Warszawa, 11 lutego 2016

prof. dr hab. Wojciech Niemiro

Instytut Matematyki Stosowanej
Uniwersytet Warszawski
i
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika, Toruń

Recenzja rozprawy doktorskiej

pt. „SELECTING RELEVANT GROUPS OF EXPLANATORY VARIABLES VIA CONVEX OPTIMIZATION METHODS WITH THE FALSE DISCOVERY RATE CONTROL"

Damiana Brzyskiego

Uwagi wstępne. Rozprawa doktorska Damiana Brzyskiego mieści się w nowatorskim, szybko rozwijającym się nurcie statystyki. Dotyczy tematyki, która jest ważna i intensywnie badana. Istotne są motywacje, pochodzące głównie z biologii i genetyki. Rozwój technologii stworzył zapotrzebowanie na nowy typ procedur statystycznych. Ogromne bazy danych, zawierające wiele „zmiennych objaśniających" (p) dla umiarkowanie licznej próbki „obiektów" (n) wymagają algorytmów wyboru modelu (selekcji istotnych zmiennych). To zagadnienie okazało się wyzwaniem dla społeczności statystyków i zaowocowało mnóstwem prac, zarówno teoretycznych jak i zorientowanych na zastosowania. Adekwatna teoria matematyczna jest w fazie rozwoju i nie ma charakteru zamkniętej całości. Ścisłe rozumowania przeplatają się z nowymi pomysłami weryfikowanymi empirycznie, symulacyjnie lub uzasadnionymi heurystycznie. Na tym tle należy rozpatrywać i oceniać dokonania Brzyskiego. W jego rozprawie równie ważne są twierdzenia, dowody, symulacje i propozycje algorytmów.

Omówienie zawartości rozprawy. Najprostszym i najważniejszym zadaniem wyboru modelu jest rozpoznanie podzbioru niezerowych współczynników regresji liniowej. Brzyski rozwija pomysł, który pojawił się stosunkowo niedawno w pracach M. Bogdan, E. J. Candesa i in. Chodzi o algorytm SLOPE, który jest pewnym uogólnieniem LASSO, pozwalającym na kontrolę frakcji "fałszywych odkryć" (FDR) na wzór klasycznej metody Benjaminiego-Hochberga. SLOPE wzbudził szerokie zainteresowanie wśród specjalistów i jest obiektem dalszych

badań. Główny wkład Brzyskiego polega na opracowaniu modyfikacji metody SLOPE (nazwanej gSLOPE), w której dokonuje się selekcji grup zmiennych, a nie pojedynczych zmiennych. Ten kierunek ma ważne uzasadnienia i motywacje. Jak przekonująco argumentuje Brzyski w swojej rozprawie, w sytuacji gdy zmienne objaśniające są silnie skorelowane w grupach, wybór pojedynczej współrzędnej ma niewielki sens, należy traktować grupy jako całości.

Centralnym wynikiem rozprawy jest Twierdzenie 5.3.1. Stwierdza ono, że algorytm gSLOPE pozwala kontrolować FDR na poziomie grup, przy założeniu ortogonalności macierzy planu, $X^T X = I_p$. Jest to wynik analogiczny do Twierdzenia 1.3 z pracy Bogdan, Berga, Su i Candesa z 2013 (w późniejszej wersji opublikowanej w *Ann. Appl. Statist.* 2014), zacytowanej jako Twierdzenie 2.2.1 we wstępnej części rozprawy. Zaproponowana w rozprawie modyfikacja SLOPE (wzór (5.1)) polega na zastąpieniu w składniku "regularyzującym" normy $J_\lambda(b)$ przez "grupową" normę $J_\lambda([[b]]_I)$, gdzie $[[b]]_I = (||b_{I_1}||_2, \ldots, ||b_{I_m}||_2)^T$. Dodatkowo wprowadzone są wagi w_i, dające możliwość właściwego uwzględniania różnych liczności grup. O ile w twierdzeniu Bogdan, Berga, Su i Candesa parametry λ są wyznaczone przez kwantyle rozkładu normalnego, $\lambda_i = \Phi^{-1}(1 - qi/2p)$, to w recenzowanej rozprawie są one zastąpione przez maksima ważonych kwantyli rozkładów χ_{l_i}, gdzie l_i są licznościami grup. Brzyski bardzo uważnie przeanalizował dowód wcześniejszego twierdzenia Bogdan, Berga, Su i Candesa i wprowadził takie modyfikacje algorytmu, żeby otrzymać w pełni analogiczną tezę, na poziomie grup zamiast pojedynczych współrzędnych. Ponadto Brzyski zadbał o to, żeby jego algorytm gSLOPE zachował jedną z podstawowych zalet podstawowej wersji SLOPE, mianowicie istnienie efektywnej metody obliczeniowej. Zauważa mianowicie, że do obliczania gSLOPE można zastosować metodę *proximal gradient*. Kwestie związane z numerycznym obliczaniem gSLOPe są omówione w podrozdziale 5.6, zaś w Dodatku C Procedura 3 (*proximal gradient*) jest przez Autora uzupełniona przez podanie kryterium stopu. W kolejnych podrozdziałach następujących po głównym Twierdzeniu 5.3.1, Autor uzupełnia teoretyczny wynik serią wskazówek, jak dobierać parametry algorytmu w praktyce. Chodzi tu o wybór wag (Podrozdział 5.5), wybór λ_i w sytuacji bardziej realistycznej niż ortogonalna (Podrozdział 5.6, Procedury 4 i 5), estymacja wariancji błędu (Podrozdział 5.7, Procedura 6).

Rozdział 6 jest poświęcony badaniom symulacyjnym i stanowi ważny składnik rozprawy. Autor zastosował następującą metodologię. Macierz planu X o wymiarach $n = 5402$ i $p = 12677$ jest dużym podzbiorem rzeczywistej bazy danych NFBC. W tej macierzy wyróżniono $m = 2522$ grup zmiennych, znajdując skupienia silnie skorelowanych kolumn przy pomocy algorytmu klasteryzacji HCA. Wektor β ustalono arbitralnie, wybierając po jednej niezerowej współrzędnej z grup "istotnych", odpowiednio "rzadkich". Następnie wygenerowano w sztuczny sposób wektor zmiennych objaśnianych y. Do tak spreparowanych danych zastosowano gSLOPE i badano zachowanie FDR. Należy podkreślić, że przeprowadzone doświadczenia dotyczyły sytuacji realistycznej, w której nie są spełnione założenia Twierdzenia 5.3.1 i parametry algorytmu są dobierane heurystycz-

nie. Wyniki doświadczeń potwierdzają użyteczność gSLOPE i w szczególności zachowanie poziomu FDR.

Rozdział 7 nieco odbiega od głównego tematu rozprawy. Rozpatrywana jest inna metoda selekcji zmiennych tzw. *Ordered Dantzig Selector* (ODS) i wersja „grupowa" (gODS). Klasyczna procedura *Dantzig Selector* wynaleziona przez Candesa i Tao w 2005 r. polega na minimalizacji $||b||_1$ przy ograniczeniach $||X^T(y - Xb)||_\infty \leqslant C$. Zapropnowane modyfikacje polegają na zastąpieniu normy $||\cdot||_1$ przez $J_{\lambda,I,W}$ (sortowana, grupowana i ważona norma ℓ_1). Jak rozumiem, obie modyfikacje (ODS i gODS) są dziełem Brzyskiego. Interesujące Twierdzenie 7.1.1 pokazuje równoważność ODS i SLOPE przy założeniu ortogonalności planu. Rozumiem, że równoważność gODS i gSLOPE w tej sytuacji jest problemem otwartym, nieprawdaż?

Ocena osiągniętych wyników. Jak już wspomniałem, istotnymi składnikami rozprawy są nie tylko 1) twierdzenia matematyczne ale, w równym stopniu, 2) propozycje algorytmów uzasadnionych heurystycznie, których użyteczność sprawdza się doświadczalnie.

1) Zacznę od wyników matematycznych. Najważniejsze z nich to Twierdzenie 5.3.1. Postać tego twierdzenia i metoda dowodu są podobne do wcześniejszego wyniku Bogdan, Berga, Su i Candesa. Potrzebne w ogólniejszej sytuacji modyfikacje są jednak dalekie od oczywistości. Dowód Twierdzenia 5.3.1 nie jest co prawda nowatorski, ale wymagał od Autora głębokiego zrozumienia zagadnienia i solidnej pracy przygotowawczej (seria pomocniczych wyników w Rozdziale 4). Potrzebne narzędzia matematyczne to przede wszystkim teoria optymalizacji wypukłej. Rozprawa Brzyskiego świadczy o znakomitym opanowaniu przez Autora współczesnej, zaawansowanej wiedzy z tej dziedziny i o umiejętności jej twórczego wykorzystania. Podobne uwagi dotyczą w jeszcze większym stopniu Twierdzenia 7.1.1. Jest to wynik ściśle teorio-optymalizacyjny. Dowód opiera się na ciągu subtelnych i trudnych wyników pomocniczych w podrozdziale 7.1. Pozwolę sobie w tym miejscu na ogólniejszą refleksję. Jest dla mnie zadziwiające, że (nie tylko w recenzowanej rozprawie) problemy optymalizacyjne w statystyce wysuwają się na pierwsze miejsce, spychając w cień treści probabilistyczne. Wracając do rozprawy Brzyskiego, zainteresowany probabilistyką czytelnik znajdzie tu eleganckie (choć łatwe) Twierdzenie 5.4.1 (które pełni istotną rolę w ustawieniu symulacji).

2) Heurystyczne rozważania w Podrozdziałach 5.5, 5.6 i 5.7 prowadzą do sformułowania pełnego algorytmu gSLOPE, gotowego do użycia w realistycznych zastosowaniach. Sprawdzianem użyteczności są symulacje w Rozdziale 6. Uważam przeprowadzone przez Brzyskiego doświadczenia symulacyjne za poprawnie zaprojektowane. Wynikające z nich wnioski za dobrze uzasadnione. Podoba mi się zasadnicza koncepcja użycia rzeczywistych danych w macierzy X i sztucznej konstrukcji zmiennej odpowiedzi y. Z jednej strony, jest to naśladowanie

realistycznej sytuacji, z drugiej strony daje możliwość porównania wyników testowanego algorytmu z tzw. „*ground truth*".

Uwagi krytyczne. Rozwijana przez Brzyskiego tematyka nie jest moją ścisłą specjalnością. Patrząc z pewnej perspektywy mogę nie doceniać technicznych szczegółów i trudności, widzę natomiast niewykończone i niezadowalające elementy teorii w trakcie jej powstawania. Główne twierdzenie rozprawy, 5.3.1, odpowiada na inne pytanie, niż to, które nas naprawdę interesuje. Podstawową motywacją algorytmu gSLOPE jest sytuacja $p \gg n$ (czy nawet $m > n$?). Tymczasem założenia twierdzenia mogą być spełnione tylko dla $p \leqslant n$. Autor oczywiście zdaje sobie z tego sprawę i po przeanalizowaniu przypadku „ortogonalnego" przechodzi do przypadku „niemal ortogonalnego". W tej sytuacji, matematyczne dowody są zastąpione heurystyką. Rozważania w Podrozdziale 5.6 „*Near orthogonal situation*" bazują, jak się wydaje, na śmiałym spostrzeżeniu, że jeśli mamy $\mathbb{E}X_i^T X_j = \delta_{ij}$ dla wektorów losowych $X_1, \ldots, X_p$, to ich realizacje są „niemal ortogonalne" w przestrzeni n-wymiarowej, gdzie $n \ll p$. Zapewne, stoją za tym spostrzeżeniem słuszne intuicje, które mogą być sformalizowane i uściślone. Na razie muszę poprzestać na stwierdzeniu, że z heurystycznych rozważań w Podrozdziale 5.6 wynikają konkretne recepty na wybór λ_i (Procedury 4 i 5), i te recepty sprawdzają się w doświadczeniach symulacyjnych. Jest to solidny argument, choć nie jest to matematyka. Podobna uwaga dotyczy konserwatywnego charakteru oszacowań FDR w algorytmie gSLOPE. Rozpatrywany w Twierdzeniu 5.3.1 wybór λ_i daje ścisłą nierówność, ale oszacowanie z góry bywa dość odległe od „docelowej wartości" FDR. Inna „recepta" podana we wzorze na λ^{mean} (5.20) zapewne pozwala lepiej kontrolować FDR, ale ten fakt nie jest matematycznie udowodniony, tylko uzasadniony symulacyjnie. Uważam przy tym, że przemieszanie matematyki z heurystyką jest konieczne na obecnym etapie rozwoju teorii.

Redakcja rozprawy i drobne uwagi. Praca jest napisana starannie, z dbałością o jasność wykładu. Rozważania o charakterze technicznym są poprzedzone wyjaśnieniem motywacji. Zauważyłem tylko niewielką liczbę drobnych błędów, niekonsekwencji i gorzej zredagowanych fragmentów.

- Nie jest dla mnie jasne dlaczego Autor na str. 10 przytacza argumenty na rzecz użycia normy $||X_{I_i} b_{I_i}||$ zamiast $||b_{I_i}||$, podczas gdy w zasadniczym algorytmie gSLOPE używa $||b_{I_i}||$ (str. 23, wzór (5.1).
- Proposition 3.1.1: chyba powinno być $x_{\pi^{-1}}$. Brakuje też słowa *matrix* przed P_π.
- W Lemacie 4.2.2 brakuje mi jawnego zaznaczenia, że b^* jest rozpatrywane jako funkcja y przy ustalonych innych elementach problemu, λ, D.
- Na str. 26 na dole mówi się swobodnie o niecentralnych rozkładach χ, a dopiero na str. 27 pojawia się definicja, i to tylko w przypadku rozkładów centralnych.

- W Twierdzeniu 5.6.1, odwołanie do definicji zbioru C_λ w Podrozdziale 4.2 utrudnia czytelność. Może chodzi o Podrozdział 4.1? Ale i tutaj wymiary się chyba nie zgadzają: $C_\lambda \subset \mathbb{R}^p$? Lepiej powtórzyć definicję, dopasowaną do kontekstu.

- W dowodzie Tw. 5.3.1 na str. 28 mamy współrzędne $1, \ldots, m_0$ nieistotne a $m_0 + 1, \ldots, m$ istotne. Na str. 33 na odwrót: współrzędne $1, \ldots, s$ są istotne a $s + 1, \ldots, m$ nieistotne. Dlaczego? Formalnie to dopuszczalne, ale czytanie utrudnia.

- Może warto byłoby zacytować pracę Liu i in. „The Group Dantzig Selector" (AISTATS 2010) i skomentować jej wkład. Wydaje się związana z tematyką rozprawy.

Wnioski. Rozprawa Brzyskiego stanowi wartościowy wkład do statystyki i sądzę, że zostanie doceniona przez specjalistów. Matematyczna część pracy jest solidna, ścisła i zawiera nietrywialne twierdzenia. Moje „uwagi krytyczne" wskazują tylko, że wartość rozprawy nie ogranicza się do twierdzeń i dowodów. Rozprawa spełnia warunki stawiane pracom doktorskim z matematyki. Wnoszę o dopuszczenie Pana Damiana Brzyskiego do dalszych etapów przewodu doktorskiego.

Warszawa, 11 lutego 2016

Wojciech Niemiro