Recenzja dorobku dr Wojciecha Rejchela

Rozpocznę recenzję od uwag dotyczących sposobu recenzowania. Wytłumaczę dlaczego jest to ważne w tym przypadku, jakie są trudności. W cytowaniu prac dr Rejchela będę używał jego oznaczeń, czyli od [A1] do [A8]. Nie będę tutaj kopiował tych cytowań. Używam skrótu IRR, czyli irrepresentable conditions oraz oczywiście LASSO. Sformułowania z prac oraz autoreferatu zapisuję kursywą. Tytuły prac także są zapisane kursywą.

Metodologia oceny

Metodologię oceny tematyki w pewnym stopniu określa zdanie z autoreferatu, rozdział 4.1 Wprowadzenie *Moje badania naukowe skupione są na konstrukcji i badaniu metod potrafiących anlizować i interpretować złożone zbiory danych... jest aktualnie intensywnie badana w statystyce i uczeniu maszynowym...* Wybór dziedziny jako matematyki jest wobec tego zdania zawężający. Statystykę rozumianą przez statystyków najlepiej symbolizuje trójkąt z książki *Computer Age Statistical Inference Algorithms, Evidence, and Data Science* Bradley Efron, Trevor Hastie, 2016 w Cambridge University Press, strona 448. Trójkąt którego wierzchołki zdefiniowane są jako mathematics, applications, computation. Zapisałem nazwy angielskie dlatego, że applications w statystyce bliższe jest znaczeniu użyteczności niż pojęciu zastosowania. Widzimy zatem jasno, że osoba, która oczekuje sprawiedliwej oceny dorobku napotyka trudność formalną. W Polsce brak dziedziny Statystyka i nawet jeśli ustawodawca zawęża prawo do sprawiedliwej oceny należy dokonać szczególnie starannej recenzji.

To, że dr Rejchel odwołuje się do takiej oceny wynika z punktu 4.1.3. Osiągnięcia. W punkcie 1 pisze o *efektywności obliczeniowej, o procedurach, o konstruktywnych algorytmach.* Zatem odwołuje się do użyteczności i sprawności obliczeniowej. Punkt 2 to ciągle opis osiągnięć opisywanych z punktu widzenia procedur (*nowe, skuteczne*). Oczywiście są sformułowania nawiązujące do matematyki *Badamy ich statystyczne własności.*

Sam dr Rejchel nakierowuje recenzenta na ten sposób recenzji punkt 4.1.3. Osiągnięcia punkt 2 *nasze wyniki pomagają lepiej zrozumieć.* To jest niewątpliwie perspektywa statystyka. *Wyniki pomagają* a nie są niezbędnym składnikiem. Aby zrozumieć ten punkt widzenia wystarczy zajrzeć do prac genialnego Leo Breimana, *Random Forests* albo Breiman, L. (1998a), *Arcing classifiers* (discussion paper). Annals of Statistics, 26, 801- 824. W zasadzie brak matematycznych uzasadnień. Uzasadnienia pojawiły się 15 lat później, Scornet, Erwan; Biau, Gérard; Vert, Jean-Philippe *Consistency of random forests.* Ann. Statist. 43 (2015), no. 4, 1716-1741.

Z punktu widzenia Statystyki metodologia oceny właściwa dla recenzenta oraz dla kompetencji Rady jest zawężająca. Ponieważ każdy ma prawo do rzetelnej oceny postaram się ją nieco rozszerzyć.

Modele niskowymiarowe i warunek IRR

Rozpocznę od wyników tzw. modeli niskowymiarowych, np. pozycje [Twierdzenie 3.1 A8] i [Twierdzenie 4.1, A6]. Moją jedyną krytyką całości dorobku jest

umieszczenie tych prac do oceny. Pozostałe wyniki zupełnie wystarczają mi do oceny. Sam autor nie kryje, że wyniki podobne były dowodzone w różnych modelach str 21 i 22 autoreferatu. Przyjrzyjmy się już dokładnie analizując dowody w [A8]. Użyta jest analiza dla funkcji wypukłych. Autor zauważa *Notice that convexity of the loss function and methods from the convex empirical process theory (Pollard, 1991; Niemiro, 1992, 1993; Geyer, 1996) play crucial roles in our argumentation.* Z drugiej strony Twierdzenie 2.2[A8] to w zasadzie Lemat 2 Zhou czyli konsekwencja użycia wyniku Knight oraz Fu (2000) zaś Twierdzenie 3.1 to modyfikacja Twierdzenia 2 [44] Zou, H. (2006), *The adaptive lasso and its oracle properties*, J. Amer. Statist. Assoc.

Korzyść dla recenzenta w czytaniu tych prac polega na odkrywaniu źródeł motywacji dla kolejnych prac, np. wpływ promotora prof. Niemiro. Choć najważniejsza wydaje mi się refleksja nad warunkiem IRR, który pojawia się w omawianych tutaj wcześniejszych pracach [A8] i [A6]. W sformułowaniach twierdzeń, dokładnie w założeniach, Twierdzenia 1,2,3,4,6,7 występuje współczynnik odwrotności F_∞. Jego wprowadzenie w autoreferacie jest zestawione z IRR, który w zasadzie kompletnie jest identyfikowalny poprzez klasyczne twierdzenia typu Wiosek 2.4 [A8]. Jeśli dołożymy do F_∞ jeszcze κ_a compatibility factor modyfikację współczynnika zgodności który występuje w Twierdzeniu 8(Twierdzenie 4 [A1]) to uzyskamy komplet tych *wartości (które) określają jak bardzo prawdziwy model T różni się od pozostałych* str 7 autoreferat. Co więcej, autor pokazuje przykład gdzie IRR nie zachodzi na potwierdzenie słuszności wprowadzonych miar F_∞ i κ_a.

Wniosek: Matematyka użyta w pracach jest rzetelna.

Modele wysokowymiarowe

Wyniki pogrupuję według matematycznych (probabilistycznych) metod. Rozpoczniemy od Twierdzeń 1,2,3,6,8, określiłbym je jako twierdzenia o koncentracji. Obejmuje on prace [A1], [A2], [A4], [A5]. Wyjaśnię skąd nazwa dla tej grupy prac oraz dokonam podsumowania z punktu widzenia matematyki (probabilistyki). Następnie przejdę do procedur statystycznych, których jakość została umotywowana powyższymi twierdzeniami. Na koniec przedstawię pozostałe wyniki prac [A3] i [A7].

1 Twierdzenia o koncentracji

Postaramy się poszukać różnych sformułowań wyników dotyczących koncentracji. Rozpoczynamy od dowódu Twierdzeń 1,2 (Twierdzenia 1,2 [A1]) W [A1] możemy zidentyfikować miejsce cytowania tej metody i przeczytać *This part of the paper is devoted to exponential inequalities for subgaussian random vectors. They are interesting by themselves and can be used in different problems than we consider. In the current paper they are main probabilistic tools that are needed to prove Theorems 1-3. Specifically, in Lemma 2 (iii) we generalize the Wallace inequality for χ^2 distribution Wallace (1959) to the subgaussian case using the inequality for the moment generating function in Lemma 2 (ii). The last inequality is proved by the decoupling technique as in the proof of theorem 2.1 in Hsu et al. (2012).)* Przy okazji Twierdzenie 5 z autoreferatu to krok dowodowy przy głównym założeniu Assumption 2. Komentarz do tego założenia podam w

podrozdziale Procedury i ich użyteczność - RankLasso. Dowód Twierdzenia 5 jest oparty o własności warunkowej wartości oczekiwanej.

W dowodzie Twierdzenia 3 (Twierdzenie 3.1 [A4]) podobnie wykorzystywane są nierówności dla zmiennych subgausowskich, *if one uses Markov's inequality and the fact that* $\varepsilon_1, \ldots, \varepsilon_n$ *n are independent and sub-Gaussian.*

Dr Rejchel formułuje Twierdzenie 6, które jest połączeniem Twierdzeń 2 i 5 z [A2]. W zasadzie trudno bezpośrednio dopatrzeć się twierdzenia o koncentracji. Jednak już sformułowania Lematów 13, 14 ,15 nie pozostawiają wątpliwości. Cytat z [A2] *To prove Theorem 2 we need three auxiliary results: Lemma 13, Lemma 14 and Lemma 15. The first one is borrowed from van de Geer (2016, Corollary 8.2), while the second one is its adaptation to U -statistics.* A teraz cytat wstępu do książki van de Geer *This chapter presents probability inequalities for the (dual) norm of a Gaussian vector in* R^p *. For Gaussian vectors there are ready-to-use concentration inequalities (e.g. Borell, 1975).*

Jako ostatnie do tej grupy można zakwalifikować Twierdzenie 8 ([A1] Twierdzenie 4). Tu zacytujmy już bezpośrednio z pracy *Remark 5. The important assumptions of Theorem 4 are conditions (26) and (27). They can be proved using tools from the empirical process theory such that concentration inequalities (Massart, 2000), the Symmetrization Lemma (van der Vaart & Wellner, 1996, lemma 2.3.1) and the Contraction Lemma (Ledoux & Talagrand, 1991, theorem 4.12).*

W dowodzie Twierdzenia 4 (Twierdzenie 2 i Wniosek 3 [A5]) autorzy korzystają z analogów twierdzeń o koncentacji, *In particular, we frequently use the following Hoeffding's inequality for Markov chains (Miasojedow, 2014, Theorem 1.1).* Wprawdzie autor pisze, że *łatwo to porównać z Twierdzeniem 1*, niemniej uprzejmie nie zgadzam się z dr Rejchelem. Rozumiem, że istnieją analogi twierdzeń w sytuacji łańcuchów Markowa ale nie jest to łatwe samo w sobie.

Wniosek: sam zestaw użytych twierdzeń o koncentracji obrazuje bardzo dobre obycie z tematyką i ciekawe zastosowanie statystyczne.

2 Procedury i ich użyteczność

Procedury umieściłbym w środku trójkąta o którym wspominałem w rozdziale Metodologia oceny. W zasadzie dorobek można by próbować oceniać wyłącznie z tej perspektywy, zatem ocena twierdzeń i ich dowodów jest oceną 1/3 wysiłku. Przy okazji warto przypomnieć, że jedno z najsłynniejszym twierdzeń statystycznych (1995) Twierdzenie Benjamini Hochberga nosi nazwę procedury. Oczywiście czytając wyniki i procedury identyfikujące prawdziwy model w pracach dr Rejchela natychmiast przychodzi na myśl to genialne i w zasadzie "dowodowo" nietrudne twierdzenie. Użyte narzędzia w pracach dr Rejchela są zdecydowanie trudniejsze. Procedury utożsamiam nieco na wyrost z algorytmami. Sama procedura - algorytm jeśli jest „samo wyjaśniający" jest jednocześnie użyteczny i staje się początkiem do jej badania z punktu widzenia numerycznego, wówczas można mówić np. o jego efektywności. O użyteczności mówię również w węższym sensie, czyli do jakich modeli ma zastosowanie.

RankLasso [A2]

Procedura RankLasso [A2] jest matematycznie sprytnie skonstruowana. Twierdzenie 1 w [A2] pokazuje na czym polega pomysł. Zamiast badać „dane surowe"

dane porządkujemy i na nich dokonujemy analizy. Niemniej wydaje mi się, że ma ona ograniczone znaczenie (użyteczność). Przyczyną tego jest założenie (Assumption 2). Jest ono nieco słabsze niż założenie, że predyktory mają rozkład eliptyczny. Rozkład eliptyczny dziedziczy pewne wady rozkładu gaussowskiego, który jest jednym z przykładów rozkładu eliptycznego.

Wniosek: Pomimo ładnych matematycznych własności rozkładu eliptycznego to rodziny kopuł d-vine, c-vine są bardziej użyteczne np. w badaniu ryzyka. Nie jest jednak wykluczone, że jednak się mylę co do przyszłości. Praca bardzo dobrze zredagowana.

Screening - Selection procedure [A1]

Wydaje mi się, że ciekawszą procedurą jest SS procedure, czyli Screening - Selection procedure. Ma ona zastosowanie (użyteczność) dla modeli GLM. Wygląda jak z dokumentacji R czy SAS i chyba bardziej jest w nurcie współczesnych propozycji statystycznych "mieszających" różnorakie nurty. Zatem w pierwszym kroku mamy Lasso w drugim tworzymy rodzinę zagnieżdżoną i *wybieramy model minimalizujący uogólnione kryterium informacyjne* tzw. GIC. Jak piszą autorzy artykułu [A1]*A generic combination of the penalized log-likelihood (as TL or FCP) with GIC is considered in* Fan, Y., & Tang, C. Y. (2013). *Tuning parameter selection in high dimensional penalized likelihood.* Journal of the Royal Statistical Society Series B (Statistical Methodology), 75, 531-552. Wielka szkoda, że nie jest to bezpośrednio pomysł autorów. Na pociesznie Algorytm 1 w pracy [A1] jest na pewno „książkowy". Po drugie Fan & Tang nie mają twierdzenia o koncentracji w tak klasycznej formie i zupełnie nie wiem dlaczego ten aspekt nie został podkreślony. Z punktu probabilistyki to bardzo dobry wynik. Autorzy [A1] pokazują, że procedura Screening - Selection *jest elastyczna i może być skutecznie używana również poza GLM.*

Przy okazji tej procedury podejmijmy problem współautorstwa. Pochwalam wspólny wysiłek czołówki statystyków... co w zasadzie wykluczyło ich udział w ocenie dorobku. Otóż tematyka Lasso przyciągnęła matematyków z różnych dziedzin np. z z analizy harmonicznej. Wystarczy porównać bibliografię pracy *Screening for Sparse Online Learning* J. Liang, C. Poon arXiv 2021 dotyczącego algorytmu Online Screening LASSO. W zasadzie bibliografia jest rozłączna z bibliografią z pozycji [A1] co pokazuje, że jest potrzeba dużych komunikujących się ze sobą zespołów.

Wniosek: Podsumowując procedura ma wszystkie cechy aby być wysoko ocenioną z punktu statystyki.

Progowe Lasso, TL procedure [A1], [A5]

Procedura TL z pracy [A1] i [A5] jest w nurcie wielu wyników, zatem nie będę jej szczegółowo opisywał. Stosuje się ją od wielu lat zarówno do estymacji jak i aproksymacji. Tak jak jest ona przedstawiona w [A1] jest ona niekonstruktywna... choć może wykorzystując bootstrap albo cv dałoby się procedurę empirycznie poprawić. Użyteczność w [A1], to klasa GLM. W [A5] to nie jest wyłącznie procedura progowa. Złożoność tej procedury jest dziedziczona z samego opisu modelu Isinga, *we consider a model selection problem in high-dimensional binary Markov random fields.*

Procedura Lasso SD [A4]

Użyteczność tej metody sprowadza się do *the high-dimensional linear model with sub-Gaussian errors.* Autorzy twierdzą, że *We propose a new model selection algorithm (called 'LassoSD'), which combines a step-down multiple testing approach with penalized minimization.* Co ciekawe porównują go do *the algorithm with the threshold (7) mimics the Bonferroni correction, while thresholds in (8) are analogues of the Holm method. Therefore, in further parts of the paper we call them the Bonferroni and Holm procedure, respectively.* Stąd naturalne pytanie o procedurę Benjamini Hochberga. Wydaje mi się, że takie *indywidualne* podejście do predyktorów procedury Lasso SD pomoże rozwiązać otwarty problem "separowalności" modeli. Ponadto, twierdzenie jest w sensie ścisłym procedurą, gdyż algorytm jest konstruktywany. *Theorem 3.1 suggests how parameters λ, δ and α of the algorithm should be chosen and confirms that it is a constructive procedure.*

Wniosek: Lasso SD może być „produktem eksportowym" dlatego omówię dodatkowo tę procedurę z punktu widzenia efektywności, czyli kompletnie z punktu widzenia statystyki.

3 Efektywność Lasso SD

Skupimy się w zasadzie wyłącznie nad wynikami symulacji Lasso SD. Dr Rejchel pisze *pracowaliśmy z modelami liniowymi.* To zdanie wyznacza naturalną linię „obrony" algorytmu i opartego na nim programu. Sprawdzamy go w sytuacji, gdy tester zna odpowiedź skąd wziął dane. Autorzy [A4] porównują swoją procedurę z innymi na „rynku" wychodząc od sytuacji opisanej w rozdziale 4.1. Simulated data set [A4], czyli z p wielowymiarowym rozkładem normalnym plus błąd normalny $N(0, \sigma^2)$. Dobieranych jest pięć modeli p duże $(2000, 3000)$ ale istotne są tylko pierwsze $3 \leqslant p_0 \leqslant 10$ współrzędne. Współczynniki regresji najczęściej są równe 2.

Autorzy dokładnie opisują sposób uzyskanych wyników. Program ma trzy kroki W pierwszym kroku *to calculate the Lasso we use the package 'glmnet' [Friedman J, Hastie T, Tibshirani R.] in the 'R' software [R Core Team. R].* I to jest miejsce które zawsze mnie niepokoi. Pakiety mają swoje ograniczenia. Studenci wielokrotnie znajdywali luki... także w programie komercyjnym SAS. Następne kroki programu muszą być napisane już przez autorów. Tutaj użyteczność procedury jest ważna, po to żeby zapewnić czytelność programu. Przy okazji inne procedury używają innych pakietów. Gdy zaakceptujemy te trudności przechodzimy do samych badań.

Badana jest efektywność pięciu procedur (B, H, TL, SOS,MS), Lasso SD ma dwie wersje B i H. Możemy zauważyć za autorami *In Table 1(model 1), we see that Lasso SD with both thresholds (B and H) as well as SOS perform very well in model selection.* Przy okazji Lasso SD (B i H) jest najlepsze. Nie będę szczegółowo opisywać wyników. Tabel jest pięć i jak często w statystyce każda metoda ma swój obszar zastosowań. Zatem to cały czas jest pole do zagospodarowania aby porozdzielać obszary zastosowań poszczególnych procedur. Pierwsza próba jest zrobiona dla Lasso SD (H). *The remedy for it can be the Holm procedure, that is also based on testing after the Lasso, but uses smaller and decreasing thresholds.*

Analizowany przykład danych realnych *4.2 Real data set.* to *diabetes da-*

ta set from the 'care' R package. Predyktorów jest 10 ale bierzemy interakcje i uzyskując 64 predyktory. Tutaj porównujemy procedurę z klasyczną metodą najmniejszych kwadratów. W zasadzie to dwa parametry pokazują jakość procedur, błąd *RE* oraz wymiar *dimension*. Znowu sama tabelka jest dość jasna ale dla mnie problemem jest sam wybór danych. Owszem procedura jest stosowalna w modelach liniowych ale może same dane nie są liniowe.

Wniosek: Mam przekonanie, że warto pogłębić w różnych kierunkach ten najlepszy produkt.

Problem błędnej specyfikacji modelu

W pracy [A3] opisany jest problem błędnej specyfikacji modelu klasyfikującego, binarnego (X, Y), $Y \in \{-1, 1\}$. Najlepszym klasyfikatorem jest klasyfikator bayesowski f_B oparty o prawdopodobieństwo warunkowe $\eta(x) = P(Y = 1|X = x)$. Efektywność metody klasyfikacyjnej f to porównanie f z f_B. Rozważane klasyfikatory $f = f_\theta$ są liniowe z kwadratową funkcją straty oraz funkcją straty charakterystyczną dla regresji logistycznej. W autoreferacie można przeczytać *Te dwa podejścia dają nadspodziewanie dobre rezultaty*, zob. Tabela 1 i Tabela 2 w [A3]. W pracy [A3] badana jest jednak funkcja η, która ma strukturę liniową

$$\eta(x) = g(\beta_0 + \sum_{j=1}^{p} \beta_j x_j),$$

gdzie g jest nieznana. Chyba wynik nie jest aż tak niespodziewany. Praca [A3] nawiązuje do wyników Zhang, T. *Statistical behavior and consistency of classification methods based on convex risk minimization.* Ann. Stat. 2004, 32, 56-85. Wprawdzie autorzy piszą *In this paper we do not follow this way and work* niemniej trzeba przeczytać pracę aby zrozumieć *In particular, (16) follows from ([Zhang], Theorem 2.1) applied for* f_b^{quad} *and Example 3.1.* Dowód Twierdzenia 2 nie jest wyrafinowany nie licząc wkładu Zhanga.

Wniosek: Opis badań jest równie długi jak same uzasadnienia.

U statystyki

To co łączy tą pracę z pozostałymi wynikami to widoczny wcześniejszy etap w rozwoju naukowym. Data wysłania 2015 do Neurocomputing z IF 5.7, cytowanych jest 43 pozycji. Dr Rejchel do zagadnień regresji porządkowej wprawdzie także używa twierdzeń o koncentracji dla U statystyk ale uznałem, że lepiej jego wyniki opisać oddzielnie gdyż dotyczą jednak odmiennej tematyki. Jeden z kroków dowodowych: *We start with Symmetrization Lemma [A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, Lemma 2.3.1]* To piękna klasyczna teoria, trochę brakuje mi dystansu do Twierdzenia 3 w [A3]. Postaci twierdzeń o koncentracji jest jednak dużo, np. Twierdzenia Talagranda w wersji Bousqueta.

Wniosek: Jeszcze raz podkreślę wniosek: sam zestaw użytych twierdzeń o koncentracji pokazuje bardzo dobre obycie z tematyką i ciekawe zastosowanie probabilistyczne. Dr Rejchel w swoim autoreferacie poddał krytyce procedurę w oparciu o U-statystyki mającą umocowanie w twierdzeniach. Podsumowuje on w

następujący sposób *mają one ograniczoną przydatność poza regresją porządkową.* W zasadzie ta autorefleksja pokazuje dojrzałość dr Rejchela, jego rozumienie i nadzieję dla rozwoju statystyki.

Wniosek dotyczący dorobku

Dorobek od strony matematycznej wraz z perspektywą statystyki w kierunku użyteczności i efektywności spełniają zestaw zobowiązań zwyczajowych i ustawowych do uzyskania habilitacji.

Miscellaneous

Pozostałe informacje wyczerpują zestaw zobowiązań zwyczajowych i ustawowych do uzyskania habilitacji. Potwierdzam że zapoznałem się z:

1. Pozostałe osiągnięcia naukowo-badawcze
2. Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni¡ instytucji naukowej lub instytucji kultury, w szczególności
3. Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę
4. Pozostałe osiągnięcia naukowo-badawcze i inna działalność
5. Dokumentem: Wykaz osiągnięć naukowych albo artystycznych, stanowiących znaczny wkład w rozwój określonej dyscypliny, z punktem III Informacje Naukometryczne

Wniosek końcowy

Biorąc pod wzgląd zestaw zobowiązań zwyczajowych i ustawowych do uzyskania habilitacji uznaję, że dr Rejchel wszystkie je wypełnił.

Dziedziul

Karol Dziedziul

16 maj 2022 Gdańsk