prof. dr hab. Zbigniew Szkutnik
Wydział Matematyki Stosowanej AGH
al. Mickiewicza 30, 30-059 Kraków

Recenzja rozprawy doktorskiej mgr Agnieszki Prochenki pt. „Delete or merge regressors algorithm"

(wykonana na zlecenie Instytutu Podstaw Informatyki PAN w Warszawie)

1. Forma i tematyka rozprawy

Rozprawa poświęcona jest konstrukcji i badaniu własności algorytmów wyboru zmiennych w modelu liniowym oraz uogólnionym modelu liniowym w przypadku, gdy występują zmienne objaśniające zarówno typu ciągłego jak i jakościowego. Wybór modelu oznacza tu wybór zmiennych objaśniających z nadmiarowej listy potencjalnych zmiennych objaśniających, a w przypadku zmiennych jakościowych także łączenie tych poziomów zmiennych, których efekty nie różnią się istotnie. Chodzi przy tym o zbudowanie modelu „oszczędnego", tzn. zawierającego możliwie mało parametrów, ale zachowującego dobre własności prognostyczne. Jest to jeden z klasycznych i ważnych problemów statystyki, który w ostatnich latach nabrał dodatkowego znaczenia i wygenerował nowe wyzwania wraz z pojawieniem się tzw. danych i problemów wysokowymiarowych, w których liczba p potencjalnych zmiennych objaśniających bywa wielokrotnie wyższa od liczby n obserwowanych przypadków (np. w problemach genetycznych). Klasyczne metody estymacji i doboru zmiennych nie działają w takich przypadkach. Powstały więc liczne algorytmy wprowadzające do procesu estymacji różnie definiowane funkcjonały kary za złożoność modelu - wśród nich najbardziej znana procedura Lasso i jej liczne modyfikacje. Celem jest przy tym konstrukcja metod, które przy założeniu, że prawdziwy model zawiera stosunkowo niewiele istotnych parametrów mają tzw. własność wyroczni, tzn. z prawdopodobieństwem dążącym do jeden, gdy liczność próby rośnie do nieskończoności, wybierają model prawdziwy, a dodatkowo estymatory parametrów wybranego modelu mają rozkłady asymptotyczne takie, jakby zbiór istotnych parametrów był z góry znany. Ze względu na dużą złożoność obliczeniową i konieczność rozpatrzenia bardzo dużej liczby podmodeli (podzbiorów wyjściowego zbioru potencjalnych zmiennych objaśniających) istotnym wyzwaniem staje się przy tym, zwłaszcza dla danych wysokowymiarowych, strona obliczeniowo/implementacyjna. Dla danych ze zmiennymi jakościowymi opisano w literaturze w szczególności tzw. grupowe Lasso (Yuan, Lin 2006), które włącza lub odrzuca całe zmienne jakościowe bez łączenia ich poziomów oraz algorytm CAS-ANOVA (Bondell, Reich 2009) i algorytm z pracy Oelkera i in. (2014), które także łączą poziomy zmiennych jakościowych przy braku istotnych różnic ich efektów. Wszystkie wspomniane algorytmy oparte są na idei Lasso, wykorzystującej odpowiednio zdefiniowane funkcje kary oparte na normie L_1. Wiadomo, że metody oparte na idei Lasso mają

w skończonych próbach tendencję do wybierania modeli o zbyt wysokim wymiarze. Badania opisane w rozprawie miały na celu konstrukcję algorytmów, które zachowując asymptotyczną własność wyroczni nie miałyby tej wady w próbach skończonych.

Rozprawa składa się z siedmiu rozdziałow na 95 stronach i napisana jest w języku angielskim. Pierwszy rozdział, to wprowadzenie do problemu, przegląd wybranej literatury przedmiotu i sformułowanie problemu badawczego. W rozdziale drugim wprowadzony jest formalizm potrzebny do precyzyjnego opisu i parametryzacji rozpatrywanych modeli i krótko omówiona jest idea metod opartych na Lasso. W rozdziale trzecim zdefiniowany jest algorytm DMR4lm, rozwiązujący problem wyboru modelu liniowego gdy $p < n$ i podane są pewne warunki, przy których ma on asymptotyczną własność wyroczni. Zaproponowany jest też algorytm DMRnet4lm dla przypadku gdy $p > n$, ale bez teoretycznej analizy jego własności. W skrócie polega on na wstępnym przesiewaniu listy p zmiennych przy pomocy grupowego Lasso, co redukuje listę zmiennych do liczności mniejszej od n, po którym stosowany jest algorytm DMR4lm. Powstaje w ten sposób lista modeli-kandydatów o różnych wymiarach, z której wybiera się ostateczny model stosując kryterium najlepszej predykcji na poziomie grup modeli o tym samym wymiarze, a kryterium informacyjne lub kroswalidację na poziomie grupy modeli o różnych wymiarach. W rozdziale 4 omawiane są wersje tych algorytmów dla uogólnionych modeli liniowych. Analiza teoretyczna ogranicza się tu do wykazania zgodności algorytmu DMR4glm. Rozdziały 5 i 6 poświęcone są opisowi eksperymentów numerycznych, odpowiednio dla modeli liniowych i modeli logistycznych, w których porównano algorytmy zaproponowane w rozprawie z algorytmami znanymi z literatury. Rozpatrzono przy tym zarówno po trzy zbiory danych rzeczywistych, jak i dane symulowane, podobne do używanych wcześniej w literaturze. Przy pomocy 10-krotnej kroswalidacji (dla danych rzeczywistych) lub dodatkowo wygenerowanej grupy kontrolnej (dla danych symulowanych) wyznaczano dla każdej metody średni wymiar wybranego modelu i średni błąd prognozy, które były używane do porównania różnych metod. W ostatnim rozdziale podsumowano wyniki przeprowadzonych badań.

Część wyników opisanych w rozprawie, a dotyczących algorytmu DMR4lm, a także idea algorytmu DMR4glm_wald została opublikowana wspólnie z A. Maj-Kańską i P. Pokarowskim w roku 2015 w wydawanym przez IMS (Institute of Mathematical Statistics) bardzo dobrym czasopiśmie *Electronic Journal of Statistics*.

Tematyka rozprawy dotyczy ważnych problemów statystyki matematycznej i obliczeniowej, z ważnymi zastosowaniami praktycznymi. Jak widać choćby z cytowanej literatury, jest to żywy nurt w światowej literaturze statystycznej. Rozprawa realizuje dobrze zaplanowany projekt badawczy w ważnej i nowoczesnej tematyce.

2. Główne wyniki

Z punktu widzenia możliwych zastosowań, konstrukcja wszystkich algorytmów (DMR4lm, DMR4glm, DMR4gml_wald i ich wersje z wstępnym przesiewaniem dla przypadku $p > n$) i ich efektywna implemetacja jest dokonaniem ważnym, a wstępna analiza ich własności w eksperymentach symulacyjnych potwierdziła, przynajmniej w przypadku DMR4lm i DMR4glm, osiągnięcie postawionego celu badawczego, jakim była konstrukcja metod mających asymptotyczną własność wyroczni, ale wybierających modele o wymiarze średnio niższym niż Lasso.

Najważniejsze teoretyczne wyniki rozprawy to dowody zgodności procedury DMR4lm i asymptotycznej normalności estymatorów w wybranym modelu (Th. 1 i Corollary 1 oraz Th. 2 w rozdziale 3.3) i dowód zgodności procedury DMR4glm (Th. 3 w rozdz. 4.6).

3. Dyskusja i ocena rozprawy

Uzyskane wyniki są w mojej ocenie formalnie poprawne i ważne z punktu widzenia praktycznych zastosowań rozważanych metod. Przedstawione dowody świadczą o sprawności Autorki w posługiwaniu się się technikami analitycznymi i probabilistycznymi, a zaproponowane metody i ich implementacja pokazują sprawność Autorki w myśleniu algorytmicznym. Trochę trudno czytało się niektóre fragmenty dowodów, bo Autorka nie zawsze jasno przedstawia ich idee. Sądzę też, o ile czegoś nie przeoczyłem, że można je nieco uprościć, o czym bardziej szczegółowo piszę w dalszej części recenzji.

Brakuje mi w rozprawie dyskusji założeń, przy których udowodniono własności asymptotyczne algorytmów. Czy można rozsądnie oczekiwać ich spełnienia w typowych sytuacjach, np. w rozważanych przykładach numerycznych? Czy da się je wyrazić w terminach macierzy planu $\mathbf{X}$ i wektora parametrów?

Pewnym niedostatkiem ocenianej pracy jest też zupełne zignorowanie zagadnień związanych z możliwą hierarchicznością modelu i jej implikacjami dla procedury wyboru modelu. Hierarchiczność w modelu ze zmiennymi jakościowymi pojawia się naturalnie, gdy oprócz efektów głównych tych zmiennych w modelu występują też ich interakcje. Połączenie dwóch poziomów zmiennej jakościowej jest metodologicznie dopuszczalne tylko wtedy, gdy nie tylko wartości efektów głównych tych poziomów nie różnią się istotnie, ale nie ma też istotnych różnic między żadnymi interakcjami, w których te poziomy występują. Wcześniejsze algorytmy, np. CAS-ANOVA, zakładały, że w modelu występują tylko efekty główne, ignorując specjalną strukturę modeli z interakcjami. Wydaje się, że aby uniknąć trudności interpretacyjnych dla otrzymanych modeli, założenie to powinno być jawnie sformułowane także dla opisanego w pracy algorytmu DMR. Tymczasem, w rozdziale 6.1.3 ocenianej pracy analizowany jest znany z literatury dotyczącej grupowego Lasso przykład modelu logitowego dla prawdopodobieństwa wystąpienia tzw. „donor sites" w łańcuchu DNA, w którym występują też interakcje. W mojej ocenie, stosowanie DMR do tych danych jest metodologicznie wątpliwe. Z prezentowanych w rozprawie wyników dla tego modelu nie można niestety odczytać, jak często następowało w wybieranych modelach łamanie hierarchicznej struktury modelu. Oczywiście, asymptotycznie metoda działa i, jak udowodniono, wybiera z prawdopodobieństwem jeden właściwy model, ale w skończonych próbach łatwo wpaść w kłopoty interpretacyjne. Wątpliwości takie miałem także w przykładzie z rozdz. 5.1.4 gdzie, tak jak to rozumiem, mamy do czynienia z modelem zagnieżdżonym (działki zagnieżdżone w lokalizacjach), a efekt działki bardziej naturalnie byłoby modelować jako efekt losowy, co dałoby model mieszany.

Należy podkreślić, że zagadnienie uwzględnienia hierarchicznej struktury modelu liniowego w procesie wyboru zmiennych było dyskutowane w literaturze. W zagadnieniach regresji z ciągłymi predyktorami, hierarchiczna struktura modelu uwzględniana jest przez tzw. „zasadę dziedziczenia", która wymaga, aby człony algebraicznie wyższego rzędu mogły wystąpić w modelu tylko wtedy, gdy występują w nim wszystkie odpo-

wiednie człony rzędów niższych. Uwzględnienie tej zasady w procedurach doboru zmiennych opisano np. w pracach Chipmana (1996), Yuana i in. (2007, 2009) i Choia i in. (2010) dla procedur bayesowskich, Lasso i LARS. Dla problemu wyboru i grupowania poziomów czynników w modelach ANOVA, Post i Bondell (2013) zaproponowali algorytm GASH-ANOVA (od Grouping And Selection using Heredity in ANOVA) i pokazali, że uwzględnienie hierarchicznej struktury modelu może dać istotne korzyści w skończonych próbach. Trochę szkoda, że ten ważny aspekt został w pracy pominięty.

Pewne pytania nasuwają się także w związku z opisem algorytmów DMRnet4lm i DMRnet4glm. Wydaje się, że parametry $o, m, \lambda_1, \ldots, \lambda_m$ i maksymalny wymiar p^- analizowanych modeli były wybierane dość arbitralnie, prawdopodobnie na podstawie wyników obserwowanych w eksperymentach symulacyjnych. Nie znalazłem odpowiedzi na nasuwające się pytanie jak wyniki zależą od wartości tych parametrów. Druga wątpliwość dotyczy wstępnego przesiewania przy pomocy grupowego Lasso. Wiadomo, że wyniki Lasso zależą od przyjętej parametryzacji i dlatego w literaturze (np. Bühlmann, van de Geer 2011, rozdz. 4.3) są sugestie, aby stosować Lasso dopiero po wprowadzeniu w podprzestrzeniach odpowiadających efektom głównym baz ortonormalnych i związanej z nimi parametryzacji modelu, co zapewnia w szczególności niezależność wyników Lasso od rodzaju kontrastów użytych w wyjściowej parametryzacji modelu. Autorka pracuje w dyskutowanych algorytmach z oryginalnymi kontrastami prostymi. A gdyby tak przed grupowym Lasso wykonać opisane wyżej przeparametryzowanie? Dekompozycja QR i tak jest liczona w kroku DMR4lm. Czy i jak wpłynęłoby to na własności algorytmu? Oczywiście, jakieś decyzje trzeba było podjąć, ale ciekaw jestem opinii Autorki w tej sprawie.

4. Uwagi szczegółowe

Lemat 2 jest wykorzystywany w dowodzie lematu 8 do wykazania równości zdarzeń na dole str. 29. Do dowodu twierdzenia 1 wystarczy jednak górne oszacowanie dla $P(T \notin \mathcal{M})$, a więc w ostatnim wzorze na str. 29 wystarczy trywialna inkluzja $\subset$ zamiast ostatniej równości. To oznacza, że lemat 2 jest tu zbędny. Odwołanie do niego następuje wprawdzie jeszcze raz na str. 43 w dowodzie twierdzenia 3, ale moim zdaniem jest on zbędny także tam, a w efekcie nie jest potrzebny nigdzie. Dowód twierdzenia 3 jest opisany trochę myląco, co wynika z kilku niedokładności, zakładam, że redakcyjnych, których efekty się kumulują. Po pierwsze, w punkcie 3 opisu algorytmu DMR4glm jest napisane, że wektor $\mathbf{h}$ wartości statystyk testu ilorazu wiarogodności jest porządkowany rosnąco, a powinno być malejąco, przy przyjętej definicji statystyk LRT (małe wartości powodują odrzucanie hipotez). Po drugie, zbiór $\mathcal{E}_1$ jest zdefiniowany dwukrotnie na dwa różne sposoby na str. 40 i 42. Poprawna definicja ze str. 42 powinna pojawić się na str. 40 zaraz po „Let us denote", a równość $\{T \notin J\} = \mathcal{E}_1$ nie jest definicją, lecz jest dowodzona na str. 43. Fragment dowodu opisany na str. 43, w którym następuje odwołanie do lematu 2, jest opisany dość niejasno, ale moim zdaniem jest on zbędny, bo do dowodu twierdzenia 3 wystarczy zamiast równości tylko trywialna inkluzja $\{T \notin J\} \subset \mathcal{E}_1$, przy której mamy, jak na str. 41, $\{\hat{T} \neq T\} \subset \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_{3a}$, gdy $tr \leq ac\delta$ i na podstawie założenia (4.3) można przyjąć $a = 1/4$. Na str. 41 i 42 pokazano też, że $\mathcal{E}_1 \subset \mathcal{E}_{3\frac{1}{4}}$ i $\mathcal{E}_2 \cup \mathcal{E}_{3\frac{1}{4}} \subset \mathcal{C}(cr)$ i że $P(\mathcal{C}(cr)) \leq \sigma^2 p/cr$, co kończy dowód twierdzenia 3.

Rozprawa jest napisana zasadniczo poprawnym i starannym językiem i jest też w większości starannie zredagowana, choć są też niedokładności. Na przykład: „indexes" powinny w języku rozprawy naukowej przyjąć raczej formę „indices", IRLS jest rozwijane niepoprawnie albo jako „iterated reweighted least squares" (str. 6), albo jako „iteratively related least squares" (linia 1 na str. 36) albo poprawnie (dwie linie niżej), liczba ograniczeń to $p - q$ a nie q (10_{10}), zbędna transpozycja w (2.14), błędna zapowiedź prezentacji szczegółów „in the Appendix" (tak było w artykule, w którym ten fragment był opublikowany, ale nie w rozprawie) na str. 19, ustalenie rozmiaru modelu prawdziwego w 23^1 mimo zapowiadanego w 17^7 jego uzmiennienia, mała zamiast dużej δ w 23^{13}, brak minusa w ostatnim wzorze na str. 23, zbędne założenie $t \prec p$ w twierdzeniu 2 (i niemożliwe do spełnienia, skoro p jest skończone), pierwszy minus w 27^3 powinien być plusem, a drugi plus w 27_7 minusem (w 27_4 znak jest już poprawny), przestawione a i b w B_{1-x} w 28_5, „belong to" zamiast „are included in" w 41^5, nie ten Schwarz w 41^7 i brak zamykającego nawiasu w 42_3.

5. Konkluzja

Oceniana rozprawa doktorska mgr Agnieszki Prochenki jest oryginalnym i interesującym wkładem Autorki do nowoczesnej statystyki matematycznej i obliczeniowej. Stwierdzam, że rozprawa spełnia formalne, merytoryczne i zwyczajowe wymogi stawiane pracom doktorskim z zakresu informatyki w dziedzinie nauk matematycznych i wnoszę o dopuszczenie jej do publicznej obrony.

Zbigniew Szkutnik

Szkutnik

Kraków, 03.10.2016.

Literatura

Bondell HD, Reich BJ (2009) Simultaneous factor selection and collapsing levels in ANOVA, *Biometrics* **65**, 169-177.

Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer.

Chipman H (1996) Bayesian variable selection with related predictors, *The Canadian Journal of Statistics* **24**, 17-36.

Choi N, Li W, Zhu J (2010) Variable selection with the strong heredity constraint and its oracle property, *Journal of the American Statistical Association* **105**, 354-364.

Oelker M-R, Gertheiss J, Tutz G (2014) Regularization and model selection with categorical predictors and effect modifiers in generalized linear models, *Statistical Modelling* **14**, 157-177.

Post JB, Bondell HD (2013) Factor selection and structural identfication in the interaction ANOVA model, *Biometrics* **69**, 70-79.

Yuan M, Joseph V, Lin Y (2007) An efficient variable selection approach for analyzing designed experiments, *Technometrics* **49**, 430-439.

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society B* **68**, 49-67.

Yuan M, Joseph V, Zou H (2009) Structured variable selection and estimation, *The Annals of Applied Statistics* **3**, 1738-1757.