii i w finansach.

3. Konluzja. Uważam, że rozprawa magistra Michała Kosa **spełnia wymagania ustawowe** stawiane pracom doktorskim. Wnoszę o dopuszczenie jej autora do dalszych etapów przewodu doktorskiego.

Maciej Wilczyński