prof. dr hab. Zbigniew Szkutnik
Wydział Matematyki Stosowanej AGH
al. Mickiewicza 30, 30-059 Kraków
e-mail: szkutnik@agh.edu.pl

R e c e n z j a
rozprawy habilitacyjnej i dorobku naukowego doktora Wojciecha Rejchela

Dr Wojciech Rejchel uzyskał magisterium (2005) i doktorat (2011) z matematyki na Wydziale Matematyki i Informatyki Uniwersytetu Mikołaja Kopernika w Toruniu. Zarówno tematyka pracy magisterskiej jak i doktorskiej dotyczyły problemów statystyki matematycznej. W szczególności, rozważany w pracy doktorskiej (przygotowanej pod opieką prof. Niemiry) tzw. problem regresji rangowej jest też ważnym elementem przedstawianej rozprawy habilitacyjnej.

Omówienie rozprawy habilitacyjnej

Rozprawa habilitacyjna dr. Rejchela „*M-estymatory z karą w wyborze modelu*" składa się z ośmiu powiązanych tematycznie prac [A1-A8] (używam numeracji z autoreferatu) opublikowanych w latach 2016-2021 w bardzo dobrych i dobrych czasopismach międzynarodowych, głównie z dziedziny statystyki matematycznej i uczenia maszynowego, o obecnej punktacji MEiN od 40 (1 praca) do 140 (4 prace). Trzy z tych prac są samodzielne. Z przedstawionych oświadczeń dr. Rejchela i współautorów pozostałych prac (P. Pokarowski, J. Mielniczuk, M. Bogdan, K. Furmańczyk i B. Miasojedow) wynika, że udział dr. Rejchela w ich powstaniu, w szczególności w dowodach wyników teoretycznych, był we wszystkich przypadkach istotny, a w większości przypadków wiodący.

Wspólnym elementem wszystkich prac jest, zgodnie z tytułem rozprawy, wykorzystywanie formalizmu M-estymatorów z funkcją kary. Autor wykorzystuje ten formalizm do rozwiązywania problemów wyboru modelu i jego estymacji, zarówno w klasach regularnych uogólnionych modeli liniowych (GLM), jak i w mniej regularnych klasach modeli semiparametrycznych, jak np. model pojedynczego indeksu (ang. *single index model*) czy też model regresji rangowej. Rozpatrywane przy tym nietrywialne i ważne z punktu widzenia zastosowań problemy to konstrukcja i efektywna implementacja algorytmów i badanie własności teoretycznych otrzymywanych tak estymatorów. W szczególności bada się zdolność algorytmów do wyboru zmiennych istotnych (tu celem jest uzyskanie tzw. zgodności selekcyjnej przy realistycznych założeniach) oraz własności predykcyjne wybranych i wyestymowanych modeli. Szczególnie ważny z punktu widzenia zastosowań (np. w genetyce) i intensywnie badany w ostatnich latach w czołowych ośrodkach na świecie jest przy tym przypadek tzw. rzadkich modeli wysokowymiarowych, w którym liczba p potencjalnych zmiennych objaśniających jest

istotnie większa od dostępnej liczby obserwowanych przypadków n, ale liczba t zmiennych istotnych jest mniejsza od n. Najnowsze wyniki dr. Rejchela mieszczą się w tym żywym i ważnym nurcie badań.

Do klasy M-estymatorów należą w szczególności estymatory typu LASSO, w których funkcja kary jest proporcjonalna do normy ℓ_1 estymowanego wektora. Wiadomo, że LASSO dobrze radzi sobie z przesiewaniem zbioru zmiennych objaśniających, ale zwykle zostawia ich zbyt dużo, a zgodność selekcyjna wymaga mocnych i trudno weryfikowalnych założeń. Zaproponowano więc w literaturze wiele modyfikacji procedury LASSO, które miały to poprawić. W tym nurcie mieszczą się prace [A4] i [A1].

We wspólnej z Furmańczykiem pracy [A4] zaproponowano algorytm LassoSD, w którym po przesianiu zmiennych przez LASSO wybiera się ostateczny podzbiór zmiennych stosując metody testowania wielokrotnego - idea pierwotnie zaproponowana przez Buneę i in. (2006). W [A4] zastosowano ją do przypadku wysokowymiarowego modelu regresji liniowej z planem stałym i z błędem subgaussowskim, wykazując zgodność selekcyjną przy założeniach słabszych niż wymagane przez samo LASSO i możliwość jawnego doboru parametrów algorytmu. Tę ostatnią cechę autorzy nazywają *konstruktywnością*.

W wieloautorskiej pracy [A1], rozwijając i uogólniając na przypadek modeli GLM i tzw. kontrastów wypukłych (obejmujący np. maszyny wektorów podpierających SVM używane w uczeniu maszynowym) wcześniejsze wyniki Pokarowskiego i Mielniczuka (2015), zaproponowano nowy algorytm SS wyboru i estymacji modelu, w którym LASSO jest wykorzystywane do wstępnego przesiewania i uporządkowania zmiennych, a ostateczny wybór modelu z zagnieżdżonej rodziny oparty jest na uogólnionym kryterium informacyjnym. Wynik jest interesujący, bo otrzymany algorytm jest selekcyjnie zgodny przy stosunkowo słabych założeniach i dla szerokiej klasy modeli, a dodatkowo w przypadku subgaussowskich modeli liniowych jest konstruktywny. Ze względu na ogólność otrzymanych wyników, ta praca wydaje mi się najbardziej obiecująca z punktu widzenia możliwości ich dalszych zastosowań.

We wspólnej z Miasojedowem pracy [A5] zastosowano LASSO do analizy modelu Isinga - szczególnego modelu grafowego do opisu i analizy zależności między wieloma zmiennymi binarnymi. Szczególną cechą tego modelu jest trudność w wyznaczeniu stałej normującej, co wymusza stosowanie różnych aproksymacji funkcji wiarogodności. W [A5] zastosowano aproksymację opartą na metodzie Monte Carlo łańcuchów Markowa (MCMC). Oryginalną cechą pracy [A5] jest szczegółowa ilościowa analiza zgodności selekcyjnej zaproponowanego algorytmu opartego na łącznym zastosowaniu LASSO i MCMC w rzadkich modelach Isinga. Chodzi przy tym o identyfikację par zmiennych warunkowo zależnych, które są reprezentowane przez krawędzie grafu. Zgodność selekcyjną udowodniono w [A5] przy założeniach nieco słabszych od przyjmowanych dla innych algorytmów opisanych w literaturze.

We wspólnej z M. Bogdan pracy [A2] rozważano tzw. model pojedynczego indeksu z nieznaną funkcją wiążącą i nieznanym rozkładem błędu, do którego zastosowano metodę LASSO z funkcją straty opartą na rangach. Wczesna wersja tej podstawowej idei pojawiła się w pracy Zhu i Zhu (2009) i była badana w innych pracach przy różnych restrykcyjnych założeniach. W [A2] zaproponowano jej wersję progową (niezerowe współczynniki z LASSO są zerowane, gdy ich moduły są zbyt małe) i ważoną

(kara ℓ_1 zastąpiona jest ważoną karą ℓ_1 z adaptacyjnym wyborem wag.) Przy bardzo istotnie osłabionych założeniach wykazano zgodność selekcyjną metody progowej i podano oszacowania błędu estymacji metody ważonej. Warto podkreślić brak założeń o funkcji wiążącej (poza monotonicznością) i o rozkładzie błędu. Ta metodologia działa jednak tylko dla losowych zmiennych objaśniających, bo tylko wtedy udaje się pokazać odpowiedni związek wektora estymowanego przez RankLASSO i prawdziwego wektora współczynników w predyktorze liniowym.

Własności LASSO w pewnym szczególnym modelu semiparametrycznym w zagadnieniu klasyfikacji binarnej były badane w pracy [A3] (wspólna z Furmańczykiem) pod kątem estymacji i selekcji zmiennych. Uzyskano tam w szczególności ładne górne oszacowania na błąd względny klasyfikatora opartego na LASSO (w stosunku do błędu optymalnego klasyfikatora bayesowskiego) dla przypadku logistycznej i kwadratowej funkcji strat, gdy prawdziwa funkcja wiążąca w modelu nie jest znana. Podobne wyniki zostały uzyskane nieco wcześniej przez Kubkowskiego i Mielniczuka (2020) dla lipschitzowskich funkcji strat, co nie obejmowało przypadku funkcji kwadratowej. W [A3] podano także warunki przy których w opisanej wyżej sytuacji dostaje się zgodność selekcyjną.

W pracy [A7] rozważana jest tzw. regresja porządkowa zwana też problemem rangowania. Chodzi tu o wskazanie, który z dwóch porównywanych obiektów jest „lepszy" (porządek wyznacza nieobserwowana zmienna zależna) na podstawie obserwowanych zmiennych niezależnych. Wyznaczenie reguły rangującej dokonuje się w tym przypadku także na podstawie próby uczącej przez minimalizację empirycznego ryzyka z dodaną karą ℓ_1, ale ryzyko empiryczne nie jest tu sumą niezależnych zmiennych losowych, lecz U-statystyką rzędu 2, co wymaga przy badaniu własności procedur użycia teorii tzw. U-procesów. Główny wynik w A7 to eleganckie oszacowanie ryzyka względnego i błędu estymacji w tw. 1.

Pozostałe prace [A6] i [A8] wchodzące w skład rozprawy dotyczą przypadku niskowymiarowego, w którym wymiar modelu p jest ustalony, gdy liczba obserwacji n rośnie. W [A8] uzyskano wyniki dotyczące zgodności selekcyjnej i własności estymatorów dla klasy problemów z wypukłą funkcją strat i karą LASSO - standardową i adaptacyjną z wagami opartymi na wstępnym estymatorze parametrów. Pokazano w ten sposób, że techniki dowodowe używane wcześniej przez innych autorów do analizy modeli ze szczególnymi funkcjami strat można zaadaptować do ogólnego przypadku wypukłego. Wyniki z [A8] dotyczą sytuacji, w której funkcjonał ryzyka empirycznego jest sumą niezależnych zmiennych losowych. Analogiczne do [A8] wyniki dotyczące sytuacji, w której funkcjonał ryzyka empirycznego jest U-statystyką opisano w pracy [A6]. Obejmują one w szczególności problem rangowania.

Ważną zaletą wszystkich przedstawionych prac jest ich część implementacyjno/symulacyjna. Pokazuje ona bowiem, że zaproponowane algorytmy nie tylko mają pożądane własności teoretyczne, ale mogą też być efektywnie zaimplementowane, działają w wielu przypadkach lepiej od dotychczas stosowanych i pozwalają na ciekawe zastosowania do rzeczywistych zbiorów danych.

Wyniki opisane w pracach wchodzących w skład rozprawy wymagały zastosowania

zaawansowanych narzędzi z różnych dziedzin matematyki (statystyka matematyczna, analiza wypukła, teoria procesów empirycznych i U-procesów) i dowodzą dojrzałej kompetencji matematycznej doktora Rejchela. Są też one bardzo dobrze osadzone w nurcie światowych badań i jasno opisane w artykułach [A1-A8] i przedstawionym autoreferacie wraz z dojrzałą dyskusją na tle wyników innych autorów. W autoreferacie raziło mnie tylko nadużywanie przez autora anglicyzmów („sekcja" zamiast „rozdział", „aktualne wartości" zamiast „rzeczywiste wartości", itp.)

Pewien niedosyt pozostawia słaby, jak na razie, oddźwięk prac wchodzących w skład rozprawy. 8 tych prac ma dotychczas tylko 10 cytowań, z czego 4 to autocytowania, a pozostałe są raczej mało istotne. Może to być częściowo wynikiem stosunkowo krótkiego czasu jaki upłynął od ich publikacji.

Omówienie pozostałego dorobku

Dorobek publikacyjny niewchodzący w skład rozprawy habilitacyjnej obejmuje 11 prac [B1-B11] (wg numeracji z autoreferatu), z których dwie [B1, B2] ukazały się przed uzyskaniem przez W. Rejchela stopnia doktora. Prace [B1, B2, B4, B8, B9, B10] dotyczą problemu rangowania, prace [B3, B5, B7] dotyczą estymatorów największej wiarogodności otrzymanych metodami Monte Carlo, a aplikacyjne prace [B6, B11] poświęcone są porównywaniu jakości czterech metod używanych do produkcji map cyfrowych.

W tej grupie główne, w mojej ocenie, i najczęściej (40 razy wg WoS, 30 III 2022) cytowane osiągnięcie to samodzielna praca [B10], w której badane są własności reguł rangujących otrzymywanych przez minimalizację tzw. ryzyka empirycznego (w sensie używanym w uczeniu maszynowym). Rozwijając wcześniejsze idee Clémençona, Lugosiego i Vayatisa (2008) polegające na motywowanej efektywnością algorytmów odpowiedniej wypukłej modyfikacji funkcji straty i wykorzystaniu teorii U-procesów (indeksowanych funkcjami procesów, które mają reprezentację w terminach U-statystyk) uzyskano w [B10] nieasymptotyczne oszacowania względnego ryzyka wypukłego (ang. *excess convex risk*) badanych reguł rangujących i pokazano kiedy tempa zbieżności do zera tego ryzyka są szybsze od $n^{-1/2}$, gdzie n jest wielkością próby uczącej. Odbywa się to kosztem zmniejszenia złożoności klas funkcyjnych, na których minimalizuje się tzw. empiryczne ryzyko wypukłe, co musi oczywiście spowodować wzrost błędu aproksymacji względem minimalizacji na klasie wszystkich funkcji mierzalnych. Nie jest to analizowane ilościowo w [B10] (podobnie zresztą jak u Clémençona i in.), co nie pozwala przenieść uzyskanych temp zbieżności względnego ryzyka wypukłego na tempa zbieżności wyjściowego ryzyka względnego, które są właściwym wskaźnikiem jakości konstruowanych reguł rangujących. Istniejące wyniki dotyczące zbieżności do zera błędu aproksymacji (ale bez temp!) pozwalają co najwyżej wywnioskować zgodność konstruowanych reguł rangujących. To powoduje, że rozważanie temp zbieżności względnego ryzyka wypukłego traci nieco na wadze. Niemniej jednak, są to wyniki na bardzo dobrym poziomie matematycznym, wykorzystujące nietrywialne narzędzia oparte na entropii metrycznej klas funkcji i teorii procesów empirycznych. Warianty podobnych wyników uzyskane nieco innymi metodami i dające się zastosować do

szerszych klas algorytmów rangujących opisane są w [B4, B8]. Wypukła modyfikacja funkcji straty dla problemu rangowania z rodziną liniowych reguł rangujących jest rozważana wspólnie z W. Niemirą w pracy [B1], gdzie zbadano zgodność i asymptotyczną normalność wektora estymowanych współczynników funkcji rangującej.

Na dobrym poziomie matematycznym są także współautorskie prace [B3, B5, B7], w których badano zgodność i rozkłady asymptotyczne różnych aproksymacji estymatorów największej wiarogodności wyznaczanych metodami Monte Carlo, wykorzystując przy tym ich strukturę martyngałową. Jest to grupa metod pierwotnie zaproponowanych w latach 90 (np. Geyer, Thompson (1992)) i przydatnych w estymacji złożonych modeli stochastycznych, w szczególności modeli z trudnymi do wyznaczenia stałymi normujacymi. Choć prace [B3, B5, B7] nie znalazły na razie oddźwięku (brak cytowań, poza 3 autocytowaniami), to uważam je za interesujące, także z punktu widzenia możliwych zastosowań praktycznych opisanych np. w [A5].

Prace [B11, B6] są opisem zastosowania elementarnych metod statystycznych (parametryczny i nieparametryczny problem dwóch prób, ANOVA) do pewnego praktycznego problemu z dziedziny geodezji, więc nie będę ich tu szerzej omawiał.

Dr Rejchel wygłosił 44 referaty na konferencjach naukowych (w tym 7 referatów zaproszonych). Około połowy z tych konferencji miało charakter międzynarodowy. Jest osobą dobrze znaną w polskim środowisku statystyków matematycznych i aktywnym uczestnikiem corocznych konferencji „Statystyka Matematyczna" w Będlewie. Był kierownikiem grantu FUGA z NCN (2014-2016) i wykonawcą w czterech innych grantach. Poza głównym zatrudnieniem na Wydziale Matematyki i Informatyki UMK w Toruniu prowadził też w latach 2014-2016 i 2017-2018 badania na Wydziale Matematyki, Informatyki i Mechaniki UW w ramach stażu podoktorskiego i podoktorskiego stypendium badawczego.

Wskaźniki cytowań (16 prac zarejestrowanych w WoS było cytowanych 61 razy, w tym 14 autocytowań, indeks Hirscha = 3, WoS 27.04.2022) nie są może imponujące, ale w mojej ocenie wystarczające. Mimo sporej całkowitej liczby cytowań indeks Hirscha jest stosunkowo niski, bo większość z tych cytowań (41) dotyczy jednej pracy [B10].

Konkluzja

Podsumowując stwierdzam, że rozprawa habilitacyjna doktora Wojciecha Rejchela oraz jego pozostały dorobek stanowią znaczny wkład w rozwój statystyki matematycznej i spełniają wymogi Ustawy w tym zakresie. Bez wątpliwości popieram wniosek o nadanie mu stopnia doktora habilitowanego w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie matematyka.

Kraków, 11 maja 2022 r.

Literatura

Bunea F, Wegkamp MH, Auguste A (2006) Consistent variable selection in high dimensional regression via multiple testing, *Journal of Statistical Planning and Inference* **136**, 4349–4364.

Clémençon S, Lugosi G, Vayatis N (2008) Ranking and empirical minimization of U-statistics, *Annals of Statistics* **36**, 844–874.

Gejer CJ, Thompson EA (1992) Constrained Monte Carlo maximum likelihood for dependent data, *Journal of the Royal Statistical Society B* **54**, 657–699.

Kubkowski M, Mielniczuk J (2020) Selection consistency of Lasso-based procedures for misspecified high-dimensional binary model and random regressors, *Entropy* **22**, 153.

Pokarowski P, Mielniczuk J (2015) Combined ℓ_1 and greedy ℓ_0 penalized least squares for linear model selection, *Journal of Machine Learning Research* **16**, 961-992.

Zhu LP, Zhu LX (2009) Nonconcave penalized inverse regression in single-index models with high dimensional predictors, *Journal of Multivariate Analysis* **100**, 862-875.