Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika w Toruniu
ul. Chopina 12/18
87-100 Toruń
za pośrednictwem:
Rady Doskonałości Naukowej
pl. Defilad 1
00-901 Warszawa
(Pałac Kultury i Nauki, p. XXIV, pok. 2401)

Wojciech Rejchel
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika w Toruniu
ul. Chopina 12/18
87-100 Toruń

Wniosek

z dnia 28.10.2021 r.

o przeprowadzenie postępowania w sprawie nadania stopnia doktora habilitowanego w dziedzinie **nauk ścisłych i przyrodniczych** w dyscyplinie[1] **matematyka**

Określenie osiągnięcia naukowego będącego podstawą ubiegania się o nadanie stopnia doktora habilitowanego

M-estymatory z karą w wyborze modelu

Wnioskuję – na podstawie art. 221 ust. 10 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 zm.) – aby komisja habilitacyjna podejmowała uchwałę w sprawie nadania stopnia doktora habilitowanego w głosowaniu ~~**tajnym**~~**/jawnym***[2]

Zostałem poinformowany, że:

Administratorem w odniesieniu do danych osobowych pozyskanych w ramach postępowania w sprawie nadania stopnia doktora habilitowanego jest Przewodniczący Rady Doskonałości Naukowej z siedzibą w Warszawie (pl. Defilad 1, XXIV piętro, 00-901Warszawa).

Kontakt za pośrednictwem e-mail:kancelaria@rdn.gov.pl, tel. 226566098 lub w siedzibie organu. Dane osobowe będą przetwarzane w oparciu o przesłankę wskazaną w art. 6 ust. 1 lit. c) Rozporządzenia UE 2016/679 z dnia z dnia 27 kwietnia 2016 r. w związku z art. 220 - 221 oraz art.232–240 ustawy z dnia 20 lipca 2018 roku - Prawo o szkolnictwie wyższym i nauce, w celu przeprowadzenie postępowania o nadanie stopnia doktora habilitowanego oraz realizacji praw i obowiązków oraz środków odwoławczych przewidzianych w tym postępowaniu.

Szczegółowa informacja na temat przetwarzania danych osobowych w postępowaniu dostępna jest na stronie www.rdn.gov.pl/klauzula-informacyjna-rodo.html

.....
(podpis wnioskodawcy)

[1] Klasyfikacja dziedzin i dyscyplin wg. rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 20 września 2018 r. w sprawie dziedzin nauki i dyscyplin naukowych oraz dyscyplin w zakresie sztuki (Dz. U. z 2018 r. poz. 1818).

[2] * Niepotrzebne skreślić.

Załączniki: wersja elektroniczna wniosku oraz następujące dokumenty (również w wersji elektronicznej)

1. Dane wnioskodawcy.
2. Kopia dokumentu potwierdzającego posiadanie stopnia doktora.
3. Autoreferat.
4. Wykaz osiągnięć naukowych albo artystycznych, stanowiących znaczny wkład w rozwój określonej dyscypliny.
5. Prace wchodzące w skład osiągnięcia naukowego.
6. Potwierdzenie otrzymania stypendium badawczego z grantu 2015/17/B/ST6/01878.
7. Potwierdzenie otrzymania stażu doktorskiego 2014/12/S/ST1/00344.
8. Zaproszenie od prof. Luoqinga Li do odbycia wizyty naukowej na Uniwersytecie Hubei w Wuhan (Chiny).
9. Powołanie na promotora pomocniczego w rozprawie mgr. P. Truszczyńskiego.
10. Powołanie na promotora pomocniczego w rozprawie mgr. P. Krasuskiego.
11. Potwierdzenie otrzymania grantu N N201 391237.
12. Potwierdzenie otrzymania indywidualnej nagrody rektora UMK w 2020 r.
13. Oświadczenia współautorów prac [A1, A2, A3, A4, A5].

Wojciech Rejchel
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika
ul. Chopina 12/18
87-100 Toruń

28.10.2021r.
Załącznik 4.

Wykaz osiągnięć naukowych albo artystycznych, stanowiących znaczny wkład w rozwój określonej dyscypliny

I. INFORMACJA O OSIĄGNIĘCIACH NAUKOWYCH ALBO ARTYSTYCZNYCH, O KTÓRYCH MOWA W ART. 219 UST. 1. PKT 2 USTAWY

Jako osiągnięcie naukowe wskazuję cykl powiązanych tematycznie artykułów naukowych pod wspólnym tytułem

M-estymatory z karą w wyborze modelu

Publikacje wchodzące w skład osiągnięcia (wszystkie powstały po doktoracie):

[A1] P. Pokarowski, W. Rejchel, A. Sołtys, M. Frej, J. Mielniczuk (2021). „Improving Lasso for model selection and prediction", Scandinavian Journal of Statistics, p. 1-33, https://doi.org/10.1111/sjos.12546
Impact Factor: 1.396, Punktacja MEiN (wcześniej MNiSW, max=200pkt): 140
Sekcja 3, w szczególności Twierdzenie 4 z dowodem, jest moim indywidualnym wkładem w powstanie pracy. Ponadto wspólnie z P. Pokarowskim oraz M. Frejem udowodniłem Twierdzenie 2. Mój wkład w redakcję pracy był podobny do pozostałych autorów (nie licząc M. Freja).

[A2] W. Rejchel, M. Bogdan (2020). „Rank-based Lasso - efficient methods for high-dimensional robust model selection", Journal of Machine Learning Research, vol. 21, p. 1-47.
Impact Factor: 3.654, Punktacja (max=200pkt): 140
Koncepcja pracy, część teoretyczna i numeryczna zostały przygotowane wspólnie. Mój indywidualny wkład polegał na udowodnieniu wszystkich wyników teoretycznych (Twierdzenie 2, Wniosek 3, Twierdzenie 5, Twierdzenie 7, Twierdzenie 8, Lemat 9, Twierdzenie 10, Twierdzenie 11 oraz pomocnicze wyniki w dodatku).

[A3] K. Furmańczyk, W. Rejchel (2020). „Prediction and variable selection in high-dimensional misspecified binary classification", Entropy, 22, 543.
Impact Factor: 2.524, Punktacja (max=200pkt): 100

Koncepcja pracy, Sekcja 3 oraz Sekcja 6 zostały opracowane wspólnie. Mój indywidualny wkład to Sekcja 4 oraz Sekcja 5, w szczególności sformułowanie i udowodnienie Twierdzenia 3, Wniosku 2 oraz Wniosku 3.

[A4] K. Furmańczyk, W. Rejchel (2020). „High-dimensional linear model selection motivated by multiple testing", Statistics, vol. 54, p. 152-166.
Impact Factor: 1.051, Punktacja (max=200pkt): 70
Wspólnie opracowaliśmy koncepcję pracy, zaprojektowaliśmy algorytm oraz przygotowaliśmy część praktyczną pracy (Sekcja 4). Odegrałem wiodącą rolę w opracowaniu części teoretycznej (Sekcja 2 oraz Sekcja 3), w szczególności w dowodzeniu Twierdzenia 3.1.

[A5] B. Miasojedow, W. Rejchel (2018). „Sparse estimation in Ising model via penalized Monte Carlo methods", Journal of Machine Learning Research, vol. 19, p. 1-26.
Impact Factor: 4.091, Punktacja (max=50pkt): 50
Wspólnie opracowaliśmy koncepcję pracy, zaprojektowaliśmy algorytm, zaplanowaliśmy część teoretyczną pracy (Sekcja 2 oraz Sekcja 3), a także praktyczną pracy (Sekcja 4 oraz Sekcja 5). Moim indywidualnym wkładem było sformułowanie i udowodnienie głównych wyników teoretycznych pracy (Twierdzenie 2 oraz Wniosek 3), jak również rezultatów pomocniczych w dodatku.

[A6] W. Rejchel (2017). „Model selection consistency of U-statistics with convex loss and weighted Lasso penalty", Journal of Nonparametric Statistics, vol. 29, p. 768-791.
Impact Factor: 0.630, Punktacja (max=50pkt): 20
Jestem jedynym autorem pracy.

[A7] W. Rejchel (2017). „Oracle inequalities for ranking and U-processes with Lasso penalty", Neurocomputing, vol. 239, p. 214-222.
Impact Factor: 3.241, Punktacja (max=50pkt): 30
Jestem jedynym autorem pracy.

[A8] W. Rejchel (2016). „Lasso with convex loss function: model selection consistency and estimation", Communications in Statistics: Theory and Methods, vol. 45, p. 1989-2004.
Impact Factor: 0.311, Punktacja (max=50pkt): 15
Jestem jedynym autorem pracy.

II. INFORMACJA O AKTYWNOŚCI NAUKOWEJ ALBO ARTYSTYCZNEJ

1. Wykaz opublikowanych artykułów w czasopismach naukowych (poza wymienionymi w pkt I).

Prace opublikowane przed uzyskaniem stopnia doktora

[B1] W. Niemiro, W. Rejchel (2009). „Rank correlation estimators and their limiting distributions", Statistical Papers, vol. 50, p. 887-893
Impact Factor: 0.396, Punktacja (max=50pkt): 10

[B2] W. Rejchel (2009). "Ranking - convex risk minimization", Proceedings of World Academy of Science, Engineering and Technology, vol. 56, p. 172-178

Prace opublikowane po uzyskaniu stopnia doktora

[B3] B. Miasojedow, W. Niemiro, W. Rejchel (2021). „Asymptotics of maximum likelihood estimators based on Markov chain Monte Carlo methods", Annales de l'Institut Henri Poincaré - Probabilités et Statistiques, Vol. 57, p. 815-829
Impact Factor: 1.851, Punktacja (max=200pkt): 140

[B4] W. Rejchel (2018). „Generalization Bounds for Ranking Algorithms", rozdział w „Ensemble Classification Methods with Applications in R" (Eds. E. Alfaro, M. Gámez, N. García), Wiley, p. 135-140.
Punktacja (max=50pkt): 20

[B5] B. Miasojedow, W. Niemiro, J. Palczewski, W. Rejchel (2016). „Asymptotics of Monte Carlo maximum likelihood estimators", Probability and Mathematical Statistics, vol. 36, p. 295-310.
Impact Factor: 0.150, Punktacja (max=50pkt): 15

[B6] A. Doskocz, W. Rejchel (2016). „Evaluation of accuracy of digital map data via multiple comparisons", Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 64, p. 799-805.
Impact Factor: 1.156, Punktacja (max=50pkt): 20

[B7] B. Miasojedow, W. Niemiro, J. Palczewski, W. Rejchel (2016). „Adaptive Monte Carlo Maximum Likelihood", rozdział w Studies in Computational Intelligence, Vol. 605: Challenges in Computational Statistics and Data Mining (Eds. S. Matwin, J. Mielniczuk), Springer,
Punktacja (max=50pkt): 5

[B8] W. Rejchel (2015). „Fast rates for ranking with large families", Neurocomputing, vol. 168, p. 1104-1110,
Impact Factor: 2.392, Punktacja (max=50pkt): 30

[B9] W. Rejchel, H. Li, C. Ren, L. Li (2015). „Comments and correction on „U-processes and preference learning", Neural Computation, vol. 27, p. 1549-1553
Impact Factor: 1.626, Punktacja (max=50pkt): 25

[B10] W. Rejchel (2012). „On ranking and generalization bounds", Journal of Machine Learning Research, vol. 13, p. 1373-1392
Impact Factor: 3.420, Punktacja (max=50pkt): 50

[B11] A. Doskocz, W. Rejchel (2012). „Propozycja automatyzacji analizy dokładności baz danych map wielkoskalowych", Zeszyty Naukowe Politechniki Rzeszowskiej, seria Budownictwo i Inżynieria Środowiska, vol. 59, p. 85-93,
Punktacja (max=50pkt): 4

2. Wystąpienia na krajowych lub międzynarodowych konferencjach naukowych lub artystycznych

Wystąpienia na konferencjach naukowych przed uzyskaniem stopnia doktora

Wystąpienia zaproszone

[C1] „On rank regression, minimization of U-processes and some probabilistic inequalities"- International Conference on Trends and Perspectives in Linear Statistical Inference, LinStat2010, 27-31.07.2010 r., Tomar, Portugalia

[C2] „O regresji rangowej i jej estymatorach" - IV Forum Matematyków Polskich, 01-03.07.2010 r., Olsztyn

Wystąpienia zgłoszone

[C3] „Maszyny wektorów podpierających w regresji rangowej" - XXXVI Konferencja „Statystyka Matematyczna", 06-10.12.2010, Wisła

[C4] „Ranking - minimalizacja ryzyka wypukłego" - XXXV Konferencja Statystyka Matematyczna, 07-11.12.2009 r., Wisła,

[C5] „Estymatory regresji rangowej i ich asymptotyka" - XXXVIII Ogólnopolska Konferencja Zastosowań Matematyki, 08-15.09.2009, Zakopane

[C6] „Ranking - convex risk minimization" - International Conference on Machine Learning and Data Analysis, 26-28.08.2009, Singapur

[C7] „Rank correlation estimators and their limiting distributions" - The international conference on trends and perspectives in linear statistical inference, LinStat2008, 21-25.04.2008, Będlewo

[C8] „Ograniczenie prawdopodobieństwa błędu w boostingu" - XXXIII Konferencja „Statystyka Matematyczna", 03-07.12.2007, Wisła

Wystąpienia na konferencjach naukowych po uzyskaniu stopnia doktora

Wystąpienia zaproszone

[C9] „Efficient methods for high-dimensional robust variable selection" - 14th International Conference on Computational and Financial Econometrics (CFE 2020), 19-21.12.2020, on-line

[C10] „Szybka i odporna selekcja cech w modelach regresyjnych" - Jubileuszowy Zjazd Matematyków Polskich w stulecie PTM, 3-7.09.2019, Kraków

[C11] „High-dimensional model selection via generalized information criterion" - Joint Meeting of UMI-SIMAI-PTM, 17-20.09.2018, Wrocław

[C12] „Generalized information criterion in high-dimensional model selection" - The 23rd International Conference on Computational Statistics COMPSTAT 2018, 28-31.08.2018, Jassy, Rumunia

[C13] „Metody Monte Carlo w wysokowymiarowym modelu Isinga" - VIII Forum Matematyków Polskich, 18-22.09.2017, Lublin

Wystąpienia zgłoszone

[C14] „Szybka i odporna selekcja cech w modelach wysokowymiarowych" - XLIX Ogólnopolska Konferencja Zastosowań Matematyki, 20-25.09.2021, Zakopane

[C15] „Fast and robust procedures in high-dimensional variable selection" - Bernoulli-IMS One World Symposium 2020, 24-28.08.2020, on-line

[C16] „Variable selection in high-dimensional binary regression" - XLV Konferencja „Statystyka Matematyczna", 02-06.12.2019, Będlewo,

[C17] „Fast and robust model selection based on ranks" - IX International Workshop on Perspectives on High-Dimensional Data Analysis (HDDA IX), 24-27.06.2019, Uppsala, Szwecja

[C18] „Rank-based model selection" - XLIV Konferencja „Statystyka Matematyczna", 02-07.12.2018, Będlewo,

[C19] „Improving Lasso for model selection and prediction" - Workshop "Model selection, regularization and inference", 12 - 14.07.2018, Wiedeń, Austria

[C20] „Selekcja cech w oparciu o kryterium GIC z karą LASSO w modelach wysokowymiarowych" - XLIII Konferencja „Statystyka Matematyczna",

04-08.12.2017, Będlewo

[C21] „Penalized Monte Carlo methods in high-dimensional Ising model" - Mathematical Methods of Modern Statistics, 10-14.07.2017, Luminy, Francja,

[C22] „High-dimensional Ising model and Monte Carlo methods" - XLII Konferencja "Statystyka Matematyczna", 27.11-02.12.2016, Będlewo

[C23] „Asymptotic properties of U-processes with convex loss and weighted Lasso penalty" - The 3rd Conference of the International Society for Non-Parametric Statistics (ISNPS), 11-16.06.2016, Awinion, Francja

[C24] „Oracle inequalities for ranking with Lasso penalty" - The 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2015), 12-14.12.2015, Londyn, Wielka Brytania

[C25] „Asymptotyczne własności U-procesów z karą Lasso" - XLI Konferencja „Statystyka Matematyczna", 06-12.12.2015, Będlewo

[C26] „Własności estymatorów regresji porządkowej z karą LASSO" - V spotkanie Polskiej grupy badawczej systemów uczących się, 23-24.04 2015, Gliwice

[C27] „Regresja rangowa z karą LASSO oraz nierówności z wyrocznią" - XL Konferencja ,,Statystyka Matematyczna", 30.11-05.12.2014, Będlewo

[C28] „High-dimensional problems in ranking (ordinal regression) with Lasso" - The 21st International Conference on Computational Statistics COMPSTAT 2014, 19-22.08.2014, Genewa, Szwajcaria

[C29] „Estymatory regresji rangowej oparte na metodzie LASSO" - XXXIX Konferencja ,,Statystyka Matematyczna", 02-06.12.2013, Wisła

[C30] plakat ,,Lasso and adaptive Lasso with convex loss functions" - International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: theory and applications (ROKS 2013), 08-10.07.2013, Leuven, Belgia

[C31] plakat ,,Use of nonparametric statistics for estimation of accuracy of digital map data" - German-Polish Joint Conference on Probability and Mathematical Statistics, 06-09.06.2013, Toruń

[C32] „Lasso and adaptive Lasso with convex loss functions" - German-Polish Joint Conference on Probability and Mathematical Statistics, 06-09.06.2013, Toruń

[C33] „Wybór modelu w oparciu o metodę LASSO z wypukłą funkcją straty" - XXXVIII Konferencja ,,Statystyka Matematyczna", 03-07.12.2012, Wisła,

[C34] „Maximum likelihood estimation via Monte Carlo methods" - The 20th International Conference on Computational Statistics COMPSTAT 2012, 27-31.08.2012, Limassol, Cypr

[C35] „Oszacowanie ryzyka estymatorów regresji rangowej" - XXXVII Konferencja „Statystyka Matematyczna", 05-09.12.2011, Wisła,

[C36] „O regresji rangowej, jej estymatorach i ich własnościach" - XL Konferencja Zastosowań Matematyki, 30.08-06.09.2011, Zakopane

3. Informacja o udziale w komitetach organizacyjnych i naukowych konferencji krajowych lub międzynarodowych, z podaniem pełnionej funkcji.

- członek Komitetu Organizacyjnego XLIV Konferencji „Statystyka Matematyczna", 02-07.12.2018, Będlewo

4. Informacja o uczestnictwie w pracach zespołów badawczych realizujących projekty finansowane w drodze konkursów krajowych lub zagranicznych, z podziałem na projekty zrealizowane i będące w toku realizacji, oraz z uwzględnieniem informacji o pełnionej funkcji w ramach prac zespołów.

Projekty realizowane przed uzyskaniem stopnia doktora

a) główny wykonawca w grancie promotorskim MNiSW N N201 391237 "Statystyczne modele regresji rangowej" na realizację pracy doktorskiej 2009-2011, kierownik: prof. dr hab. W. Niemiro

b) kierownik stypendium z projektu "Stypendia dla doktorantów 2008/2009 - ZPORR"

Projekty realizowane po uzyskaniu stopnia doktora

c) wykonawca w grancie OPUS z NCN, 2018/31/B/ST1/00253, „Metody obliczeniowe dla wysokowymiarowego uczenia statystycznego", 2019-2023 (w realizacji), kierownik: dr. hab. B. Miasojedow

d) stanowisko podoktorskie w grancie OPUS z NCN, 2015/17/B/ST6/01878, „SOSnet: oszczędne modelowanie i predykcja dla danych wysokiego wymiaru", 2016-2019, kierownik: dr hab. P. Pokarowski

e) kierownik w grancie FUGA z NCN, 2014/12/S/ST1/00344, „Regresja rangowa i U-procesy z karą LASSO - selekcja cech, estymacja i nierówności z wyrocznią", 2014-2016

f) wykonawca w grancie OPUS z NCN, N N201 608740, „Asymptotic properties and inequalities for MCMC estimators", 2011-2014, kierownik: prof. dr hab. W. Niemiro

5. Członkostwo w międzynarodowych lub krajowych organizacjach i towarzystwach naukowych wraz z informacją o pełnionych funkcjach.

   - członek Komisji Statystyki Komitetu Matematyki PAN - od 2020 r.

6. Informacja o odbytych stażach w instytucjach naukowych lub artystycznych, w tym zagranicznych, z podaniem miejsca, terminu, czasu trwania stażu i jego charakteru.

   a) stypendium badawcze postdoc (opisane w II.4.d) zostało mi przyznane w ramach konkursu ogłoszonego przez kierownika grantu z NCN dr. hab. P. Pokarowskiego. Stypendium było realizowane od 10.2017 do 09.2018 na Wydziale Matematyki, Informatyki i Mechaniki UW. W jego wyniku powstała publikacja [A1],

   b) staż podoktorski FUGA (opisany w II.4.e) został mi przyznany przez NCN. Byłem jego kierownikiem, a opiekunem naukowym był prof. dr hab. W. Niemiro. Staż odbyłem na Wydziale Matematyki, Informatyki i Mechaniki UW od 10.2014 do 09.2016. W jego wyniku powstały trzy publikacje [A5, A6, A7].

7. Informacja o recenzowanych pracach naukowych lub artystycznych, w szczególności publikowanych w czasopismach międzynarodowych.

   a) recenzent artykułów naukowych dla następujących czasopism międzynarodowych: Applicationes Mathematicae, Artificial Intelligence in Medicine, Neural Computation, Neurocomputing, Scandinavian Journal of Statistics, Statistics, Statistics in Transition

   b) recenzent AMS Mathematical Reviews/MathSciNet

   c) recenzent pracy nadesłanej na 4th Conference of the International Society for Nonparametric Statistics (ISNPS), 11-15.06.2018, Salerno, Włochy

   d) recenzent prac licencjackich i magisterskich na Wydziale Matematyki i Informatyki UMK

8. Informacja o udziale w zespołach badawczych, realizujących projekty inne niż określone w pkt. II.4

   a) kierownik w grancie Wydziału Matematyki i Informatyki UMK „Własności estymatorów opartych na metodzie Lasso z wypukłą funkcją straty", 2013
   b) kierownik w grancie Wydziału Matematyki i Informatyki UMK "Estymatory największej wiarogodności a metody Monte Carlo", 2012

9. Informacja o uczestnictwie w zespołach oceniających wnioski o finansowanie badań, wnioski o przyznanie nagród naukowych, wnioski w innych konkursach mających charakter naukowy lub dydaktyczny.

   - członek Jury Konkursu PTM na najlepszą pracę studencką z teorii prawdopodobieństwa i zastosowań matematyki - 2020 r.

   - członek komisji w konkursie na stypendium badawcze dla doktoranta w grancie BEETHOVEN z NCN, 2018/31/G/ST1/02252, „Analiza wrażliwości operatorów nielokalnych z zastosowaniami do procesów skokowych", kierownik: prof. K. Bogdan - lipiec 2020 r.

III. INFORMACJE NAUKOMETRYCZNE

Researcher ID: D-4813-2014
ORCID: 0000-0003-1148-1439
Scopus author ID: 29867586300

1. Informacja o punktacji Impact Factor

   Sumaryczny Impact Factor (IF) wszystkich publikacji naukowych według listy Journal Citation Reports (JCR), zgodnie z rokiem opublikowania (w przypadku publikacji z roku 2021 zastosowano IF z roku 2020) wynosi 27.889.
   - przed uzyskaniem stopnia doktora 0.396
   - po uzyskaniu stopnia doktora 27.493

2. Informacja o liczbie cytowań publikacji wnioskodawcy, z oddzielnym uwzględnieniem autocytowań.
   - wg bazy Web of Science: 52 cytowania (bez autocytowań: 44)
   - wg bazy Scopus: 67 cytowań (bez autocytowań: 52)

3. Informacja o posiadanym indeksie Hirscha.
   - wg bazy Web of Science: 3
   - wg bazy Scopus: 3

4. Informacja o liczbie punktów MEiN (wcześniej MNiSW).

   Sumaryczna liczba punktów wg Listy wszystkich publikacji naukowych, zgodnie z rokiem opublikowania wynosi 294 dla lat 2009-2018 oraz 590 dla lat 2019-2021:
   - przed uzyskaniem stopnia doktora: 10 dla lat 2009-2018
   - po uzyskaniu stopnia doktora: 284 dla lat 2009-2018 oraz 590 dla lat 2019-2021

.....

(podpis wnioskodawcy)

Autoreferat

Wojciech Rejchel

Spis treści

1 Imię i nazwisko **2**

2 Posiadane dyplomy i stopnie naukowe **2**

3 Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych **2**

4 Omówienie osiągnięcią, o których mowa w art. 219 ust. 1 pkt. 2 Ustawy **2**

4.1 Wprowadzenie . . . 3

4.1.1 M-estymatory . . . 3

4.1.2 M-estymatory z karą . . . 4

4.1.3 Osiągnięcia . . . 5

4.2 Definicje i oznaczenia . . . 6

4.3 Wysokowymiarowe modele parametryczne . . . 7

4.3.1 Uogólnione modele liniowe . . . 7

4.3.2 Model Isinga . . . 11

4.4 Wysokowymiarowe modele semiparametryczne . . . 13

4.4.1 Estymatory oparte na rangach . . . 13

4.4.2 Algorytm Screening-Selection poza GLM . . . 16

4.4.3 Źle wyspecyfikowana klasyfikacja binarna . . . 17

4.4.4 Regresja porządkowa . . . 18

4.5 Analiza danych niskowymiarowych . . . 20

4.5.1 Niskowymiarowe modele semiparametryczne . . . 21

4.6 Pozostałe osiągnięcia naukowo-badawcze . . . 23

5 Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej **26**

6 Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę **27**

7 Pozostałe osiągnięcia naukowo-badawcze i inna działalność **27**

1 Imię i nazwisko

Wojciech Rejchel

2 Posiadane dyplomy i stopnie naukowe

2011 Doktorat (z wyróżnieniem), matematyka
Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu
Tytuł rozprawy: *Statystyczne modele regresji rangowej*,
promotor: prof. dr hab. Wojciech Niemiro

2005 Magisterium, matematyka
Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu
Tytuł pracy: *Estymacja monotonicznej gęstości*,
promotor: prof. dr hab. Tomasz Rychlik

3 Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych

2017 – 2018	Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski, *adiunkt*, stanowisko podoktorskie w projekcie OPUS z NCN
2014 – 2016	Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski, *adiunkt*, staż podoktorski FUGA z NCN
2011 – obecnie	Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu, *adiunkt*
2010 –2011	Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu, *asystent*
2009 –2010	Wydział Matematyki i Informatyki, Uniwersytet Warmińsko-Mazurski w Olsztynie, *asystent*

4 Omówienie osiągnięcią, o których mowa w art. 219 ust. 1 pkt. 2 Ustawy

Jako osiągnięcie naukowe, o którym mowa w art. 219 ust. 1 pkt. 2 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, wskazuję cykl powiązanych tematycznie artykułów naukowych pod wspólnym tytułem

M-estymatory z karą w wyborze modelu

Publikacje wchodzące w skład osiągnięcia

[A1] P. Pokarowski, W. Rejchel, A. Sołtys, M. Frej, J. Mielniczuk (2021). „Improving Lasso for model selection and prediction", Scandinavian Journal of Statistics, p. 1-33, doi.org/10.1111/sjos.12546

[A2] W. Rejchel, M. Bogdan (2020). „Rank-based Lasso - efficient methods for high-dimensional robust model selection", Journal of Machine Learning Research, vol. 21, p. 1-47.

[A3] K. Furmańczyk, W. Rejchel (2020). „Prediction and Variable Selection in High-Dimensional Misspecified Binary Classification", Entropy, 22, 543.

[A4] K. Furmańczyk, W. Rejchel (2020). „High-dimensional linear model selection motivated by multiple testing", Statistics, vol. 54, p. 152-166.

[A5] B. Miasojedow, W. Rejchel (2018). „Sparse estimation in Ising Model via penalized Monte Carlo methods", Journal of Machine Learning Research, vol. 19, p. 1-26.

[A6] W. Rejchel (2017). „Model selection consistency of U-statistics with convex loss and weighted Lasso penalty", Journal of Nonparametric Statistics, vol. 29, p. 768-791.

[A7] W. Rejchel (2017). „Oracle inequalities for ranking and U-processes with Lasso penalty", Neurocomputing, vol. 239, p. 214-222.

[A8] W. Rejchel (2016). „Lasso with convex loss function: model selection consistency and estimation", Communications in Statistics: Theory and Methods, vol. 45, p. 1989-2004.

4.1 Wprowadzenie

Moje badania naukowe skupione są na konstrukcji i badaniu metod potrafiących analizować i interpretować złożone zbiory danych. Interesuję się zarówno danymi niskowymiarowymi, jak i wysokowymiarowymi. Problem niskowymiarowy dotyczy sytuacji, gdy liczba parametrów (cech, zmiennych objaśniających) p może być znacząca (powiedzmy kilkadziesiąt), ale mniejsza niż rozmiar danych n. Przypadek wysokowymiarowy oznacza, że p jest porównywalne bądź większe niż n. Zwłaszcza ta druga sytuacja jest aktualnie intensywnie badana w statystyce i uczeniu maszynowym, gdyż tego typu dane są powszechnie spotykane w genetyce, biologii, ekonomii czy naukach społecznych. Na przykład badając związek między genami a pewną zmienną odpowiedzi, często pracujemy ze zbiorami danych zawierającymi dziesiątki (albo setki) tysięcy genów, podczas gdy rozmiar danych zwykle jest nie większy niż tysiąc.

W obu przypadkach (nisko i wysokowymiarowym) liczba cech może być duża, jednak wiemy (albo wierzymy), że badany model jest *rzadki*. Oznacza to, że większość cech stanowią cechy nieniosące żadnej dodatkowej informacji o badanym zjawisku. Dlatego naszym celem jest znalezienie tego małego zbioru zawierającego tylko cechy istotne. W ten sposób otrzymamy prosty i łatwy do interpretacji związek między cechami a zmienną odpowiedzi. Zagadnienie to nazywane jest "wyborem modelu" i może być skutecznie rozwiązane, używając M-estymatorów z karą.

4.1.1 M-estymatory

Załóżmy, że $X \in \mathbb{R}^p$ jest p-wymiarowym wektorem cech, a $Y \in \mathbb{R}$ jest zmienną odpowiedzi. Zmienna Y może być ciągła jak w regresji liniowej albo dyskretna jak w klasyfikacji. Może być również mierzona na skali porządkowej jak w regresji porządkowej.

Oznaczmy $z = (y, x)$. Niech $\theta \in \mathbb{R}^p$ oraz $\phi(\theta, z)$ będzie nieujemną funkcją straty. Zakładamy, że ϕ jest wypukła względem θ dla ustalonego z oraz mierzalna względem z dla ustalonego θ. Wypukłą funkcję

$$Q(\theta) = \mathbb{E}\phi(\theta, Z) \tag{4.1.1}$$

bedziemy nazywać funkcją ryzyka. Chcemy wyznaczyć element minimalizujący Q, który oznaczymy przez θ^*, to znaczy $\theta^* = \arg\min_{\theta \in \mathbb{R}^p} Q(\theta)$. Zwykle nie potrafimy obliczyć θ^*, gdyż nie znamy rozkładu Z. Jednakże dysponując próbką $Z_1, \ldots, Z_n$, składającą się z niezależnych kopii Z, możemy rozważyć empiryczny odpowiednik funkcji Q dany jako

$$\bar{Q}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \phi(\theta, Z_i), \tag{4.1.2}$$

który bedziemy nazywać ryzykiem empirycznym. Następnie parametr θ^* będziemy przybliżać elementem *minimalizującym*

$$\arg\min_{\theta\in\mathbb{R}^p} \bar{Q}(\theta), \tag{4.1.3}$$

który będziemy nazywać M-estymatorem. Podejście to jest popularne w statystyce i łatwo możemy wskazać wiele przykładów M-estymatorów. Rozważmy model liniowy $Y = X^T\theta^* + \varepsilon$, gdzie ε jest losowym błędem. Przy standardowych założeniach wektor (4.1.3) jest estymatorem najmniejszych kwadratów (ENK), gdy

$$\phi(\theta, y, x) = (y - \theta^T x)^2. \tag{4.1.4}$$

Natomiast dla

$$\phi(\theta, y, x) = |y - \theta^T x| \tag{4.1.5}$$

(4.1.3) staje się estymatorem najmniejszych odchyleń (z ang. least absolute deviations, LAD). Przypuśćmy teraz, że Y jest taka jak w regresji logistycznej, to znaczy $Y \in \{0, 1\}$ oraz $P(Y = 1|X = x) = 1/(1 + \exp(-x^T\theta^*))$. Wtedy biorąc

$$\phi(\theta, y, x) = -yx^T\theta + \log(1 + \exp(x^T\theta)), \tag{4.1.6}$$

będziemy minimalizować funkcję przeciwną do logarytmu wiarogodności.

Ryzyko empiryczne (4.1.2) jest sumą niezależnych zmiennych losowych. W pracach [A2, A5, A6, A7] będziemy rozważać bardziej skomplikowane wyrażenia. Będą to U-statystyki bądź aproksymacje Monte Carlo logarytmu wiarogodności. Przypadki te zostaną dokładnie opisane w dalszych sekcjach.

4.1.2 M-estymatory z karą

M-estymatory z karą definiujemy następująco

$$\arg\min_{\theta\in\mathbb{R}^p} \quad \bar{Q}(\theta) + \lambda Pen(\theta), \tag{4.1.7}$$

gdzie $\bar{Q}(\theta)$ jest określone w (4.1.2) i mierzy dopasowanie modelu. Funkcja $Pen(\theta)$ jest karą za złożoność modelu, a $\lambda > 0$ jest dodatkowym parametrem procedury.

Celem selekcji cech jest znalezienie zbioru

$$T = \{1 \leqslant j \leqslant p : \theta_j^* \neq 0\}, \tag{4.1.8}$$

który zawiera tylko cechy istotne. Na przykład w pracy [1] Akaike zaproponował, aby wybierać model minimalizujący odległość Kullbacka-Leiblera od modelu prawdziwego (4.1.8). Procedura ta sprowadza się do (4.1.7) z $\lambda = 2$ oraz $Pen(\theta) = \sum_{j=1}^p I(\theta_j \neq 0)$, gdzie $I(\cdot)$ jest indykatorem. Zatem złożoność modelu mierzona jest liczbą niezerowych współczynników, często określaną normą w l_0 i oznaczaną $|\theta|_0$. Używając podejścia bayesowskiego, podobny algorytm z $\lambda = \log n$ został zaproponowany w [29]. Inne przykłady były wprowadzone w [21] lub [11]. Naturalnie procedury te są obliczeniowo złożone i mogą być efektywnie wyznaczone jedynie w przypadkach, gdy p jest nie większe niż kilkadziesiąt.

Estymatory regresji grzbietowej (ERG) zostały wprowadzone w [16]. W tym przypadku w (4.1.7) kara jest kwadratem l_2-normy, i.e. $Pen(\theta) = |\theta|_2^2$. Celem była poprawa własności standardowego M-estymatora (4.1.3) w przypadku, gdy cechy są współliniowe bądź p jest względnie duże. Dodanie tej kary sprawia, że estymator staje się obciążony, ale jednocześnie zmniejszamy jego wariancję. Jeśli λ jest rozsądnie wybrana, to błąd ERG w estymacji i predykcji jest mniejszy niż estymatora (4.1.3). Naturalnie ERG nie potrafi wybierać cech (zerować współczynników estymatora). Estymatory "Bridge" z $Pen(\theta) = |\theta|_q$ dla $0 < q \leqslant 2$, przedstawione w [12], w naturalny sposób łączą karę w l_0 (używaną do selekcji cech) z karą w l_2 (używaną do poprawy predykcji). M-estymator z karą w l_1 nazywamy

"LASSO" (z ang. Least Absolute Shrinkage and Selection Operator) [35]. W tym przypadku estymator jest dany jako

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \quad \bar{Q}(\theta) + \lambda |\theta|_1. \tag{4.1.9}$$

Literatura dotycząca M-estymatorów z karą została w ostatnich latach zdominowana przez analizę estymatorów LASSO i ich modyfikacji. Powodem jest fakt, że jedynie dla $q = 1$ kara $Pen(\theta) = |\theta|_q$ jest jednocześnie wypukła i nieróżniczkowalna w zerze. Ta druga własność implikuje, że rozwiązania (4.1.9) są *rzadkie*, jeśli λ jest wystarczająco duże. Oznacza to, że LASSO może być użyte do selekcji cech. Co więcej, niezerowe współczynniki są estymowane, co automatycznie umożliwia predykcję. Z drugiej strony wypukłość funkcji straty ϕ i kary jest kluczowa z teoretycznego, jak i praktycznego punktu widzenia. Po pierwsze każde minimum funkcji wypukłej jest jej globalnym minimum, więc unikamy problemów z minimami lokalnymi. Po drugie własność ta pozwoliła skonstruować algorytmy efektywnie wyznaczające (4.1.9) nawet wtedy, gdy p jest duże [8, 13].

Literatura dotycząca badania własności LASSO w estymacji, predykcji oraz selekcji cech jest bogata [42, 44, 36, 4, 40, 5, 17, 33], zwłaszcza w kontekście modeli parametrycznych. Wyniki te potwierdzają dobrą jakość procedur LASSO w estymacji oraz predykcji, jednak wskazują one znaczące niedostatki w selekcji cech. Mianowicie model wybierany przez LASSO jest zwykle zbyt duży, a zgodna selekcja wymaga dodatkowego warunku (z ang. *the irrepresentable condition*) [24, 42]. Oznaczmy go "IRR". Co ważne warunek ten jest restrykcyjny (bez względu na to czy rozważany model jest nisko- czy wysokowymiarowy, zob. Sekcja 4.2). W pracach [40, 5, 17] zasugerowano, że LASSO nie powinno być nazywane *operatorem selekcji*, ale "operatorem przesiewającym" (z ang. *screening operator*) albo *seperatorem*. Przesiewanie polega na znaczącej redukcji wymiaru problemu bez utraty cech istotnych. Natomiast separacja dotyczy zdolności procedur LASSO polegającej na tym, że współczynniki LASSO odpowiadące cecho istotnym są większe (w wartościach bezwzględnych) niż współczynniki od cech nieistotnych. Obydwie własności opierają się na *estymacyjnej zgodności* procedur LASSO, którą można wykazać przy znacznie słabszych założeniach aniżeli warunek IRR [37].

Aktualnie niekwestionowany jest fakt, że LASSO powinno być traktowane jako pierwszy krok bardziej złożonego algorytmu. Zaproponowano wiele procedur statystycznych, które miały poprawić LASSO. To znaczy miały one być zgodne w selekcji przy słabszych założeniach niż IRR. Głównymi ulepszeniami są: progowe LASSO (z ang. *thresholded* LASSO) [43], adaptacyjne LASSO [44] lub algorytmy z niewypukłymi karami [41, 10]. W przypadku tych ostatnich procedury przestają być wypukłe, co sprawia, że ich analiza oraz implementacja są trudne. Progowe LASSO (TL) poprawia LASSO, odrzucając cechy odpowiadające małym, w wartości bezwzględnej, współczynnikom LASSO. Natomiast adaptacyjne LASSO działa nastepująco: najpierw wstępny estymator jest wyznaczony, na przykład LASSO. Następnie rozważamy ważoną wersję LASSO, w której wagi konstruowane są na podstawie estymatora z poprzedniego kroku. Jednakże procedury z [44, 43, 41, 10] nie mogą być w praktyce stosowane bezpośrednio, gdyż do bycia zgodnymi w selekcji wymagają one użycia nieznanych parametrów. Na przykład nie wiadomo jak wybrać próg w algorytmie TL, zob. Sekcja 4.3.1. W praktyce problemy te próbuje się rozwiązać, używając metody walidacji krzyżowej albo kryteriów informacyjnych (podobnych do procedur z [1, 29]). Inne podejście jest oparte na metodzie "knockoffs" z [2]. Naturalnie użycie wspomnianych rozwiązań zwiększa złożoność obliczeniową algorytmu.

4.1.3 Osiągnięcia

Zajmuję się badaniem M-estymatorów z karą w modelach nisko- oraz wysokowymiarowych. Mogą one być regularne jak uogólnione modele liniowe (GLM). Jednakże nieraz rozważam modele trudniejsze i bardziej nieregularne, na przykład model (4.4.1) podany poniżej. Często pracuję z karą LASSO,

jednakże interesuję się również procedurami z karą w l_0. Moje osiągnięcia zawarte są w pracach [A1-A8] i mogą być streszczone następująco:

1. W pracach [A1, A4, A5] badane są wysokowymiarowe modele parametryczne (GLM oraz model Isinga). Wraz ze współautorami proponujemy M-estymatory z karą, które w tych przypadkach są selekcyjnie zgodne. Zaletami procedur z [A1] i [A4] są *konstruktywność* (przynajmniej w wysokowymiarowym modelu liniowym z suggaussowskim błędem), efektywność obliczeniowa oraz selekcyjna zgodność przy słabych założeniach. Mówimy, że algorytm jest konstruktywny, o ile jego selekcyjna zgodność nie wymaga znajomości nieznanych parametrów. Zaproponowane przez nas procedury oparte są jedynie na *znanych wartościach* takich, jak n, p itp. Natomiast W pracy [A5] proponujemy algorytm używający metod Monte Carlo do aproksymacji trudnej do wyznaczenia stałej normującej w modelu Isinga. Zaproponowane podejście jest lepsze (w teorii i praktyce) niż jego konkurenci.

2. W pracach [A2, A3, A7] oraz [A1, Section 3] badamy wysokowymiarowe modele semiparametryczne. Często zdarza się, że "duże" zbiory danych zawierają liczne "nieregularności" i nie spełniają założeń modeli parametrycznych bądź warunki te są trudne do zweryfikowania. W [A1, A2] opracowujemy nowe procedury, które potrafią pracować z nieregularnymi zbiorami danych. Badamy również ich statystyczne własności. Natomiast w pracach [A3, A7] analizujemy znane i skuteczne procedury, których teoretycznym własnościom nie przyjrzano się wystarczająco dokładnie. Nasze wyniki pomagają lepiej zrozumieć przyczyny dobrej skuteczności tych procedur, jak również ich ograniczenia.

3. Badamy również własności M-estymatorów z karą w przypadku niskowymiarowym. Analiza ta zawarta jest w [A6, A8] oraz Dodatku z [A2]. Aktualnie zagadnienie to nie jest tak intensywnie badane, jak sytuacja wysokowymiarowa. Jednakże, moim zdaniem, przypadek ten jest ważny, gdyż nie każdy zbiór danych, z którym pracujemy, jest wysokowymiarowy. Jest wiele ciekawych problemów, w których liczba parametrów jest znacznie mniejsza niż rozmiar próbki. Oczywiście można stwierdzić, że każdy rezultat dla danych wysokiego wymiaru może być zredukowany do przypadku niskowymiarowego. Jednakże, jak pokażemy później, taka redukcja często prowadzi do nieoptymalnych wyników.

4. Teoretyczne rezultaty z prac [A1-A8] są dopełnione analizą numeryczną, w której badamy selekcyjne i predykcyjne własności procedur. Jeśli algorytmy są niekonstruktywne, to podajemy metody wyboru ich *nieznanych* parametrów.

4.2 Definicje i oznaczenia

Notacja użyta w ośmiu wspomnianych artykułach nie jest spójna. W tej części zostanie ona ujednolicona. Mam nadzieję, że nie będzie to prowadziło do nieporozumień, gdy poszczególne wyniki będą porównywane z ich pierwotnymi wersjami z [A1-A8].

W modelu wysokowymiarowym p moze być znacznie większe niż n. Dlatego zakłada się, że $p = p_n$. Zatem powinniśmy również pisać $X_n, Q_n, \lambda_n, \hat{\theta}_n$ itd., co chyba bardziej przeszkadza niż pomaga. Wobec tego będziemy pomijać ten dolny index.

Niech $\mathbf{X} = (X_{\cdot 1}, \ldots, X_{\cdot p}) = (X_1, \ldots, X_n)^T$ będzie $n \times p$ macierzą eksperymentu, a J będzie dowolnym podzbiorem *modelu pełnego* $F = \{1, 2, \ldots, p\}$, $J^c = F \setminus J$. Na J można patrzeć jak na ciąg zero-jedynkowy na F, stąd $|J| = |J|_1$ oznacza liczność J. Niech θ_J będzie podwektorem wektora θ, którego elementy są wyznaczone przez współczynniki z J. Podobnie $\mathbf{X}_J$ jest podmacierzą macierzy $\mathbf{X}$ z kolumnami wyznaczonymi przez współczynniki z J. Niech $G_J = \mathbf{X}_J(\mathbf{X}_J^T\mathbf{X}_J)^{-1}\mathbf{X}_J^T$ będzie rzutem ortogonalnym na podprzestrzeń wyznaczoną przez kolumny z $\mathbf{X}_J$.

Następnie dla $\theta \in \mathbb{R}^p$ i $q \geqslant 1$ niech $|\theta|_q = (\sum_{j=1}^p |\theta_j|^q)^{1/q}$ będzie ℓ_q-normą. Ponadto $\theta^*_{\min} = \min_{j \in T} |\theta^*_j|$ jest najmniejszym (według wartości bezwzględnej) niezerowym współczynnikiem wektora θ^*. Nośnik wektora θ oznaczamy jako $supp(\theta) = \{1 \leqslant j \leqslant p : \theta_j \neq 0\}$.

Dla $\xi > 1$ oraz $q \geqslant 1$ definiujemy stożek

$$\mathcal{C}(\xi) = \{\theta \in \mathbb{R}^p : |\theta_{T^c}|_1 \leqslant \xi |\theta_T|_1\}. \tag{4.2.1}$$

Niech $H = \mathbb{E} X_1 X_1^T$. *Współczynnik odwrotności* (z ang. cone invertibility factor (CIF)) dany jest jako

$$F_q(\xi) = \inf_{0 \neq \theta \in \mathcal{C}(\xi)} \frac{t^{1/q} |H\theta|_\infty}{|\theta|_q}\,, \tag{4.2.2}$$

gdzie $t = |T|$ i T jest zdefiniowane w (4.1.8). W przypadku deterministycznych wektorów cech przyjmiemy $H = \mathbf{X}^T\mathbf{X}/n$. CIF został wprowadzony w [40] i jest podobny do *ograniczonej wartości własnej* [4] czy *współczynnika zgodności* [36]

$$\kappa(\xi) = \inf_{0 \neq \theta \in \mathcal{C}(\xi)} \frac{t\theta^T H \theta}{|\theta_T|_1^2}\,. \tag{4.2.3}$$

Krótko mówiąc, wartości te określają jak bardzo prawdziwy model T różni się od pozostałych. Można również patrzeć na te wartości jak na miary współzależności cech, ponieważ współczynniki te są uogólnieniem najmniejszej wartości własnej macierzy $\mathbf{X}^T\mathbf{X}/n$, która jest zerem w przypadku, gdy $p > n$. W [A1] oraz [A5] liczniki w (4.2.2) są nieznacznie zmienione, zob. [A1, wzór (13)] oraz [A5, wzór (11)], odpowiednio. Ponadto w [A1, A4] używamy większych (więc lepszych) wartości (z ang. *sign-restricted pseudo-cone invertibility factors* (SCIF)), które mają zmienione stożki (4.2.1), zob. [A1, wzór (12)]. Co więcej, w [A1] rozważany stożek zależy od a zamiast ξ, gdzie $a = (\xi - 1)/(\xi + 1)$.

We Wprowadzeniu wspomnieliśmy o warunku IRR, który teraz przytoczymy. Wygląda on następująco

$$|H_{T^c,T} H_T^{-1} \operatorname{sign}(\theta^*)|_\infty \leqslant 1, \tag{4.2.4}$$

gdzie $H_T = (H_{jj})_{j \in T}$ oraz $H_{T^c,T} = (H_{jk})_{j \notin T, k \in T}$. Dokładne porównania pomiędzy (4.2.2), (4.2.3) i (4.2.4) można znaleźć w [37, 40, 5]. Tutaj podamy jedynie prosty przykład, w którym (4.2.4) nie zachodzi. Niech $\theta^* = (1, 1, 0)$ oraz b będzie odpowiednio duże, powiedzmy $b^2 > 1/4$. Załóżmy, że

$$H = \begin{pmatrix} 1 & 0 & b \\ 0 & 1 & b \\ b & b & 1 \end{pmatrix}.$$

Wtedy (4.2.4) jest nie prawdziwe. Zauważmy, że H jest dodatnio określona, gdy $b^2 < 1/2$.

Powiemy, że zmienna losowa ε ma rozkład subgaussowski ze współczynnikiem $\sigma > 0$, o ile dla dowolnego $u \in \mathbb{R}$

$$\mathbb{E} \exp(u\varepsilon) \leqslant \exp(\sigma^2 u^2/2). \tag{4.2.5}$$

Wektor $X_1 \in \mathbb{R}^p$ jest subgaussowski ze współczynnikiem σ, o ile dla każdego $u \in \mathbb{R}^p$ mamy nierówność $\mathbb{E} \exp(u^T X_1) \leqslant \exp(\sigma^2 u^T u/2)$.

Na koniec dla dwóch liczb rzeczywistych a, b oznaczmy $a \vee b = \max(a, b)$ oraz $a \wedge b = \min(a, b)$.

4.3 Wysokowymiarowe modele parametryczne

Niniejsza część odnosi się do prac [A1, A4] oraz [A5]. Proponujemy w nich M-estymatory z karą, które są selekcyjnie zgodne w rozważanych modelach parametrycznych.

4.3.1 Uogólnione modele liniowe

W tej części zakładamy, że wektory cech są deterministyczne, więc losowe są tylko zmienne odpowiedzi. Ponadto rozważane dane spełniają założenia GLM, to znaczy

$$\mathbb{E} Y_i = \nabla \gamma(x_i^T \theta^*), \quad i = 1, \ldots, n, \tag{4.3.1}$$

gdzie $\nabla\gamma$ jest pochodną znanej funkcji γ. GLM zostały wprowadzone w [23]. W (4.3.1) używamy małych liter w odniesieniu do wektorów cech, aby podkreślić, że są one nielosowe. Rozważana tutaj funkcja straty może być przedstawiona następująco

$$\phi(\theta, y, x) = \gamma(x^T\theta) - yx^T\theta + const, \tag{4.3.2}$$

gdzie *const* jest wyrażeniem, które nie zależy od θ i może być pominięte w dalszej analizie. Dwa główne przykłady GLM to: model liniowy z funkcją straty (4.1.4) oraz model logistyczny z funkcją straty (4.1.6). Dodatkowo zakładamy, że scentrowane zmienne odpowiedzi $\varepsilon_i = Y_i - \mathbb{E}Y_i$ mają rozkład subgaussowski z tym samym współczynnikiem σ, zob. (4.2.5).

Pierwszym wynikiem w [A1] jest selekcyjna zgodność progowego LASSO (TL) w wysokowymiarowych GLM. Niech $\tau > 0$ będzie danym progiem. Wtedy estymatorem zbioru T, zwróconym przez TL, jest $\hat{T}_{TL} = \{1 \leqslant j \leqslant p : |\hat{\theta}_j| > \tau\}$, gdzie $\hat{\theta}$ jest estymatorem LASSO z (4.1.9). Zauważmy, że w [A1] kolumny macierzy $\mathbf{X}$ są znormalizowane $|x_{\cdot j}|_2 = 1$. Chcąc ujednolicić notację z innymi publikacjami, założymy, że $|x_{\cdot j}|_2 = \sqrt{n}$.

Twierdzenie 1 (A1, Twierdzenie 1) *Przypuśćmy, że dla liczb $a_1, a_2 \in (0,1)$ mamy*

$$2a_1^{-2}a_2^{-1}\sigma^2\frac{\log p}{n} \leqslant \lambda^2 \leqslant (1+a_1)^{-2}F_\infty^2(a_1)\tau^2 < (1+a_1)^{-2}F_\infty^2(a_1)(\theta^*_{\min})^2/4. \tag{4.3.3}$$

Wtedy

$$\mathbb{P}(\hat{T}_{TL} \neq T) \leqslant 2\exp\Big(-\frac{(1-a_2)a_1^2 n\lambda^2}{2\sigma^2}\Big).$$

W przypadku wysokowymiarowych modeli liniowych analogiczny wynik został otrzymany w Twierdzeniu 8 z [40] przy założeniach, które wydają się *minimalne* dla selekcyjnej zgodności. Zatem Twierdzenie 1 uogólnia [40, Twierdzenie 8] do wysokowymiarowych GLM.

Twierdzenie 1 pozwala konstruktywnie wybrać λ, na przykład można wziąć lewą stronę nierówności (4.3.3). Jednakże drugi parametr procedury jest zwarty w przedziale $\tau \in [(1+a_1)\lambda F_\infty^{-1}(a_1), \theta^*_{\min}/2)$, którego obydwa końce są nieznane, gdyż $F_\infty(a_1)$ oraz $\theta^*_{\min}$ są nieznane. Wobec tego nie wiemy, jak wybrać τ, co sprawia, że algorytm TL jest niekonstruktywny. Podobne wyniki dla algorytmów z niewypukłymi karami można znaleźć w [10, Wniosek 3 i Wniosek 5]. Natomiast rezultaty dla adaptacyjnego LASSO zostały podane w [5, Wniosek 7.7]. Niestety wybór parametrów tych procedur również wymaga znajomości nieznanych parametrów. W [A1] oraz [A4] proponujemy nowe metody, które są konstruktywne.

Algorytm SS (Screening - Selection)

Ta dwukrokowa procedura została przedstawiona w [A1]. W pierwszym kroku (przesiewanie) należy wyznaczyć estymator LASSO $\hat{\theta}$ z parametrem λ oraz uporządkować nierosnąco jego niezerowe współczynniki (w odniesieniu do wartości bezwzględnych) $|\hat{\theta}_{j_1}| \geqslant \ldots \geqslant |\hat{\theta}_{j_s}|$, gdzie $s = |\mathrm{supp}(\hat{\theta})|$. Używając tego porządku, konstruujemy rodzinę zagnieżdżoną $\mathcal{J} = \Big\{\{j_1\}, \{j_1, j_2\}, \ldots, \mathrm{supp}(\hat{\theta})\Big\}$. W drugim kroku (selekcja) wybieramy model minimalizujący uogólnione kryterium informacyjne (z ang. *Generalized Information Criterion,* GIC) z parametrem $\lambda^2/2$ w rodzinie $\mathcal{J}$, mianowicie

$$\hat{T}_{SS} = \arg\min_{J\in\mathcal{J}} \Big\{\bar{Q}_J + \lambda^2/2|J|\Big\},$$

gdzie $\bar{Q}_J = \bar{Q}(\hat{\theta}_J)$ oraz $\hat{\theta}_J = \arg\min_{\theta_J} \bar{Q}(\theta_J)$ jest M-estymatorem wyznaczonym tylko dla $\{x_{\cdot j}, j \in J\}$. Zatem na algorytm SS można patrzeć jak na algorytm TL z adaptacyjnym, opartym na GIC, wyborem progu.

Algorytm SS jest uproszczeniem i uogólnieniem algorytmu *Screening–Ordering–Selection* (SOS) z [26], który został zaproponowany w kontekście normalnych modeli liniowych. My pokazujemy, że

drugi krok (z ang. *ordering*) w algorytmie SOS może być wykonany, korzystając z separowalności LASSO zamiast używania t-statystyk. Jest to obserwacja kluczowa, ponieważ pozwala łatwo dostosować ten nowy algorytm do modeli innych niż liniowe normalne. Naturalnie nowy algorytm jest również obliczeniowo szybszy.

Zanim wykażemy selekcyjną zgodność algorytmu SS, potrzebujemy kilku dodatkowych oznaczeń. Zdefiniujmy

$$\delta_k = \min_{J\subset T,\, |T\setminus J|=k} |(I - G_J)X_T\theta_T^*|_2^2,$$

dla $k = 1, \ldots, t-1$. Przypomnijmy, że $t = |T|$ oraz $G_J = \mathbf{X}_J(\mathbf{X}_J^T\mathbf{X}_J)^{-1}\mathbf{X}_J^T$. Przeskalowana odległość Kullbacka-Leiblera (K-L) między T a jego podmodelami została wprowadzona w [30, 31] jako $\delta = \min_{1\leqslant k\leqslant t-1} \delta_k/k$. Różne warianty odległości K–L były używane w analizie zgodności algorytmów selekcyjnych [26, Sekcja 3.1], ale wydaje się, że δ prowadzi do optymalnych wyników [31, Twierdzenie 1]. Ponadto zdefiniujmy kule $\mathbb{B} = \bigcup_{J:J\supset T, r(X_J)=|J|\leqslant \bar{t}} \{\theta_J : |X_J(\theta_J^* - \theta_J)|_2^2 \leqslant \delta_{t-1}\}$, gdzie $r(X_J)$ jest rzędem macierzy X_J oraz $t \leqslant \bar{t} < n \wedge p$. Zatem $\mathbb{B}$ składa się z wektorów *rzadkich*. Ponadto załóżmy, że $\bar{Q}(\cdot)$ jest *ściśle wypukła* w θ^*, co oznacza że istnieje $c \in (0, 1]$ takie, że dla wszystkich $\theta \in \mathbb{B}$ mamy

$$\bar{Q}(\theta) \geqslant \bar{Q}(\theta^*) + (\theta - \theta^*)^T\nabla\bar{Q}(\theta^*) + \frac{c}{2}(\theta^* - \theta)^T X^T X/n(\theta^* - \theta). \tag{4.3.4}$$

Co więcej, dla danego $1/2 < a_1 < 1$ definiujemy $a_2 = 1 - (1 - \log(1 - a_1))(1 - a_1)$, $a_3 = 2 - 1/a_1$ oraz $a_4 = \sqrt{a_1 a_2}$.

Twierdzenie 2 (A1, Twierdzenie 2) *Niech $\varepsilon_i, i = 1, \ldots, n$ będą subgaussowskie z σ oraz $\bar{Q}$ będzie ściśle wypukła w θ^* jak w* (4.3.4)*. Załóżmy, że dla $a_1 \in (1/2, 1)$ mamy*

$$\frac{2\sigma^2 \log p}{a_3 a_2 a_1 cn} \vee \frac{\sigma^2 t}{(1 - a_1)^2 cn} \leqslant \lambda^2 < \frac{c\delta_{t-1}}{16n(\bar{t} - t)} \wedge \frac{c\delta}{n(1 + \sqrt{2(1 - a_1)})^2} \wedge \frac{F_\infty^2(a_4)(\theta_{\min}^*)^2}{4(1 + a_4)^2}.$$

Wtedy

$$\mathbb{P}(\hat{T}_{SS} \neq T) \leqslant 4.5 \exp\Big(- \frac{a_2(1 - a_1)cn\lambda^2}{2\sigma^2}\Big).$$

Załóżmy, że $t = o(\log p)$. Twierdzenie 2 stanowi, że λ może być wybrana jako

$$\lambda = \sqrt{2\sigma^2 \log p/n}(1 + o(1)) \tag{4.3.5}$$

dla subgaussowskich modeli liniowych oraz

$$\lambda = \sqrt{\log p/(2cn)}(1 + o(1)) \tag{4.3.6}$$

dla modeli logistycznych. Rozważmy subgaussowski model liniowy i przypuśćmy, że σ^2 jest znane, co jest powszechnym założeniem w literaturze badającej własności procedur selekcyjnych [40, 5, 10]. Wtedy Twierdzenie 2 pozwala wybrać λ konstruktywnie jak w (4.3.5). Niestety algorytm SS przestaje być kostruktywny, gdy opuszczamy subgaussowskie modele liniowe, co widać w (4.3.6). Pomimo tego założenia algorytmu SS są i tak słabsze niż jego konkurentów, zob. [A1, Uwaga 2].

Zgodność selekcyjna w Twierdzeniu 2 jest wykazana dla GLM. Wynik ten można poprawić, jeśli rozważymy tylko subgaussowskie modele liniowe. W [A1, Twierdzenie 3] pokazujemy, że w takiej sytuacji λ może być wybrana jak w (4.3.5) bez dodatkowego założenia $t = o(\log p)$.

W klasycznym (niskowymiarowym) przypadku selekcyjna zgodność jest wykazana dla bayesowskiego kryterium informacyjnego [29], które jest podobne do drugiego kroku algorytmu SS, ale λ^2 zachowuje się jak $\log n/n$. Zatem w [A1] uzasadniamy użyteczność kryteriów informacyjnych również w przypadku wysokowymiarowym. Ważną różnicą jest fakt, że w sytuacji wysokowymiarowej parametr λ^2 powinien być proporcjonalny do $\log p/n$, a nie do $\log n/n$.

Nasza analiza teoretyczna algorytmu SS ogranicza się do przypadku, gdy używamy LASSO tylko z jednym parametrem λ. Jednakże praktyczne implementacje LASSO potrafią efektywnie wyznaczać wartości estymatorów nawet dla siatki parametrów, jak w pakiecie "glmnet" [13]. Zatem zaproponowaliśmy również "siatkową" modyfikację algorytmu SS, którą nazwaliśmy "SSnet". W pierwszym kroku wyznaczamy LASSO dla siatki parametrów λ. Następnie końcowy model jest wybierany, używając GIC, podobnie jak w algorytmie SS. Symulacje przedstawione w [A1] sugerują, że predykcyjnie optymalny parametr λ dla normalnych modeli liniowych wynosi $\sqrt{2.5\sigma^2 \log p/n}$, co jest bliskie wartości w (4.3.5). Dla modelu logistycznego wartość ta wynosi $\sqrt{2 \log p/n}$.

Algorytm LassoSD

Najbardziej czasochłonnym krokiem w algorytmie SS jest LASSO. Jednakże krok selekcyjny również może być złożony obliczeniowo, gdyż musimy wyznaczyć M-estymator dla każdego elementu rodziny $\mathcal{J}$. W [A4] zaproponowaliśmy procedurę, która wymaga policzenia tylko jednego M-estymatora w kroku selekcyjnym. Nazwaliśmy ją "LassoSD", gdyż łączy ona LASSO ze *zstępującym wielokrotnym testowaniem hipotez* (z ang. **S**tep-**D**own multiple testing). LassoSD jest selekcyjnie zgodne i konstruktywne w wysokowymiarowym subgaussowskim modelu liniowym, podobnie jak algorytm SS.

W algorytmie LassoSD rozpoczynamy od wyznaczenia LASSO z parametrem λ i wybrania tylko tych współczynników LASSO, które są większe niż *znany* próg δ. Załóżmy, że zbiór zawierający te współczynniki oznaczymy jako S oraz $s = |S|$. Następnie wyznaczamy ENK ograniczony do $\{x_{\cdot j} : j \in S\}$ oraz odpowiednie t-statystyki $\{t_j : j \in S\}$. Później sortujemy te statystyki nierosnąco względem ich kwadratów, to znaczy $t^2_{[1]} \geqslant t^2_{[2]} \geqslant \ldots \geqslant t^2_{[s]}$, w szczególności $|t_{[1]}| = \max(|t_1|, \ldots, |t_s|)$. Niech $\gamma_1 \geqslant \ldots \geqslant \gamma_s > 0$ będzie ustalonym ciągiem progów. Estymator $\hat{T}_{LassoSD}$ prawdziwego zbioru T jest konstruowany następująco: jeśli $t^2_{[1]} < \gamma^2_1$, to $\hat{T}_{LassoSD}$ jest zbiorem pustym. W przeciwnym przypadku szukamy największego r spełniającego

$$t^2_{[1]} \geqslant \gamma^2_1, \ldots, t^2_{[r]} \geqslant \gamma^2_r$$

oraz $\hat{T}_{LassoSD}$ składa się z r cech, które odpowiadają uporządkowanym statystykom $t_{[1]}, t_{[2]}, \ldots, t_{[r]}$. Zatem na LassoSD można patrzeć jak na algorytm TL z adaptacyjnym wyborem progu. Ten wybór motywowany jest metodami wielokrotnego testowania.

W [A4] proponujemy, aby progi $\gamma_1 \geqslant \ldots \geqslant \gamma_s$ wybierać na dwa sposoby, mianowicie dla $\alpha > 0$ oraz $j = 1, ..., s$ rozważamy

a) procedurę z równymi progami

$$\gamma_j = \sqrt{2 \log \binom{p}{s} - 2 \log \alpha}\,, \tag{4.3.7}$$

b) procedurę z malejącymi progami

$$\gamma_j = \sqrt{2 \log \binom{p+1-j}{s+1-j} - 2 \log \alpha}\,. \tag{4.3.8}$$

Dowodzimy, że LassoSD z tymi progami jest selekcyjnie zgodne.

Twierdzenie 3 (A4, Twierdzenie 3.1) *Ustalmy $a \in (0,1), b > 0, \xi > 1$ oraz $\alpha \geqslant \frac{1}{p}$. Rozważmy algorytm LassoSD z parametrem λ spełniajacym*

$$\frac{\sigma(\xi+1)}{\sqrt{a}(\xi-1)}\sqrt{\frac{2\log(p)}{n}} \leqslant \lambda \leqslant \frac{(\xi+1)\theta^*_{\min} F_\infty(\xi)}{2\xi + (\xi+1) b F_\infty(\xi)}\,, \tag{4.3.9}$$

$\delta = b\lambda$ oraz progami (4.3.7) *bądź* (4.3.8)*. Przypuśćmy, że*

$$\frac{(\theta^*_{\min})^2}{\sigma^2 m} \geqslant 16 \log(p^K/K!),$$

gdzie $K = t\left(1 + \frac{2\xi}{(\xi+1)bF_1(\xi)}\right) \leqslant n$ *oraz* $m = \max_{k=0,\ldots,K-t} \max_{J\subset T^c:|J|=k} \max_{j\in T} m_j(T \cup J)$ *dla* $m_j(J) = [(\mathbf{X}_J^T\mathbf{X}_J)^{-1}]_{jj}, j \in J$. *Wtedy* $\mathbb{P}(\hat{T}_{LassoSD} \neq T)$ *może być oszacowane z góry przez*

$$2\exp\left(-\frac{(1-a)(\xi-1)^2\lambda^2 n}{2\sigma^2(\xi+1)^2}\right) + \alpha(K-t+1)\left[t\left(\frac{K}{p-t+1}\right)^t + K - t\right].$$

Załóżmy jak poprzednio, że σ^2 jest znane. Wtedy Twierdzenie 3 pokazuje, jak kontruktywnie wybrać parametry w LassoSD. Mianowicie możemy wziąć $\lambda^2 = 2\sigma^2 \log p/n(1+o(1))$ z lewej strony nierówności (4.3.9). Ponadto wybierzmy $\delta = b\lambda$ dla dowolnego $b > 0$ oraz $\alpha \geqslant 1/p$. Wybór b jest dowolny, a jego konsekwencje są pokazane w Twierdzeniu 3: zwiększając b, zmniejszamy K, które jest górnym ograniczeniem $|S|$. Jednakże również zmniejszamy mianownik prawej strony nierówności (4.3.9), co sprawia, że warunek ten staje się bardziej restrykcyjny.

Różnice między LassoSD a SS dotyczą kroku selekcyjnego obu algorytmów. SS jest oparty na GIC, więc pracuje *grupowo*, to znaczy mamy wyznaczyć M-estymatory dla każdego elementu zagnieżdżonej rodziny $\mathcal{J}$, a rodzina ta zawiera względnie niewiele *podzbiorów* zbioru $\{1,\ldots,p\}$. LassoSD oparte jest na wielokrotnym testowaniu i pracuje bardziej *indywidualnie*. Mianowicie porównujemy t-statystyki z odpowiednimi progami. Naturalnie podejście grupowe powinno mieć lepsze własności statystyczne, a podejście indywidualne powinno być obliczeniowo szybsze. W pracy [A4] zaobserwowaliśmy, że założenia LassoSD dające zgodność selekcyjną są mocniejsze niż założenia algorytmu SOS z [26]. Zatem są one również bardziej restrykcyjne niż założenia SS. W części eksperymentalnej pracy [A4] zauważyliśmy także, że LassoSD było słabsze w selekcji cech oraz predykcji. Jednakże przewaga (w teorii i praktyce) procedur opartych na GIC była relatywnie mała. Ponadto nie zaobserwowaliśmy, aby krok selekcyjny w LassoSD był obliczeniowo szybszy. Jest to związane z faktem, że pracowaliśmy z normalnymi modelami liniowymi. W tym przypadku krok selekcyjny algorytmu SOS (SS, odpowiednio) wymaga tylko dwóch (jednego, odpowiednio) rozkładu QR macierzy cech ograniczonej do nośnika LASSO, zob. [26, page 967]. Ten zabieg nie może być użyty w bardziej złożonych modelach, na przykład regresji logistycznej czy modelu liniowym z błędami o rozkładach z *ciężkimi ogonami*. Zatem krok selekcyjny w LassoSD byłby obliczeniowo szybszy w tych przypadkach.

4.3.2 Model Isinga

Problemem, na którym skupimy się w tej części, jest poszukiwanie związków między zmiennymi losowymi. Zagadnienie to jest odgrywa ważną rolę, na przykład, w biologii, genetyce czy fizyce. W tym celu często używa się pól losowych Markowa (PLM), czyli grafów nieskierowanych (V, E), gdzie $V = \{1,\ldots,p\}$ jest zbiorem wierzchołków, a $E \subset V \times V$ jest zbiorem krawędzi. Rozważmy wektor losowy $Y = (Y(1),\ldots,Y(p))$, w którym zmienna losowa $Y(s)$ jest stowarzyszona z wierzchołkiem $s \in V$. Warunkowa niezależność pomiędzy współrzędnymi (lub podzbiorami) wektora Y jest opisywana przez strukturę tego grafu. Zatem w tym przypadku wybór modelu odnosi się do poszukiwania *istniejących krawędzi* w grafie, a *wysoki wymiar* dotyczy sytuacji, w której liczba wierzchołków może być porównywalna (bądź większa) z rozmiarem próbki. Natomiast *rzadkość* oznacza, że wiemy (lub wierzymy), że liczba istniejących krawędzi jest relatywnie mała, gdy porównamy ją z liczbą wszystkich możliwych krawędzi $\frac{p(p-1)}{2}$ lub rozmiarem próbki n.

W pracy [A5] rozważamy dyskretne PLM, a dokładniej pracujemy w sytuacji, gdy $Y(s) \in \{-1, 1\}$. Jednym z najbardziej popularnych modeli matematycznych używanych w tym zagadnieniu jest model Isinga [18]. Łączny rozkład wektora Y w tym przypadku dany jest wzorem

$$p(y|\theta^*) = \frac{1}{C(\theta^*)}\exp\left(\sum_{r<s}\theta^*_{rs}y(r)y(s)\right), \tag{4.3.10}$$

gdzie suma w (4.3.10) odnosi się do wszystkich par indeksów $(r,s) \in \{1,\dots,p\}^2$ takich, że $r < s$. Wektor $\theta^* \in \mathbb{R}^{p(p-1)/2}$ jest prawdziwym parametrem, a $C(\theta^*) = \sum_{y\in\{-1,1\}^p} \exp\left(\sum_{r<s} \theta^*_{rs} y(r) y(s)\right)$ jest stałą normującą. Kluczową własnością modelu Isinga jest następujący fakt: wierzchołki r oraz s nie są połączone krawędzią (to znaczy $\theta^*_{rs} = 0$), o ile zmienne $Y(r)$ oraz $Y(s)$ są warunkowo niezależne przy ustalonych wartościach we wszystkich pozostałych wierzchołkach. Zatem odgadnięcie struktury grafu jest równoważne estymacji parametru θ^*. Jednakże model Isinga ma również istotną wadę, mianowicie stała normująca jest sumą składającą się z 2^p elementów, co sprawia, że jest ona trudna do wyznaczenia nawet wtedy, gdy p jest względnie małe.

W [A5] rozważamy wysokowymiarowy model Isinga, więc, oprócz trudnej stałej normującej, mamy jeszcze dodatkowy problem. Mianowicie liczba wierzchołków może być większa niż rozmiar danych. Proponujemy, aby te dwie trudności przezwyciężyć, używając metod Monte Carlo opartych na łańcuchach Markowa (z ang. Markov chain Monte Carlo, MCMC) połączonych z M-estymatorami z karą. Niech $Y_1,\dots,Y_n$ będą niezależnymi wektorami z (4.3.10), a $Y^1,\dots,Y^m$ będzie nowym zbiorem danych, który jest łańcuchem Markowa z rozkładem stacjonarnym o gęstości h. W szczególności ten nowy zbiór danych zawiera zmienne zależne. Nasze podejście polega na penalizowanej minimalizacji aproksymacji ujemnej wiarodności w oparciu o metody MCMC, to znaczy będziemy minimalizować

$$-\frac{1}{n}\sum_{i=1}^{n} \theta^T J(Y_i) + \log\left(\frac{1}{m}\sum_{k=1}^{m} \frac{\exp\left[\theta^T J(Y^k)\right]}{h(Y^k)}\right) + \lambda|\theta|_1, \tag{4.3.11}$$

gdzie $J(y) = (y(r)y(s))_{r<s}$ oraz $|\theta|_1 = \sum_{r<s}|\theta_{rs}|$ jest l_1-normą wektora θ. Zauważmy, że (4.3.11) jest funkcją wypukłą. Oznaczmy element ją minimalizujący przez $\hat{\theta}$.

Zanim sformułujemy główny wynik z [A5], potrzebujemy kilku dodatkowych oznaczeń: niech $\bar{p} = p(p-1)/2$, jak poprzednio T to prawdziwy zbiór, który tutaj oznacza $T = \{(r,s) : \theta^*_{rs} \neq 0\}$. Ponadto $\theta^*_{\min} = \min_{(r,s)\in T}|\theta^*_{rs}|$ oraz $t = |T|$. Dla $\xi > 1$ przez $K_1(\xi), K_2(\xi), \dots$ oznaczamy liczby, które zależą jedynie ξ. Co więcej, $Y^1,\dots,Y^m$ jest łańcuchem Markowa na $\{-1,1\}^p$, który jest generowany przez próbnik Gibbsa. Jego gęstość początkowa jest oznaczona jako q, podczas gdy rozkład stacjonarny to h. Wypiszmy teraz ważne wielkości charakteryzujące ten łańcuch $\beta_1 = \sqrt{\sum_{y\in\{-1,1\}^p} \frac{q^2(y)}{h(y)}}$, $\beta_2 = \frac{1-\kappa}{1+\kappa}$, $M = \max_{y\in\{-1,1\}^p} \frac{\exp((\theta^*)^T J(y))}{h(y)C(\theta^*)}$, gdzie $1-\kappa$ jest luką spektralną. Krótko mówiąc, te trzy wartości można postrzegać nastepująco: β_1 – jak blisko jest gęstość początkowa gęstości stacjonarnej, β_2 – „mieszanie" łańcucha, M – jak blisko jest gęstość h gęstości prawdziwej (4.3.10).

Rozważmy progowe LASSO, to znaczy $\hat{\theta}$, które minimalizuje (4.3.11), z progiem $\tau > 0$. Niech $\hat{T} = \{(r,s) : |\hat{\theta}_{rs}| > \tau\}$. W kolejnym wyniku udowadniamy zgodną selekcję zaproponowanego estymatora. Jest to połączenie [A5, Twiedzenie 2] z [A5, Wniosek 3]. W wersji przedstawionej poniżej rezultat ten może być łatwo porównany ze zgodną selekcją estymatora TL w GLM (Twierdzenie 1).

Twierdzenie 4 *Niech $\varepsilon > 0, \xi > 1$. Załóżmy, że $n \geqslant \frac{K_1(\xi)\, t^2 \log(\bar{p}/\varepsilon)}{F^2_\infty(\xi)}$, $m \geqslant \frac{K_2(\xi)\, t^2\, M^2 \log(\bar{p}\beta_1/\varepsilon)}{F^2_\infty(\xi)\beta_2}$ oraz*

$$K_3(\xi)\max\left[\log(\bar{p}/\varepsilon)/n,\ M^2\log\left(\bar{p}\beta_1/\varepsilon\right)/(m\beta_2)\right] \leqslant \lambda^2 \leqslant K_4(\xi)F^2_\infty(\xi)\tau^2 \leqslant K_4(\xi)F^2_\infty(\xi)(\theta^*_{\min})^2/4 \tag{4.3.12}$$

Wtedy mamy $\mathbb{P}(\hat{T} \neq T) \leqslant 4\varepsilon$.

Twierdzenie 4 wskazuje warunki wystarczające zgodnej selekcji naszego estymatora. Dokładne omówienie tych warunków oraz ich porównanie z innymi algorytmami znajduje się w [A5, strony 8-10]. Krótko mówiąc, procedura rozpoznaje prawdziwy model, jeśli rozmiary danych początkowych i próby MCMC są wystarczająco duże, model jest rzadki, niezerowe współczynniki w θ^* odpowiednio duże, a CIF nie jest zbyt bliski zeru. Oczywistą zaletą naszego algorytmu jest fakt, że podejście oparte na MCMC pozwala przybliżać stałą normującą z dowolną precyzją. Błędy aproksymacyjne innych

metod zależą od nieznanej struktury grafu i nie mogą być łatwo poprawione. W naszym podejściu, aby poprawić aproksymację, wystarczy zwiększyć liczność próby MCMC. Naturalną wadą naszego algorytmu jest konieczność wykonania dodatkowych symulacji. Sprawia to, że nasza procedura jest obliczeniowo bardziej złożona, ale jednocześnie pozwala to lepiej znajdować prawdziwy model.

Porównajmy teraz Twierdzenie 4 z Twierdzeniem 1, w którym wykazaliśmy zgodną selekcję progowego LASSO w GLM. Warunki dotyczące λ oraz τ, podane w (4.3.3) oraz (4.3.12), są podobne. Oczywiście w Twierdzeniu 4 liczba estymowanych parametrów to $\bar{p} = p(p-1)/2$, a także musimy wziąć pod uwagę dodatkową próbę MCMC.

Porównując GLM z podejściem MCMC w modelu Isinga, możemy zauważyć istotną różnicę, która sprawia, że analiza tego drugiego przypadku jest znacznie trudniejsza. Mianowicie w GLM ryzyko empiryczne można rozłożyć na dwa składniki: pierwszy z nich jest liniowy i losowy, a drugi nieliniowy i deterministyczny, zob. (4.3.2) lub [A1, wzór (10)]. W tej sekcji mamy podobny rozkład, ale składnik nieliniowy również jest losowy, por. drugie wyrażenie w (4.3.11). Dlatego musieliśmy zmodyfikować CIF, a także pokonać kilka dodatkowych trudności, na przykład fakt, że CIF stał się losowy czy rozważyć zdarzenie Ω_2 w [A5, Lemat 6]. Konsekwencją jest pojawienie się dwóch dodatkowych założeń dotyczących n oraz m w Twierdzeniu 4.

Naturalnie wybór progu τ podany w Twierdzeniu 4 nie jest kostruktywny. W implementacji naszego algorytmu w [A5, Sekcja 4] pokonaliśmy ten problem w sposób analogiczny jak w [A1], to znaczy wybraliśmy próg adaptacyjnie w oparciu o GIC.

4.4 Wysokowymiarowe modele semiparametryczne

W Sekcji 4.3 rozważamy modele parametryczne. Są one dosyć "regularne", na przykład w GLM zakłada się znajomość funkcji odpowiedzi $\nabla\gamma$ oraz używa się jej w procesie estymacji. Ponadto wymaga się, aby błędy ε były subgaussowskie, czyli "ogony" ich rozkładów mają się zbliżać do zera w tempie wykładniczym. Jednakże często zdarza się, że badany zbiór danych nie spełnia tych założeń bądź są one trudne do sprawdzenia.

Dlatego teraz będziemy badać własności M-estymatorów z karą, które pracują w mniej regularnych sytuacjach, mianowicie

$$Y_i = g(X_i^T\theta^*, \varepsilon_i), \quad i = 1, \ldots, n, \tag{4.4.1}$$

gdzie g jest *nieznaną* funkcją odpowiedzi, ε_i jest błędem losowym, a $X_i \in \mathbb{R}^p$ jest wektorem cech, który jest losowy w tej sekcji. Zatem zakładamy, że cechy wpływają na zmienną odpowiedzi poprzez funkcję g iloczynu skalarnego $X_i^T\theta^*$. Jednakże nie zakładamy nic o postaci funkcji g (poza monotonicznością względem pierwszej zmiennej) ani o rozkładzie błędu ε_i. W szczególności nie wymagamy istnienia momentów ε_i. Model (4.4.1) jest zwykle nazywany modelem z jednym indeksem (z ang. *single index model*, SIM).

W pracach [A2, A3] zakładamy, że $\mathbb{E}X_1 = 0$, $H = \mathbb{E}X_1X_1^T$ jest dodatnio określona oraz $H_{jj} = 1$ dla $j = 1, \ldots, p$.

4.4.1 Estymatory oparte na rangach

Istnieje wiele procedur potrafiących pracować w modelu (4.4.1). Metody te głównie opierają się na *odpornych* funkcjach straty, na przykład funkcji (4.1.5), i będą obiektem naszych badań w kolejnych sekcjach. Jednakże ich użycie w analizie dużych zbiorów danych jest często ograniczone przez ich złożoność obliczeniową bądź potrzebę opracowania specjalnych algorytmów optymalizacyjnych. W pracy [A2] chcieliśmy zaproponować procedurę, która będzie obliczeniowo szybka i odporna w selekcji cech (względem rozkładów błędów oraz nieznanej funkcji g). Nasze podejście jest bardzo proste i opiera się

na zamianie wartości zmiennych odpowiedzi Y_i na ich scentrowane rangi. Rangi R_i definiujemy jako $R_i = \sum_{j=1}^n \mathbb{I}(Y_j \leqslant Y_i), i = 1, \ldots, n$, gdzie $\mathbb{I}(\cdot)$ jest indykatorem. Rozważana funkcja straty

$$\phi(\theta, Y_1, \ldots, Y_n, X_i) = (R_i/n - 0.5 - X_i^T\theta)^2 \tag{4.4.2}$$

jest trochę bardziej skomplikowana niż zwykle używana $\phi(\theta, y_i, x_i)$ w (4.1.2) oraz (4.1.9). Zatem w naszym podejściu nie używamy prawdopodobnie *zaburzonych* zmiennych Y_i, lecz opieramy się jedynie na porządku między nimi. Procedura, nazwana "RankLasso", nie wymaga specjalnie dedykowanych algorytmów i może być wykonana, używając efektywnych implementacji LASSO w „R" [28], na przykład pakiety „lars" [8] czy „glmnet" [13].

Pierwszym problemem, który omawiamy w [A2], jest wskazanie związku między nośnikiem θ^* (czyli zbiorem T) a nośnikiem $\theta_0 = \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}\,\bar{Q}(\theta)$, który jest parametrem estymowanym przez RankLasso. Funkcja $\bar{Q}(\cdot)$ jest jak w (4.1.2) ze stratą (4.4.2).

Twierdzenie 5 (A2, Twierdzenie 1) *Zakładamy, że rozkład X_1 jest absolutnie ciągły oraz X_1 jest niezależne od ε_1. Przypuśćmy ponadto, że dla każdego $\theta \in \mathbb{R}^p$ istnieje $d_\theta \in \mathbb{R}$ takie, że*

$$\mathbb{E}(\theta^T X_1 | \beta^T X_1) = d_\theta \beta^T X_1. \tag{4.4.3}$$

Ponadto niech dystrybuanta F zmiennej odpowiedzi Y_1 będzie rosnąca oraz g w (4.4.1) będzie rosnąca względem pierwszego argumentu. Wtedy istnieje dodatnia wartość γ_{θ^*} taka, że $\theta_0 = \gamma_{\theta^*}\theta^*$.

Twierdzenie 5 stanowi, że $\{j : \theta_j^* \neq 0\} = \{j : (\theta_0)_j \neq 0\}$, więc RankLasso może być pomocne w znalezieniu zbioru T. Jednakże, jak już wspomnieliśmy wcześniej, selekcyjna zgodność tej procedury wymaga restrykcyjnego warunku IRR. Zatem w [A2] badamy progową i ważoną wersję RankLasso, a także uzasadniamy ich zgodność w selekcji cech przy słabych założeniach. Są one bardzo podobne do warunków dla "standardowego" progowego czy ważonego LASSO podanych w Twierdzeniu 1 czy [40, 5, 17]. Oznaczmy te modyfikacje jako "RankTL" i "RankW", odpowiednio.

Kolejny wynik jest połączeniem [A2, Twierdzenie 2] z [A2, Twierdzenie 5]. Ten drugi rezultat w pracy [A2] jest podany w uproszczonej wersji. Tutaj przedstawiamy go w ogólnej formie, co pozwoli porównać go z Twierdzeniem 1.

Twierdzenie 6 *Niech $a \in (0, 1)$ oraz $\xi > 1$ będą ustalone. Przypuśćmy, że założenia Twierdzenia 5 są spełnione, wektor $(X_1)_T$ jest subgaussowski z σ_0, a zmienne $(X_1)_j$ są subgaussowskie z σ_j dla $j \notin T$ oraz $\sigma = \max(\sigma_0, \sigma_j, j \notin T)$. Ponadto przypuśćmy, że $n \geqslant \frac{K_1 t^2 \sigma^4 (1+\xi)^2 \log(p/a)}{F_\infty^2(\xi)}$ oraz*

$$\frac{K_2(\xi+1)^2\sigma^4\log(p/a)}{(\xi-1)^2\kappa n} \leqslant \lambda^2 \leqslant \frac{(\xi+1)^2 F_\infty^2(\xi)\tau^2}{16\xi^2} \leqslant \frac{(\xi+1)^2 F_\infty^2(\xi)(\theta_0^{\min})^2}{64\xi^2} \tag{4.4.4}$$

gdzie K_1, K_2 są stałymi, κ jest namniejszą wartością własną macierzy $H_T = (H_{jk})_{j,k\in T}$, a także $\theta_0^{\min} = \min_{j\in T} |(\theta_0)_j|$. Wtedy istnieje stała K_3 taka, że $\mathbb{P}(\hat{T}_{RankTL} \neq T) \leqslant K_3 a$. Jeśli X_1 ma rozkład normalny $N(0, H)$, to możemy opuścić κ oraz σ w powyższych wyrażeniach.

Twierdzenie 6 podaje warunki wystarczające zgodnej selekcji estymatora RankTL. Przypomnijmy, że nie nakładamy żadnych restrykcji na rozkłady błędów, a także funkcja g jest nieznana. Porównajmy Twierdzenie 6 z Twierdzeniem 1, które pracuje przy mocniejszych założeniach wymaganych w GLM. Założenia (4.4.3) oraz fakt, że g jest rosnąca, których potrzebujemy w Twierdzeniu 5, nie są wymagane w Twierdzeniu 1. W Twierdzeniu 6 dodatkowo zakładamy, że wektory cech są subgaussowskie oraz rozmiar danych jest odpowiednio duży. W Twierdzeniu 1 nie wymagamy tych dwóch warunków, gdyż pracujemy w nim z cechami deterministycznymi. Moim zdaniem kluczową różnicą jest występowanie $\theta_0^{\min}$ w (4.4.4), gdy w analogicznym warunku (4.3.3) mamy $\theta_{\min}^*$. Przypomnijmy, że

$\theta_0 = \gamma_{\theta^*}\theta^*$ oraz γ_{θ^*} jest zwykle znacząco mniejsze niż jeden. Zatem (4.4.4) staje się bardziej restrykcyjne niż (4.3.3). Można na to spojrzeć jak na cenę, którą musimy zapłacić za pracę w znaczniej mniej regularnym przypadku niż GLM. Krótko mówiąc, różnica ta polega na konieczności użycia większego rozmiaru próby przy pracy z procedurami opartymi na rangach. Obserwacja ta jest powszechna wśród estymatorów w modelu (4.4.1). Na przykład podobne restrykcje pojawiają się w przypadku estymatora LAD z LASSO.

W [A2] rozważamy także drugą modyfikację RankLasso, czyli ważone RankLasso. W tym przypadku minimalizujemy

$$\bar{Q}(\theta) + \lambda_a \sum_{j=1}^{p} w_j |\theta_j|, \tag{4.4.5}$$

gdzie $\lambda_a > 0$ oraz wagi są wybrane następująco: dla dowolnej liczby $K > 0$ oraz estymatora RankLasso $\hat{\theta}$ połóżmy $w_j = |\hat{\theta}_j|^{-1}$ dla $|\hat{\theta}_j| \leqslant \lambda_a$ i $w_j = K$ w przeciwnym przypadku.

Kolejny wynik jest połączeniem [A2, Twierdzenie 2] z [A2, Twierdzenie 7]. Prezentujemy go w ogólniejszej wersji niż w [A2], podobnie jak to miało miejsce w przypadku Twierdzenia 6.

Twierdzenie 7 *Niech* $a \in (0,1), \xi > 1$ *będą ustalone oraz* $\lambda_a := \frac{4\xi\lambda}{(\xi+1)F_\infty(\xi)}, K \leqslant F_\infty(\xi)$. *Przypuśćmy, że założenia Twierdzenia 5 są spełnione, wektor* $(X_1)_T$ *jest subgaussowski z* σ_0, *zmienne* $(X_1)_j$ *są subgaussowskie z* σ_j *dla* $j \notin T$, *a także* $\sigma = \max(\sigma_0, \sigma_j, j \notin T)$. *Ponadto załóżmy, że* $n \geqslant \frac{K_1 t^2 \sigma^4 (1+\xi)^2 \log(p/a)}{F_\infty^2(\xi)}$ *oraz*

$$\frac{K_2(\xi+1)^2\sigma^4 \log(p/a)}{(\xi-1)^2 \kappa n} \leqslant \lambda^2 \leqslant K_3(\xi) \min\left[(\theta_0^{\min})^2 F_\infty^2(\xi), \kappa^2/t^2\right]$$

gdzie K_1, K_2 *są stałymi,* $K_3(\xi)$ *jest liczbą zależącą tylko od* ξ, κ *jest najmniejszą wartością własną macierzy* $H_T = (H_{jk})_{j,k\in T}$ *oraz* $\theta_0^{\min} = \min_{j\in T} |(\theta_0)_j|$. *Wtedy z prawdopodobieństwem przynajmniej* $1 - K_4 a$ *istnieje element* $\hat{\theta}^a$ *minimalizujący funkcję* (4.4.5) *taki, że* $\hat{\theta}^a_{T^c} = 0$ *oraz*

$$|\hat{\theta}^a_T - (\theta_0)_T|_1 \leqslant \frac{K_5(\xi) t \lambda}{\kappa}\,, \tag{4.4.6}$$

gdzie K_4 *jest stałą, a* $K_5(\xi)$ *zależy tylko od* ξ.

Twierdzenie 7 potwierdza fakt, że ważone RankLasso jest lepsze niż RankLasso. Istotnie ważone RankLasso nie potrzebuje warunku IRR, aby znaleźć prawdziwy model. Ponadto używając [A2, Twierdzenie 2], otrzymujemy oszacowanie błędu w estymacji w l_1-normie dla RankLasso, które jest podobne do (4.4.6). Jednakże oszacowanie to będzie zależało od $F_1(\xi)$, ktore odnosi się do całej macierzy cech H. Natomiast κ z (4.4.6) dotyczy tylko podmacierzy H_T. Nasz model jest rzadki, więc należy oczekiwać, że κ jest (znacznie) mniejsze niż $F_1(\xi)$.

Zauważmy, że ważone RankLasso jest procedurą niekonstruktywną, gdyż jego wagi zależą od nieznanych parametrów. W części eksperymentalnej w [A2] wskazaliśmy wybór wag, który jest konstruktywny i działa w sposób satysfakcjonujący. Ponadto wykazaliśmy wyższość modyfikacji RankLasso nad jego wersją pierwotną w sytuacji, gdy cechy są zależne. Nasze badania numeryczne pokazują także, że RankLasso działa znacząco lepiej niż LADLasso, które jest popularną metodą w odpornej selekcji cech [3, 9].

Dowody własności RankLasso są trudniejsze niż standardowego LASSO, gdyż tutaj ryzyko empiryczne $\bar{Q}(\theta)$ nie jest sumą niezależnych zmiennych losowych. Problem ten pokonaliśmy, używając teorii U-statystyk. W szczególności w [A2, Lemat 14] udowodniliśmy nierówność wykładniczą dla U-statystyk. Wynik ten jest interesujący sam w sobie i może być wykorzystany w innych problemach. Co ważne w [A2, Lemat 14] nie wymagamy, aby jądra U-statystyk były ograniczone (w odróżnieniu od [7, Twierdzenie 4.1.13]).

4.4.2 Algorytm Screening-Selection poza GLM

W Sekcji 4.3 badaliśmy selekcyjną zgodność algorytmu SS w GLM. W [A1, Sekcja 3] pokazujemy, że procedura ta jest elastyczna i może być skutecznie używana również poza GLM. Na przykład w modelu liniowym z błędami niekoniecznie subgaussowskimi albo w binarnej klasyfikacji z niekoniecznie logistyczną funkcją straty.

Wypukłość funkcji straty ϕ wciąż jest głównym założeniem, ale, podobnie jak w Sekcji 4.3, potrzebujemy również *ścisłej wypukłości* funkcji ryzyka $Q(\theta)$ w θ^*. Własność ta jest zdefiniowana jak w (4.3.4), ale $\bar{Q}(\cdot)$ jest zastąpione przez $Q(\cdot)$. Co więcej, będziemy również wymagać ścisłej wypukłości z l_2-kulą $\mathbb{B}$ podmienioną przez l_1-kulę $B_1(r) = \{\theta : |\theta - \theta^*|_1 \leqslant r\}$ z $r = \theta^*_{\min}$. Oznaczmy odpowiednie współczynniki c z (4.3.4) jako c_2 oraz c_1, odpowiednio. Zauważmy, że różniczkowalność i ścisła wypukłość odnosi się do ryzyka $Q(\theta) = \mathbb{E}\bar{Q}(\theta)$, a nie ryzyka empirycznego $\bar{Q}(\theta)$. Zatem własności te mogą zachodzić nawet wtedy, gdy funkcja straty ϕ jest nieróżniczkowalna, na przykład dla wartości bezwzględnej bądź kwadratowej perceptronowej, zob. [A1, Uwaga 6].

W naszych rozważaniach ważną rolę odgrywają dwa procesy empiryczne:

$$Z(r) = \sup_{\theta \in B_1(r)} \left|\bar{Q}(\theta) - Q(\theta) - [\bar{Q}(\theta^*) - Q(\theta^*)]\right| \tag{4.4.7}$$

oraz

$$U_J(r) = \sup_{\theta \in B_{2,J}(r)} \left|\bar{Q}(\theta) - Q(\theta) - [\bar{Q}(\theta^*) - Q(\theta^*)]\right|, \tag{4.4.8}$$

gdzie $B_{2,J}(r) = \{\theta_J : |X_J(\theta^* - \theta_J)|_2^2 \leqslant r^2\}$ jest kulą o promieniu $r > 0$, a J jest (rzadkim) podzbiorem $\{1, \ldots, p\}$ takim, że $T \subset J, r(X_J) = |J| \leqslant \bar{t}$. Zamiast nierówności wykładniczych dla wektorów subgaussowskich z [A1, Lemat 2], potrzebujemy wykładniczych oszacowań dla procesów (4.4.7) oraz (4.4.8) następującej postaci: istnieje $L > 0$ oraz stałe $K_1, K_2 > 0$ takie, że dla każdego $0 < r \leqslant \theta^*_{\min}$ oraz $z \geqslant 1$ mamy

$$P\left(\frac{Z(r)}{r} > K_1 L z \sqrt{\log(2p)/n}\right) \leqslant \exp\left(-K_2 \log(2p)\, z^2\right). \tag{4.4.9}$$

Drugie oszacowanie jest podobne: istnieje $L > 0$ oraz stałe $K_3, K_4 > 0$ takie, że dla każdego $0 < r \leqslant \sqrt{\delta_{t-1}}$ (δ_{t-1} jest zdefiniowane w (4.3.1)), $z \geqslant 1$ oraz J takiego, że $T \subset J, r(X_J) = |J| \leqslant \bar{t}$ mamy

$$P\left(\frac{U_J(r)}{r} > K_3 L z \sqrt{|J|/n}\right) \leqslant \exp(-K_4 |J| z^2). \tag{4.4.10}$$

Teraz napiszmy główny wynik z [A1, Sekcja 3].

Twierdzenie 8 (A1, Twierdzenie 4) *Ustalmy $a_1, a_2 \in (0,1)$ oraz niech K_i będą stałymi. Załóżmy, że $Q(\cdot)$ jest ściśle wypukła w sensie omówionym powyżej oraz* (4.4.9), (4.4.10) *zachodzą. Ponadto przypuśćmy, że*

$$K_1 \max\left(\frac{\log p}{a_1^2}, \frac{\log p}{c_2}, \frac{t}{a_2 c_2}\right)\frac{L^2}{n} \leqslant \lambda^2 \leqslant K_2 \min\left[\frac{c_2\delta}{n(1+\sqrt{2a_2})^2}, \frac{c_2\delta_{t-1}}{n(\bar{t}-t)}, (1-a_1)^2 c_1^2 \kappa_{a_1}^2 (\theta^*_{\min})^2/t\right]. \tag{4.4.11}$$

Wtedy

$$P\left(\hat{T}_{SS} \neq T\right) \leqslant K_3 \exp\left[-\frac{K_4 n \lambda^2}{L^2} \min\left(a_1^2, a_2 c_2\right)\right].$$

W Twierdzeniu 8 szacujemy wykładniczo błąd selekcyjny algorytmu SS. Rozszerzamy zatem Twierdzenie 2 do szerokiej klasy wypukłych funkcji straty. W szczególności funkcje te nie muszą być różniczkowalne jak w regresji kwantylowej czy maszynach wektorów podpierających. Warto zauważyć, że Twierdzenie 8 zastosowane do GLM daje nieznacznie słabsze rezultaty niż Twierdzenie 2, które jest poświęcone tylko GLM. Dokładne porównanie prezentujemy w [A1, Uwagi 7 oraz 8]. Krótko mówiąc, w Twierdzeniu 8 otrzymaliśmy gorsze stałe oraz c_1^2 pojawia się po prawej stronie (4.4.11).

Ważnym założeniem w Twierdzeniu 8 są warunki (4.4.9) oraz (4.4.10). Można pokazać, że są one spełnione, używając narzędzi z teorii procesów empirycznych: nierówności koncentracyjnych [22], lematu o symetryzacji [38, Lemat 2.3.1] czy lematu o kontrakcji [19, Twierdzenie 4.12]. Zdumiewający jest fakt, że do wykazania (4.4.9) bądź (4.4.10) potrzebujemy tylko jednego dodatkowego założenia. Mianowicie wystarczy, że funkcja straty ϕ spełnia warunek Lipschitza z L. Naturalnie strata logistyczna i wartość bezwzględna są funkcjami Lipschitza z $L = 2$ oraz $L = 1$, odpowiednio. Własność ta jest również spełniona w przypadku straty kwadratowej perceptronowej, ale tutaj L będzie zależało od n.

4.4.3 Źle wyspecyfikowana klasyfikacja binarna

W [A3] badamy zachowanie estymatorów w (prawdopodobnie) źle wyspecyfikowanej klasyfikacji wysokowymiarowej. Sytuacja ta często pojawia się w praktyce, na przykład pracując z danych dotyczącymi klasyfikacji, często błędnie zakłada się, że dane te pochodzą z modelu regresji logistycznej. Następnie stosuje się estymator LASSO (albo inny) z logistyczną funkcją straty (4.1.6). Inne podejście do binarnej klasyfikacji, które jest często używane z względu na łatwość obliczeniową, polega na potraktowaniu zmiennych Y_i jakby były liczbami i zastosowaniu estymatora LASSO z kwadratową funkcją straty (4.1.4). Te dwa podejścia czasami dają nadspodziewanie dobre rezultaty w selekcji cech i predykcji, jednakże przyczyny tego zjawiska nie zostały dotychczas zbyt głęboko zbadane.

W klasyfikacji celem jest przewidzenie bądź odgadnięcie nieznanej klasy (etykiety) obiektu na podstawie jego obserwowanych cech. Obiekty opisane są przez wektory losowe (X, Y), gdzie $X \in \mathbb{R}^p$ jest wektorem cech, a $Y \in \{-1, 1\}$ jest etykietą obiektu. Klasyfikator jest to funkcja mierzalna $f : \mathbb{R}^p \to \mathbb{R}$, która wyznacza etykietę obiektu następująco: jeśli $f(x) \geqslant 0$, to przewidujemy, że $y = 1$. W przeciwnym przypadku zgadujemy, że $y = -1$.

Jakość klasyfikatora f mierzona jest jego ryzykiem (prawdopodobieństwem błędnej klasyfikacji)

$$R(f) = P(Y = 1, f(X) < 0) + P(Y = -1, f(X) \geqslant 0). \tag{4.4.12}$$

Niech $\eta(x) = P(Y = 1|X = x)$. Jest jasne, że $f_B(x) = \operatorname{sign}(2\eta(x) - 1)$ mimimalizuje ryzyko (4.4.12) w rodzinie wszystkich klasyfikatorów. Funkcję tę nazwiemy klasyfikatorem Bayesa, a jego ryzyko oznaczymy jako $R_B = R(f_B)$.

W [A3] rozważamy liniowe klasyfikatory $f(x) = f_{\hat{\theta}}(x) = x^T\hat{\theta}$, gdzie $\hat{\theta}$ jest estymatorem LASSO (4.1.9) z funkcją straty ϕ. Będziemy badać, jak blisko może być ryzyko funkcji $f_{\hat{\theta}}$ oraz f_B. Zatem będziemy analizować ryzyko względne funkcji $f_{\hat{\theta}}$ zdefiniowane jako $\mathcal{E}(\hat{\theta}, f_B) = \mathbb{E}_D R(\hat{\theta}) - R_B$, gdzie $\mathbb{E}_D$ jest wartością oczekiwaną względem danych $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Ponadto bedziemy używać $R(\theta)$ zamiast $R(f_\theta)$.

Badamy własności predykcyjne klasyfikatorów, więc będziemy jawnie rozważać *rozszerzone* wersje parametrów i wektorów cech. Mianowicie każdy wektor θ zawiera wyraz wolny, a każdy wektor x zawiera dodatkową współrzędną odnoszącą się do wyrazu wolnego. Ponadto zdefiniujmy θ^*_{quad} oraz θ^*_{log} jako elementy minimalizujące (4.1.1) z kwadratową i logistyczną funkcją straty, odpowiednio. Podobnie definiujemy $\hat{\theta}_{quad}$ oraz $\hat{\theta}_{log}$ jako estymatory LASSO w (4.1.9) z kwadratową i logistyczną funkcją straty, odpowiednio. Ponadto dla $c > 0$ rozważmy zdarzenie $\Omega_{quad} = \{|\hat{\theta}_{quad} - \theta^*_{quad}|_1 \leqslant c\}$. Zdarzenie Ω_{log} jest określone analogicznie.

Twierdzenie 9 (A3, Twierdzenie 2) *Załóżmy, że cechy są subgaussowskie ze współczynnikami σ_j oraz $\sigma = \max_{1 \leqslant j \leqslant p} \sigma_j$. Ponadto zmienna losowa $X^T\theta^*_{quad}$ ($X^T\theta^*_{log}$, odpowiednio) ma gęstość h_{quad} (h_{log}, odpowiednio), która jest ciągła na przedziale $U = [-2\sigma c\sqrt{\log p}, 2\sigma c\sqrt{\log p}\,]$ oraz $\tilde{h}_{quad} = \sup_{u \in U} h_{quad}(u)$,*

$\tilde{h}_{log}$ jest zdefiniowane analogicznie. Wtedy

$$\mathcal{E}(\hat{\theta}_{quad}, f_B) \leqslant 2P(\Omega^c_{quad}) + 4\sigma\tilde{h}_{quad}c\sqrt{\log p} + 2/p + \sqrt{\mathbb{E}\left[2\eta(X) - 1 - X^T\theta^*_{quad}\right]^2}, \quad (4.4.13)$$

$$\mathcal{E}(\hat{\theta}_{log}, f_B) \leqslant 2P(\Omega^c_{log}) + 4\sigma\tilde{h}_{log}c\sqrt{\log p} + 2/p + \sqrt{2\mathbb{E}KL\left[\eta(X), \eta_{log}(X^T\theta^*_{log})\right]}, \quad (4.4.14)$$

gdzie $\eta_{log}(u) = 1/(1+\exp(-u))$ oraz $KL(\cdot,\cdot)$ jest odległością Kullbacka-Leiblera.

Ryzyko względne estymatora $\hat{\theta}_{quad}$ (dla $\hat{\theta}_{log}$ analogicznie) można rozłożyć następująco

$$\mathbb{E}_D R(\hat{\theta}_{quad}) - R(\theta^*_{quad}) \quad + \quad R(\theta^*_{quad}) - R_B. \quad (4.4.15)$$

Drugi składnik w (4.4.15) jest ryzykiem aproksymacji i porównuje zdolność predykcyjną „najlepszego" liniowego klasyfikatora θ^*_{quad} z klasyfikatorem Bayesa. Natomiast pierwsze wyrażenie w (4.4.15) nazywane jest ryzykiem estymacji i opisuje, jak proces estymacjji wpływa na własności predykcyjne klasyfikatorów. Oszacowania (4.4.13) oraz (4.4.14) w Twierdzeniu 9 ściśle łączą się z tym rozkładem. Mianowicie ostatnie wyrażenia po prawych stronach (4.4.13) i (4.4.14) dotyczą ryzyka aproksymacji, a pozostałe ryzyka estymacji.

Rozważamy klasyfikatory liniowe, zatem ryzyko aproksymacji jest małe, jeśli "prawda" może być dobrze przybliżona w sposób liniowy. Dla estymatorów z logistyczną stratą fakt ten jest opisany przez ostatni składnik w (4.4.14), który mierzy odległość między prawdziwym prawdopodobieństwem sukcesu a tym z regresji logistycznej. W szczególności jeśli prawdziwy model jest logistyczny, to ostatni składnik w (4.4.14) znika. Ostatnie wyrażenie w (4.4.13) odnosi się do ryzyka aproksymacji w przypadku kwadratowej straty. Mierzy ono, jak dobrze warunkowa wartość oczekiwana $\mathbb{E}(Y|X)$ może być opisana przez „najlepszą" (według kwadratowej straty) liniową funkcję $X^T\theta^*_{quad}$.

Pierwsze trzy składniki po prawych stronach (4.4.13) oraz (4.4.14) odnoszą się do ryzyka estymacji. Wartości te są małe, jeśli pokażemy, że prawdopodobieństwo zdarzeń Ω^c_{quad} oraz Ω^c_{log} są małe, a także ciąg c maleje wystarczająco szybko do zera. Taką analizę przeprowadziliśmy w [A3, Twierdzenie 3]. Krótko mówiąc, pokazaliśmy, że dla c proporcjonalnego do $\frac{t\sigma^2\sqrt{\log p}}{F_1(\xi)\sqrt{n}}$ rozważane prawdopodobieństwo zachowuje się jak $1/p$.

W [A3] własności selekcyjne estymatorów w źle wyspecyfikowanej klasyfikacji są również badane. W szczególności w [A3, Sekcja 5] analizujemy estymatory LASSO z kwadratową stratą zastosowane do danych klasyfikacyjnych. Nie opisujemy tych wyników dokładnie w tej sekcji, gdyż są one podobne do rezultatów z [A2], na przykład [A2, Twierdzenie 5].

4.4.4 Regresja porządkowa

W (4.1.9) zdefiniowaliśmy M-estimatory jako penalizowane elementy minimalizujące sumę niezależnych zmiennych losowych. W [A7] rozważamy ogólniejszy przypadek, to znaczy penalizowane U- statystyki, czyli sumy zależnych zmiennych losowych. Estymatory te mogą być użyte, między innymi, do danych z modelu (4.4.1) albo do regresji porzadkowej (rankingu), gdzie zmienna Y jest mierzona na skali porządkowej. Tutaj skupimy się na regresji porządkowej, ale nasze rozważania można łatwo rozszerzyć, zob. [A7]. Zacznijmy od krótkiego opisu regresji porządkowej.

Rozważmy dwa niezależne obiekty o jednakowym rozkładzie $Z_1 = (X_1, Y_1)$ oraz $Z_2 = (X_2, Y_2)$. Zmienne Y_1 oraz Y_2 są nieznanymi zmiennymi, które definiują porządek między obiektami. Mianowicie obiekt z_1 jest "lepszy" (szybszy, silniejszy itp.) niż obiekt z_2, jeśli $y_1 > y_2$. Nasze zadanie polega na przewidzeniu porządku między obiektami na podstawie wektorów cech X_1 oraz X_2. Będziemy konstruować funkcje $f : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, zwane regułami rangującymi, które przewidują porządek następująco: jeśli $f(x_1, x_2) > 0$, to przewidujemy, że $y_1 > y_2$.

Niech $\phi : \mathbb{R} \to [0, \infty)$ będzie ustaloną funkcją straty. Odpowiednikami funkcji ryzyka (4.1.1) oraz ryzyka empirycznego (4.1.2) są $Q(f) = \mathbb{E}\phi\left[\text{sign}(Y_1 - Y_2)f(X_1, X_2)\right]$ oraz U-statystyka (rzędu dwa)

$$\bar{Q}(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \phi\left[\text{sign}(Y_i - Y_j)f(X_i, X_j)\right], \tag{4.4.16}$$

odpowiednio. Sumowanie w (4.4.16) odbywa się względem wszystkich par różnych indeksów $(i, j) \in \{1, \ldots, n\}^2$. Jesli Y jest mierzone w skali porządkowej, to zamieniamy $\text{sign}(Y_i - Y_j)$ na $2\mathbb{I}(Y_i > Y_j) - 1$ w definicjach $Q(\cdot)$ oraz $\bar{Q}(\cdot)$. Regułę Bayesa, która minimalizuje $Q(\cdot)$ pośród wszystkich funkcji mierzalnych, oznaczamy przez f_B (jak w Sekcji 4.4.3).

W [A7] rozważamy rodzinę liniowych kombinacji ustalonego zbioru funkcji bazowych. Tutaj, dla prostoty, skupimy się na liniowych regułach rangujacych $\mathcal{F} = \{f_\theta(x_1, x_2) = (x_1 - x_2)^T\theta : \theta \in \mathbb{R}^p\}$. Zdefiniujmy $\theta^* = \arg\min_\theta Q(\theta)$, gdzie $Q(\theta)$ zastąpiło $Q(f_\theta)$. Ponadto θ^* jest zdefiniowana jak w Sekcji 4.1. M-estymator z karą LASSO dany jest wzorem (4.1.9), ale $\bar{Q}(\cdot)$ jest U-statystyką (4.4.16).

Zaprezentujemy główny wynik z pracy [A7] przy dodatkowych założeniach upraszczających (użyta notacja pochodzi z [A7]): $\mathbf{F}$ jest rodziną wszystkich reguł rangujących, $||f||^2 = \mathbb{E}f^2(X_1, X_2)$, $G(u) = cu^2$ dla pewnego $c > 0$, $\Theta_1 = \{\theta^*\}$, $\delta = 1/8$, $C = 1$, $\hat{C} = 1$. Zabieg ten pozwoli łatwo porównać kolejny wynik z jego odpowiednikiem w Sekcji 4.4.1. Pojawiające się poniżej liczby K_i oznaczają stałe.

Twierdzenie 10 (A7, Twierdzenie 1) *Przypuśćmy, że funkcja straty ϕ spełnia warunek Lipschitza z $L > 0$, $|X_1|_\infty \leqslant \sqrt{n/\log p}$, a także istnieje $c \in (0, 1]$ takie, że dla każdego θ mamy $Q(\theta) - Q(f_B) \geqslant c\mathbb{E}\left[f_\theta(X_1, X_2) - f_B(X_1, X_2)\right]^2$. Ponadto oznaczmy $\lambda = K_1 L\sqrt{\log p/n}$. Wtedy mamy*

$$\mathbb{P}\left(Q(\hat{\theta}) - Q(f_B) + \lambda|\hat{\theta} - \theta^*|_1 \leqslant K_2\left[\frac{t\lambda^2}{c\kappa(3)} + Q(\theta^*) - Q(f_B)\right]\right) \geqslant 1 - 3/p^2, \tag{4.4.17}$$

gdzie $\kappa(\cdot)$ jest zdefiniowana w (4.2.3).

W Twierdzeniu 10 szacujemy ryzyko względne $Q(\hat{\theta}) - Q(f_B)$ przez dwa wyrażenia. Pierwsze z nich odnosi się do ryzyka estymacji, a drugie do ryzyka aproksymacji. Można stąd wywnioskować, że nasz estymator zachowuje się prawie jak wyrocznia wiedząca zawczasu, które parametry (cechy) powinny być użyte do aproksymacji najlepszej funkcji f_B. Istotnie załóżmy, że znamy prawdziwy zbiór T. Wtedy estymowalibyśmy tylko współczynniki zawarte w T, a zamiast pozostałych położylibyśmy zera. Ryzyko względne tego estymatora można oszacować analogicznie jak w (4.4.17), zob. [B10]. Krótko mówiąc, w tym przypadku ryzyko estymacji zachowuje się jak t/n. Zatem wyrażenie $\log p$, które pojawia się w parametrze λ^2, a przez to w pierwszym składniku po prawej stronie (4.4.17), wydaje się być ceną, którą płacimy za to, że zawczasu nie znamy zbioru T.

W Twierdzeniu 10 podajemy również nierówność probabilistyczną dla l_1-odległości $|\hat{\theta} - \theta^*|_1$. Załóżmy, że najlepsza funkcja f_B może być dobrze przybliżona przez f_{θ^*} w sensie ryzyka. Wtedy odległość ta zachowuje się jak $\frac{tL\sqrt{\log p}}{c\sqrt{n}\kappa(3)}$, więc jest mała, jeśli model jest rzadki, n jest wystarczająco duże oraz $\kappa(3)$ jest niezbyt małe. Zauważmy, że p może być znacząco większe niż n.

Zauważmy, że procedury oparte na (4.4.16) nie używają aktualnych wartości zmiennych odpowiedzi Y_i. Potrzebują jedynie porządku między tymi zmiennymi. Sprawia to, że są one odporne, więc mogą być użyte w modelu (4.4.1) z nieznaną funkcją g oraz nieregularnymi rozkładami błędów. Teraz porównamy te metody z RankLasso wprowadzonym w Sekcji 4.4.1. Kluczowa przewaga tej drugiej metody dotyczy złożoności obliczeniowej. Mianowicie RankLasso używa kwadratowej funkcji straty, więc można łatwo wypisać wzory na elementy minimalizujące RankLasso względem poszczególnych współrzędnych. Dzięki temu cykliczne algorytmy „spadku gradientowego" (z ang. cyclical coordinate descent) mogą pracować efektywnie z wysokowymiarowymi zbiorami danych [13]. U-procesy z karą z odpornymi funkcjami straty nie posiadają tej własności. Kwadratowa strata jest ważna również

z teoretycznego punktu widzenia, gdyż dzięki niej analiza oraz dowody są znacznie prostsze. Jeśli pominiemy stałe i nieistotne wyrażenia, wtedy ryzyko empiryczne ze stratą (4.4.2) może być wyrażone jako $\theta^T(\mathbf{X}^T\mathbf{X}/n)\theta - 2\theta^T A$, gdzie A jest odpowiednią U-statystyką. Zatem ryzyko empiryczne zależy tylko od jednej U-statystyki (a nie od całej rodziny U-statystyk jak w tej sekcji), a także zależność ta jest liniowa. Dzięki temu dla RankLasso potrafimy otrzymać lepsze wyniki niż w Twierdzeniu 10. Na przykład założenie o ograniczoności wektora cech możemy w [A2] zastąpić byciem wektorem subgaussowskim. W (4.4.17) odległość między $\hat{\theta}$ oraz θ^* mierzymy w l_1-normie, a w [A2, Twierdzenie 2] mamy wyniki dla każdej l_q-normy, $q \geqslant 1$. W dowodzie zgodnej selekcji progowego RankLasso wystarczy wziąć l_∞-odległość w (4.4.4), która używa $F_\infty(\xi)$. Selekcyjna zgodność TL może być łatwo otrzymana z (4.4.17). Użylibyśmy w niej $\kappa(\xi)$, więc założenia Twierdzenia 6 byłyby słabsze, gdyż $F_\infty(\xi) \geqslant \kappa(\xi)$. Co więcej, wykazanie selekcyjnej zgodności ważonej modyfikacji penalizowanych U-statystyk jest wysoce nietrywialne. Jednakże tego typu wynik dla ważonego RankLasso jest podany w Twierdzeniu 7. Z drugiej strony pewne zalety Twierdzenia 10 są związane z oszacowaniami ryzyka względnego estymatorów. Jednak wyniki te dotyczą predykcji porządku miedzy zmiennymi odpowiedzi, a nie ich aktualnych wartości. Zatem mają one ograniczoną przydatność poza regresją porządkową.

4.5 Analiza danych niskowymiarowych

Przypadek niskowymiarowy dotyczy sytuacji, gdy liczba cech p jest stosunkowo duża (powiedzmy kilkadziesiąt), ale znacznie mniejsza niż rozmiar próbki n. Aktualnie przypadek ten nie jest tak intensywnie badany jak wysokowymiarowy. Jednakże, moim zdaniem, również on zasługuje na uwagę, gdyż wciąż można spotkać interesujące problemy z danymi, w których liczba parametrów jest znacznie mniejsza niż n.

Naturalnie można stwierdzić, że każdy rezultat wysokowymiarowy można zredukować do niskowymiarowego. Mówiąc bardziej precyzyjnie, w przypadku wysokowymiarowym zakładamy, że $p = p(n)$ oraz p może rosnąć znacznie szybciej niż n, na przykład wielomianowo bądź wykładniczo. Zatem każde rozważane wyrażenie (oprócz stałych) może zmieniać się z n, w szczególności T, $\theta^*_{\min}$ czy CIF. W sytuacji niskowymiarowej zakłada się, że p może rosnąć z n, ale jest znacznie mniejsze niż n, albo że p jest ustalone. Będziemy rozważać te drugie podejście. Zatem "redukcja rezultatu wysokowymiarowego do niskowymiarowego" oznacza, że "analizujemy ten wynik przy ustalonym p". Naturalnie w ten sposób sprawiamy, że założenia tego rezultatu stają się prostsze, a także łatwiejsze do sprawdzenia. Ponadto warunki te oraz otrzymane twierdzenia są łatwiejsze do zinterpretowania. Co więcej, parametry procedur (na przykład λ czy τ) często nie zależą od nieznanych wartości. Jednakże tego typu redukcje mogą prowadzić do słabszych rezultatów w porównaniu do wyników analiz dedykowanych zbiorom niskowymiarowym. Można to zaobserwować, przyglądając się pracom [A6, A8] oraz [A2, Dodatek].

Rozważmy sytuację, gdy p jest ustalone. Jedynie estymatory oraz parametry λ, λ_a, τ mogą zmieniać się z n. W [A8] rozważamy estymator LASSO (4.1.9) oraz jego ważoną wersję

$$\hat{\theta}^a = \arg\min_{\theta \in \mathbb{R}^p} \quad \bar{Q}(\theta) + \lambda_a \sum_{j=1}^{p} \frac{|\theta_j|}{|\tilde{\theta}_j|}\,, \tag{4.5.1}$$

gdzie $\tilde{\theta}$ jest wstępnym estymatorem wektora θ^* takim, że $\sqrt{n}(\tilde{\theta} - \theta^*) = O_P(1)$. W [A8] zakładamy, że fukcja straty ϕ jest wypukła oraz
(1) θ^* minimalizujace (4.1.1) istnieje i jest jedyne,
(2) $Q(\cdot)$ jest dwukrotnie różniczkowalne w θ^* oraz macierz $\Sigma = \nabla^2 Q(\theta^*)$ jest dodatnio określona,
(3) $\mathbb{E}|\partial\phi(\theta, Z)|^2 < \infty$ dla każdego θ w pewnym otoczeniu wektora θ^*, gdzie $\partial\phi(\theta, z)$ jest subgradientem funkcji ϕ w θ dla ustalonego z. W przypadku, gdy subgradient nie jest jednoznacznie wyznaczony, wymagamy jedynie, aby $\partial\phi$ było mierzalnym wyborem subgradientu, zob. [25, Dodatek].

Dwa główne wyniki z [A8] są następujące:

Twierdzenie 11 (A8, Wniosek 2.4) *Przypuśćmy, że założenia (1)-(3) są spełnione oraz* $\lambda \searrow 0$, $\sqrt{n}\lambda \to \infty$.
(a) Jeśli estymator (4.1.9) *jest selekcyjnie zgodny, to*

$$\left|\Sigma_{T^c,T}\Sigma_T^{-1}\mathrm{sign}(\theta_T^*)\right|_\infty \leqslant 1.$$

(b) Jeśli

$$\left|\Sigma_{T^c,T}\Sigma_T^{-1}\mathrm{sign}(\theta_T^*)\right|_\infty < 1,$$

to estymator (4.1.9) *jest selekcyjnie zgodny.*

Twierdzenie 12 (A8, Twierdzenie 3.1) *Przypuśćmy, że założenia (1)-(3) są spełnione. Niech dodatkowo* $n\lambda_a \to \infty$ *oraz* $\sqrt{n}\lambda_a \to 0$. *Wtedy*
(a) $\lim_{n\to\infty} \mathbb{P}\left(supp(\hat{\theta}^a) = T\right) = 1$,
(b) $\sqrt{n}\left(\hat{\theta}_T^a - \theta_T^*\right) \to_d N\left(0, \Sigma_T^{-1} D_T \Sigma_T^{-1}\right)$, *gdzie* $D = \mathbf{Var}\partial\phi(\theta^*, Z)$.

W Twierdzeniu 11 pokazujemy, że warunek IRR (por. (4.2.4)) jest koniecznym i „prawie" wystarczającym warunkiem selekcyjnej zgodności estymatora (4.1.9) nawet wtedy, gdy dane są niskowymiarowe. W Twierdzeniu 12 dowodzimy, że selekcyjna zgodność ważonej modyfikacji estymatora (4.1.9) nie wymaga tego warunku. Ponadto niezerowe współczynniki $\hat{\theta}^a$ są estymowane ze standardowym rzędem. Takie procedury często nazywane są wyrocznią, gdyż zachowują się jak idealne (niedostępne w praktyce) estymatory, które zawczasu znają zbiór T. Ponadto zauważmy, że estymator (4.5.1) jest konstruktywny, gdyż wstępny estymator $\tilde{\theta}$ oraz parametr λ_a nie zależą od nieznanych parametrów.

Wyniki podobne do tych z Twierdzenia 11 oraz Twierdzenia 12 były dowodzone w różnych modelach i przy użyciu różnych technik. Dla GLM mozna je znaleźć w [42] oraz [44], a w [39] badano model liniowy z wartością bezwzględną jako funkcją straty. Zatem nasze wyniki rozszerzają poprzednie rezultaty i ujednolicają ich dowody. Łącząc narzędzia z analizy wypukłej [27, 25, 14] i techniki penalizacyjne, zaproponowaliśmy ogólną metodę, która jest użyteczna w badaniu własności estymatorów LASSO z wypukłymi funkcjami strat. Może być ona użyta w modelach regularnych (jak GLM) albo w modelach z nieróżniczkowalnymi funkcjami straty jak w [39], albo w klasyfikacji z perceptronową stratą (jak w maszynach wektorów podpierających). Co więcej, metoda ta po niewielkiej modyfikacji może być z powodzeniem zastosowana w sytuacji, gdy składniki w $\bar{Q}(\cdot)$ są zależne, jak w pracy [A6] bądź [A2, Dodatek].

4.5.1 Niskowymiarowe modele semiparametryczne

W [A6] rozważamy M-estymatory z karą, dla których funkcja $\bar{Q}(\cdot)$ jest U-statystyką rzędu dwa, to znaczy

$$\bar{Q}(\theta) = \frac{1}{n(n-1)} \sum_{i\neq j} \phi(\theta, Z_i, Z_j), \tag{4.5.2}$$

gdzie $\phi(\theta, z_1, z_2)$ jest wypukłą funkcją straty względem θ dla ustalonych z_1, z_2. Przyjmujemy, że $Q(\theta) = \mathbb{E}\phi(\theta, Z_1, Z_2)$ oraz $\theta^* = \arg\min_{\theta\in\mathbb{R}^p} Q(\theta)$. Ważnym przykładem są procedury rozważone w [A6, Sekcja 3], które nie używają aktualnych wartości zmiennych odpowiedzi, lecz jedynie porządku między nimi. Jak już wspomnieliśmy w Sekcji 4.4.4, mogą być one użyte w regresji porządkowej. Co więcej, stosując je, możemy otrzymać estymatory odporne, które bedą skutecznie pracować w modelu (4.4.1) czy nieznacznie ogólniejszym modelu z [15]. *Odporność* jest rozumiana, jak w Sekcji 4.4.1, w odniesieniu do funkcji odpowiedzi g w (4.4.1) oraz rozkładów błędów losowych.

Głównymi założeniami w [A6] są:
(1) θ^* minimalizujący $Q(\theta) = \mathbb{E}\phi(\theta, Z_1, Z_2)$ istnieje i jest jedyny,
(2) Q jest dwukrotnie różniczkowalne w θ^* oraz macierz $\Sigma = \nabla^2 Q(\theta^*)$ jest dodatnio określona,
(3) $\mathbb{E}|\partial\phi(\theta, Z_1, Z_2)|^2 < \infty$ dla każdego θ w pewnym otoczeniu wektora θ^*.

W [A6] rozważamy elementy minimalizujące (4.5.2) z ważoną karą postaci $\sum_{j=1}^p w_j|\theta_j|$, gdzie w_j jest (być może losową) liczbą nieujemną. Estymator oznaczmy przez $\hat{\theta}^w$. Niech $\lambda_0 = \max_{j\in T} w_j$, $\lambda_1 = \min_{j\notin T} w_j$.

Twierdzenie 13 (A6, Twierdzenie 4.1) *Przypuśćmy, że (1)-(3) są spełnione. Jeśli* $\sqrt{n}\lambda_0 \to_P 0$, $\sqrt{n}\lambda_1 \to_P \infty$, *to*
(a) $\lim_{n\to\infty} \mathbb{P}\left(supp(\hat{\theta}^w) = T\right) = 1$,
(b) $\sqrt{n}\left(\hat{\theta}^w_T - \theta^*_T\right) \to_d N\left(0, \Sigma_T^{-1} D_1 \Sigma_T^{-1}\right)$, *gdzie* $D = 4\,\mathbf{Var}\,\mathbb{E}\left[\partial\phi(\theta^*, Z_1, Z_2)|Z_2\right]$.

Założenia i tezy Twierdzenia 13 oraz Twierdzenia 12 będą bardzo podobne, jeśli weźmiemy $w_j = \lambda/|\tilde{\theta}_j|$ dla pewnego wstępnego estymatora $\tilde{\theta}$ takiego, że $\sqrt{n}(\tilde{\theta} - \theta^*) = O_P(1)$. Jednakże Twierdzenie 12 dotyczy sytuacji, gdy ryzyko empiryczne $\bar{Q}(\cdot)$ jest sumą niezależnych zmiennych losowych, a ryzyko empiryczne w Twierdzeniu 13 jest sumą zależnych zmiennych. W dowodzie Twierdzenia 13 używamy metody wprowadzonej w [A8], a także pewnych narzędzi z teorii U-statystyk (na przykład oszacowania wariancji czy CTG dla U-statystyk). Odpowiednik Twierdzenia 11 również został wykazany w [A6, Twierdzenie 4.3].

Podobna problematyka była rozważana w [32]. Główne zalety naszego podejścia z [A6] to: rozważane procedury są wypukłe, więc nie mamy problemów z lokalnymi minimami. Ponadto nasze procedury są efektywnie zaimplementowane w [34]. Co więcej, nasze założenia (1)-(3) są słabsze niż ich odpowiedniki z [32]. Dokładne porównanie znajduje się na stronie 4 w [A6].

W Twierdzeniu 10 rozważamy penalizowane U-statystyki w przypadku wysokowymiarowym. Zauważmy, że prosta redukcja tego faktu do sytuacji niskowymiarowej da nam jedynie estymacyjną zgoność $\hat{\theta}$, co jest znacznie słabszą własnością niż tezy udowodnione w Twierdzeniu 13, czyli zgodna selekcja i asymptotyczna noralność.

W [A2, Dodatek] rozważamy procedurę RankLasso, czyli kolejne podejście do odpornej selekcji cech. Głównymi wynikami są : [A2, Twierdzenie 10] oraz [A2, Twierdzenie 11]. Ponownie ich dowody mocno bazują na metodzie z pracy [A8]. Pierwszy wynik dotyczy własności ważonego RankLasso i jest podobny do Twierdzenia 12 oraz Twierdzenia 13. Natomiast wynik z [A2, Twierdzenie 11] odnosi się do progowego RankLasso i jest raczej łatwym wnioskiem z [A2, Lemat 9], które jest odpowiednikiem [A8, Twierdzenie 2.2].

Jak już wspominaliśmy, [A2, Twierdzenie 10] oraz [A2, Twierdzenie 11] są lepsze niż proste redukcje ich wysokowymiarowych wersji, czyli Twierdzenie 7 oraz Twierdzenie 6, odpowiednio. Na przykład teraz zakładamy, że wektory cech mają skończone czwarte momenty, a w Sekcji 4.4.1 wymagaliśmy, aby były subgaussowskie, w szczególności ogony ich rozkładów powinny zbiegać do zera w tempie wykładniczym. Ponadto [A2, Twierdzenie 10] dotyczy każdego elementu minimalizującego, a w Twierdzeniu 7 opisujemy własności tylko pewnego elementu minimalizującego. Co więcej, progowa i ważona modyfikacja RankLasso są konstruktywne w modelu niskowymiarowym, to znaczy ich parametry nie zależą od nieznanych wartości.

W ostatnim akapicie Sekcji 4.4.4 stwierdziliśmy, że w przypadku wysokowymiarowym RankLasso z [A2] jest lepsze (teoretycznie i praktycznie) niż penalizowane U-statystyki z [A7]. W sytuacji niskowymiarowej jakość obydwu podejść jest podobna. Jeśli p jest względnie niewielkie, to przewaga obliczeniowa estymatora RankLasso przestaje być tak widoczna. Ponadto założenia z [A2, Twierdzenie 10] są nieznacznie silniejsze od swoich odpowiedników w Twierdzeniu 13. Istotnie, w [A6, Propozycja 4.4]

użyliśmy penalizowanych U-statystyk z perceptronową (więc odporną) funkcją straty i wymagaliśmy, aby $\mathbb{E}|X_1|^2 < \infty$. Jednakże w [A2, Twierdzenie 10] zakładamy, że wektory cech mają skończone czwarte momenty.

4.6 Pozostałe osiągnięcia naukowo-badawcze

Pozostałe osiągnięcia naukowo-badawcze, niewchodzące w skład ww. osiągnięcia, stanowią następujące artykuły naukowe:

Prace opublikowane przed uzyskaniem stopnia doktora

[B1] W. Niemiro, W. Rejchel (2009). „Rank correlation estimators and their limiting distributions", Statistical Papers, vol. 50, p. 887-893,

[B2] W. Rejchel (2009). „Ranking - convex risk minimization", Proceedings of World Academy of Science, Engineering and Technology, vol. 56, p. 172-178,

Prace opublikowane po uzyskaniu stopnia doktora

[B3] B. Miasojedow, W. Niemiro, W. Rejchel (2021). „Asymptotics of maximum likelihood estimators based on Markov chain Monte Carlo methods", Annales de l'Institut Henri Poincaré - Probabilités et Statistiques, Vol. 57, p. 815-829,

[B4] W. Rejchel (2018). „Generalization Bounds for Ranking Algorithms", rozdział w "Ensemble Classification Methods with Applications in R" (Eds. E. Alfaro, M. Gámez, N. Garcia), Wiley, p. 135-140.

[B5] B. Miasojedow, W. Niemiro, J. Palczewski, W. Rejchel (2016). „Asymptotics of Monte Carlo maximum likelihood estimators", Probability and Mathematical Statistics, vol. 36, p. 295-310.

[B6] A. Doskocz, W. Rejchel (2016). „Evaluation of accuracy of digital map data via multiple comparisons", Bulletin of the Polish Academy of Sciences: Technical Sciences, vol. 64, p. 799-805.

[B7] B. Miasojedow, W. Niemiro, J. Palczewski, W. Rejchel (2016). „Adaptive Monte Carlo Maximum Likelihood", rozdział w Studies in Computational Intelligence, Vol. 605: Challenges in Computational Statistics and Data Mining (Eds. S. Matwin, J. Mielniczuk), Springer,

[B8] W. Rejchel (2015). „Fast rates for ranking with large families", Neurocomputing, vol. 168, p. 1104-1110,

[B9] W. Rejchel, H. Li, C. Ren, L. Li (2015). „Comments and correction on „U-processes and preference learning", Neural Computation, vol. 27, p. 1549-1553,

[B10] W. Rejchel (2012). „On ranking and generalization bounds", Journal of Machine Learning Research, vol. 13, p. 1373-1392,

[B11] A. Doskocz, W. Rejchel (2012). „Proposition of automatization of analysis of accuracy of the large-scale digital maps databases" (in Polish), Scientific Papers of Rzeszów University of Technology - Civil and Environmental Engineering, vol. 59, p. 85-93.

Omówienie

Moje pozostałe osiągnięcia naukowo-badawcze można podzielić na trzy grupy:
- badanie własności estymatorów regresji porządkowej,
- badanie własności estymatorów opartych na metodach Monte-Carlo,
- badanie własności metod używanych w geodezji i kartografii do produkcji wielkoskalowych map cyfrowych.

Statystyczne własności estymatorów regresji porządkowej

Prace [B1, B2, B4, B8, B9, B10] koncentrują się na estymatorach regresji porządkowej. Regresja porządkowa jest nieraz nazywana rankingiem bądź regresją rangową. Problematyka ta była również rozważana w [A6] oraz [A7] i jest krótko omówiona w Subsection 4.4.4. Zauważmy, że w [B1, B2, B4, B8, B9, B10] rozważamy M-estymatory bez kary. Jedynym wyjątkiem jest Przykład 3 z [B10], w którym analizujemy estymator z karą l_2. Ponadto w [A6, A7] pracujemy z liniowymy regułami rangowymi albo z rodziną liniowych kombinacji ustalonych funkcji bazowych. Tymczasem w tych wcześniejszych pracach rozważane rodziny reguł rangujących mogą być znacznie ogólniejsze. Jedynym wyjątkiem jest [B1], w której badamy jedynie liniowe reguły rangujące. Co najważniejsze, w [A6, A7] jesteśmy zainteresowani wyborem modelu oraz predykcją, podczas gdy we wcześniejszych pracach tylko predykcyjne własności estymatorów były analizowane.

Jak wspomniałem powyżej, w [B1] rozważana jest rodzina liniowych reguł rangujących i M- estymatory bez kary (4.1.3). Dowodzimy zgodność oraz asymptotyczną normalność takich procedur. Bardziej wyrafinowane rezultaty są otrzymane w [B2, B10]. Wyniki te są nieasymptotyczne, a także badane reguły rangujące nie muszą być liniowe. Mówiąc dokładniej, w pracach tych badamy ryzyko względne estymatorów, rozumiane jako różnica między ryzykiem wypukłym estymatora i ryzykiem wypukłym najlepszej reguły w badanej rodzinie, mianowicie $Q(\hat{f}) - Q(f^*)$, gdzie $\hat{f}$ minimalizuje (4.4.16) w rodzinie $\mathcal{F}$ dla pewnej funkcji straty ϕ, a f^* minimalizuje teoretyczny odpowiednik (4.4.16) w rodzinie $\mathcal{F}$. W teorii uczenia się predykcyjne własności estymatorów są często opisywane, używając nierówności probabilistycznych postaci: dla każdego $\alpha \in (0,1)$

$$\mathbb{P}\left(\; Q(\hat{f}) - Q(f^*) \leqslant \eta \; \right) \geqslant 1 - \alpha, \tag{4.6.1}$$

gdzie $\eta > 0$ jest pewną liczbą zależącą od poziomu α, rozmiaru próbki n, rodziny reguł rangujących $\mathcal{F}$ oraz funkcji straty ψ, ale jest ona niezależna od nieznanego rozkładu P. W pracy [6] pokazano, że zależność między η oraz n może być lepsza niż $1/\sqrt{n}$, ale tylko wtedy, gdy ϕ jest $0-1$ funkcją straty. Jednakże użycie tej funkcji straty prowadzi do algorytmów, które są obliczeniowo nieefektywne. Zatem celem prac [B2, B10] było wykazanie analogicznych własności dla wypukłych funkcji straty. Pierwsze rezultaty są zawarte w pracy [B2], która została opublikowana w materiałach pokonferencyjnych. Następnie wyniki te zostały wzmocnione i rozszerzone w [B10]. W Twierdzeniu 8 z [B10] wskazane są warunki, które powinna spełniać funkcja straty oraz rodzina $\mathcal{F}$, pozwalające uzyskać w (4.6.1) rząd zbieżności szybszy niż $1/\sqrt{n}$. W szczególności, jeśli $\mathcal{F}$ jest euklidesowa ([B10, założenie A]), to otrzymujemy rząd zbieżności $\log n/n$. Jeśli rodzina $\mathcal{F}$ spełnia warunek entropii ([B10, założenie B]), to mamy $1/n^{\beta}$ oraz $\beta \in (2/3, 1)$. Otrzymane wyniki mogą być zastosowane do powszechnie używanych algorytmów regresji porządkowej, na przykład maszyn wektorów podpierających. Ponadto badana była skuteczność rozważanych estymatorów na danych rzeczywistych.

Zauważmy, że pojawia się luka pomiędzy „zwykłym" tempem zbieżności z $\beta = 1/2$ otrzymanym w [6] a wynikami z [B10], gdzie $\beta > 2/3$. Fakt ten jest interesujący nie tylko z matematycznego punktu widzenia, ale również z praktycznego, gdyż w ten sposób ograniczona jest użyteczność wyników z [B10]. Na przykład nie obejmują one algorytmu AdaBoost, który jest jednym z najbardziej popularnych

algorytmów regresji porządkowej. Należy zauważyć, że w Przykładzie 2 z [B10] rozważa się wersję tego algorytmu i wymaga się, aby liczba iteracji w AdaBoost była mała, co jest rzadko praktykowane. Problem ten jest badany w [B4, B8], w których skonstruowano nierówności probabilistyczne (4.6.1) z rzędem zbieżności $1/n^{\beta}, \beta \in (1/2, 1)$. Oszacowania te, zastosowane do AdaBoost, nie zależą od liczby iteracji w algorytmie [B8, Przykład 1]. Ponadto w dowodzie Twierdzenia 1 z [B8] użyto Twierdzenia 2 z [B8], czyli nierówności wykładniczej dla U-procesów. Podobne wyniki mogą być łatwo znalezione w literaturze, ale ten rezultat jest „dobrze dopasowany" do badanego problemu i może być łatwo użyty nie tylko w dowodzie Twierdzenia 1 z [B8], ale również w dowodzie Twierdzenia 1 [A7], w którym bada się penalizowane estymatory regresji porządkowej.

W Twierdzeniu 1 z [20] udowodniono nierowność Bernsteina dla U-procesów rzędu dwa, używając metod z teorii entropii. Następnie wyniki te zostały zastosowane do skonstruowania nierówności probabilistycznych (4.6.1) dla estymatorów regresji porządkowej. Jednakże analiza ta opierała się na zasadzie kontrakcji dla chaosu Rademachera rzędu dwa [20, Twierdzenie 8], która nie jest prawdziwa. W [B9] wskazujemy odpowiedni kontrprzykład. Ponadto w Sekcji 3 z [B9] naprawiamy tę lukę w dowodach głównych wyników z [20].

Estymatory oparte na metodach Monte Carlo

W pracach [B3, B5, B7] badamy estymatory największej wiarogodności (NW) otrzymane metodami Monte Carlo (MC). Metody NW są dobrze znanymi i często używanymi narzędziami w estymacji nieznanych parametrów modeli. Jednakże w przypadku wielu złożonych modeli statystycznych dokładne wyznaczenie tych estymatorów jest bardzo trudne, a nieraz niemożliwe. Problemy te pojawiają się, gdy rozważane gęstości są znane tylko z dokładnością do trudnej do wyznaczenia stałej normującej, na przykład w losowych polach Markowa badanych w [A5]. W pracy [B5] przybliżamy tę stałą używają metody *losowania istotnego*. W Twierdzeniu 3.1 z [B5] wskazujemy warunki dające asymptotyczną normalność estymatorów NW opartych na MC w sytuacji, gdy rozmiary próbki wyjściowej oraz MC zbiegają do nieskończonosci.

Statystyczne własności estymatorów NW opartych na MC w znacznym stopniu zależą od wyboru gęstości w losowaniu istotnym. W pracy [B7] zaproponowaliśmy adaptacyjny algorytm, w którym gęstość ta jest dopasowywana w czasie trwania symulacji. Zgodność i asymptotyczna normalność procedury jest wykazana w pracy [B7] w Propozycji 2 oraz Twierdzeniu 3, odpowiednio. Chcąc przezwyciężyć problem degeneracji wag w losowaniu istotnym, zaproponowaliśmy drugi algorytm, który oparty jest na metodzie repróbkowania i metodach MC opartych łańcuchach Markowa. Asymptotyczna normalność tej procedury wykazana jest w Twierdzeniu 4 z [B7]. Ponadto prezentujemy wyniki doświadczeń numerycznych, w których badamy własności zaproponowanych rozwiązań.

W pracach [B5, B7] zakłada się, że potrafimy sprawnie generować niezależne próbki z zadanych rozkładów. Jednakże w wielu ciekawych problemach jest to niemożliwe. Potrafimy jedynie symulować efektywnie to łańcuchy Markowa o zadanym rozkładzie stacjonarnym. W pracy [B3] naszym celem była analiza własności tego typu estymatorów. We Wniosku 3.2 w [B3] udowodniliśmy zgodność tych estymatorów, a w Twierdzeniu 3.3 z [B3] wykazaliśmy ich asymptotyczną normalność. W tym przypadku analizujemy obydwa źródła losowości, analogicznie jak w [B5]. Nasze rezultaty zostały zastosowane do modeli z trudnymi do wyznaczenia stałymi normującymi, a także do modeli z brakami danych. Otrzymane rezultaty są potwierdzone w eksperymentach numerycznych [B3, Sekcja 4].

W pracy [A5] używamy metod MC do wysokowymiarowego modelu Isinga, który jest jednym z najpopularniejszycgh modeli z trudną do wyznaczenia stałą normującą.

Wielkoskalowe mapy cyfrowe w geodezji i kartografii

W pracach [B6, B11] porównujemy metody używane do produkcji wielkoskalowych map cyfrowych.

W ostatnich dziesięcioleciach w Polsce zgromadzono znaczące zasoby danych geodezyjnych i kartograficznych. Jednakże zostały one otrzymane, używając różnych układów odniesienia oraz metod zbierania danych. Aktualnie zasoby te są integrowane w ogólnopolski system odniesienia. Kluczowym aspektem w integracji tych danych jest ich jakość. W pracach [B6, B11] badamy jakość czterech metod używanych do produkcji map cyfrowych: nowy pomiar tachimetrem, przeliczenie wcześniejszych pomiarów bezpośrednich zrealizowanych poprzez pomiary ortogonalne i biegunowe, ręczna wektoryzacja rastrowego obrazu ortofotomapy i graficzno-numeryczne przetwarzanie map analogowych.

Rozważane bazy danych zawierają liczne nieregularności, dlatego do ich analizy używamy metod rangowych. W pracy [B11] badamy błędy badanych metod w oparciu o przedziały ufności i testy statystyczne. Te wstępne wyniki zostały później wzmocnione w pracy [B6], w której bierzemy pod uwagę fakt, że porównujemy cztery metody jednocześnie. Zatem naszą analizę oparliśmy na porównaniach wielokrotnych. Głównym wynikiem badań jest ustalenie porządku między metodami względem ich jakości.

5 Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej

1. stypendium badawcze postdoc zostało mi przyznane w ramach konkursu ogłoszonego przez dr. hab. P. Pokarowskiego, kierownika grantu OPUS z NCN, 2015/17/B/ST6/01878, pt. „SOSnet: oszczędne modelowanie i predykcja dla danych wysokiego wymiaru". Stypendium było realizowane od 10.2017 do 09.2018 na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego. W jego wyniku powstała publikacja [A1],

2. staż podoktorski FUGA 2014/12/S/ST1/00344, „Regresja rangowa i U-procesy z karą LASSO - selekcja cech, estymacja i nierówności z wyrocznią" został mi przyznany przez NCN. Byłem jego kierownikiem, a opiekunem naukowym był prof. dr hab. W. Niemiro. Staż odbyłem na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego od 10.2014 do 09.2016. W jego wyniku powstały trzy publikacje [A5, A6, A7],

3. w sierpniu 2015r. przebywałem na tygodniowej wizycie naukowej na zaproszenie prof. Luoqinga Li na Uniwersytecie Hubei w Wuhan (Chiny). W wyniku tej wizyty powstała publikacja [B9],

4. współpraca z prof. dr hab. Małgorzatą Bogdan (Wydział Matematyki i Informatyki, Uniwersytet Wrocławski): wspólna publikacja [A2],

5. współpraca z dr. hab. Konradem Furmańczykiem, prof. SGGW (Instytut Informatyki Technicznej, Szkoła Główna Gospodarstwa Wiejskiego): dwie wspólne publikacje [A3, A4],

6. współpraca z dr. hab. Błażejem Miasojedowem, prof. UW (Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski), m.in. wspólne publikacje [B3, B5, B7],

7. współpraca z dr. hab. Janem Palczewskim (School of Mathematics, University of Leeds, Wielka Brytania): dwie wspólne publikacje [B5, B7],

8. współpraca z dr. Adamem Doskoczem (Wydział Geoinżynierii, Uniwersytet Warmińsko-Mazurski): dwie wspólne publikacje [B6, B11]

6 Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę

6.1 Działalność dydaktyczna

- promotor pomocniczy rozprawy mgr. Patryka Truszczyńskiego pt. „Estymacja indeksu ogona dystrybuanty metodą kwantyli blokowych", Wydział Matematyki i Informatyki UMK, promotor: prof. dr hab. Adam Jakubowski,
- promotor pomocniczy rozprawy mgr. Patryka Krasuskiego, Szkoła Doktorska Nauk Ścisłych i Przyrodniczych UMK, promotor: dr hab. Aleksander Zaigrajew, prof. UMK
- prowadzący seminarium dyplomowe dla 3 roku matematyki i ekonomii (opiekun 8 prac dyplomowych w 2013, 8 prac w 2014 oraz 5 prac w 2017) na Wydziale Matematyki i Informatyki UMK w Toruniu,
- prowadzenie wykładów na studiach doktoranckich w Szkole Doktorskiej Nauk Ścisłych i Przyrodniczych UMK oraz Academii Copernicana na UMK,
- prowadzenie wykładów oraz ćwiczeń na studiach licencjackich i magisterskich na UMK (Wydział Matematyki i Informatyki, Wydział Chemii, Wydział Nauk o Ziemi i Gospodarki Przestrzennej, Wydział Filozofii i Nauk Społecznych), Wyższej Szkole Bankowej w Toruniu (Wydział Finansów i Zarządzania) oraz UWM (Wydział Matematyki i Informatyki, Wydział Medycyny Weterynaryjnej, Wydział Nauk Technicznych, Wydział Nauki o Żywności), m.in. ze statystyki matematycznej, biostatystyki, uczenia maszynowego, analizy danych, analizy matematycznej, rachunku prawdopodobieństwa

6.2 Działalność organizacyjna

- członek Komisji Statystyki Komitetu Matematytki PAN od 2020 r.
- członek Komitetu Organizacyjnego XLIV Konferencji „Statystyka Matematyczna", 02-07.12.2018, Będlewo
- udział w pracach nad tworzeniem nowego kierunku studiów: „matematyka stosowana" na Wydziale Matematyki i Informatyki UMK, 2016 r.
- członek Jury Konkursu PTM na najlepszą pracę studencką z teorii prawdopodobieństwa i zastosowań matematyki - 2020 r.
- członek komisji konkursu na stypendium badawcze dla doktoranta w grancie BEETHOVEN z NCN, 2018/31/G/ST1/02252, „Analiza wrażliwości dla operatorów nielokalnych z zastosowaniami do procesów skokowych", kierownik prof. K. Bogdan - lipiec 2020r.

7 Pozostałe osiągnięcia naukowo-badawcze i inna działalność

7.1 Granty i projekty badawcze (niewymienione wczesniej)

1. wykonawca w grancie OPUS z NCN, 2018/31/B/ST1/00253, „Metody obliczeniowe dla wysokowymiarowego uczenia statystycznego", 2019-2023 (w realizacji), kierownik: dr hab. B. Miasojedow

2. wykonawca w grancie OPUS z NCN, N N201 608740, „Asymptotic properties and inequalities for MCMC estimators", 2011-2014, kierownik: prof. dr hab. W. Niemiro

3. główny wykonawca w grancie promotorskim MNiSW N N201 391237 „Statystyczne modele regresji rangowej" na realizację pracy doktorskiej 2009-2011, kierownik: prof. dr hab. W. Niemiro

4. kierownik projektu "Stypendia dla doktorantów 2008/2009 - ZPORR".

7.2 Nagrody i wyróżnienia

- indywidualna nagroda rektora UMK drugiego stopnia za osiągnięcia naukowe w 2020r.,
- główna nagroda (ex aequo) w konkursie na najlepszy referat wśród młodych matematyków w czasie XL Konferencji Zastosowań Matematyki, 30.08-06.09.2011, Zakopane,
- obronienie z wyróżnieniem rozprawy doktorskiej pt. „Statystyczne modele regresji rangowej", 09.03.2011, Toruń,
- główna nagroda w konkursie na najlepszy referat wśród młodych matematyków w czasie "The international conference on trends and perspectives in linear statistical inference Linstat'2008" i wygłoszenie plenarnego wykładu na następnej konferencji tego cyklu - LinStat2010,

7.3 Inna działalność

- Poza referatami przedstawionymi w „Wykazie osiągnięć naukowych albo artystycznych, stanowiących znaczny wkład w rozwój określonej dyscypliny" wygłosiłem równiez referaty na seminariach naukowych:
  - Instytutu Matematycznego PAN,
  - Instytutu Podstaw Informatyki PAN,
  - Wydziału Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski,
  - Wydziału Matematyki, Politechnika Wrocławskia,
  - Instytutu Matematycznego, Uniwersytet Wrocławski,
  - Instytutu Matematyki, Uniwersytet Marii Curie-Skłodowskiej w Lublinie,
  - Wydziału Matematyki i Statystyki, Uniwesytet Hubei, Wuhan (Chiny),
  - Wydziału Matematyki i Statystyki, Uniwersytet Rolniczy Huazhong, Wuhan (Chiny),
  - Katedry Geodezji Szczegółowej, Wydział Geodezji i Gospodarki Przestrzennej, Uniwersytet Warmińsko-Mazurskiego w Olsztynie

Bibliografia

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *In Petrov, B. N.; Csaki, F. (eds.), 2nd International Symposium on Information Theory*, pages 267–281. Budapest: Akademiai Kiado.

[2] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43:2055–2085.

[3] Belloni, A. and Chernozhukov, V. (2011). l_1 penalized quantile regression in high dimensional sparse models. *Annals of Statistics*, 39:82–130.

[4] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732.

[5] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York.

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

[7] de la Peña, V. H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. Springer-Verlag, New York.

[8] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

[9] Fan, J., Fan, Y., and Barut, E. (2014a). Adaptive robust variable selection. *Annals of Statistics*, 42:324–351.

[10] Fan, J., Xue, L., and Zou, H. (2014b). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42:819–849.

[11] Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975.

[12] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

[13] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

[14] Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript*.

[15] Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics*, 35:303–316.

[16] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

[17] Huang, J. and Zhang, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, 13:1839–1864.

[18] Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31:253–258.

[19] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin.

[20] Li, H., Ren, C., and Li, L. (2014). U-Processes and Preference Learning. *Neural Computation*, 26:2896–2924.

[21] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

[22] Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28:863–884.

[23] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

[24] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462.

[25] Niemiro, W. (1992). Asymptotics for M-estimators defined by convex minimization. *Annals of Statistics*, 20:1514–1533.

[26] Pokarowski, P. and 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.

[27] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199.

[28] R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[29] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6:461 – 464.

[30] Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107:223–232.

[31] Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65:807–832.

[32] Song, X. and Ma, S. (2010). Penalized variable selection with U-estimates. *Journal of Nonparametric Statistics*, 4:499–515.

[33] Su, W. J., Bogdan, M., and Candès, E. J. (2017). False discoveries occur early on the lasso path. *The Annals of Statistics*, 45:2133–2150.

[34] Teo, C. H., Vishwanathan, S. V. N., Smola, A., and Le, Q. V. (2010). Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11:311–365.

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

[36] van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36:614–645.

[37] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

[38] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Verlag, New York.

[39] Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal Journal of Business & Economic Statistics*, 25:347–355.

[40] Ye, F. and Zhang, C. H. (2010). Rate minimaxity of the lasso and Dantzig selector for the l_q loss in l_r balls. *Journal of Machine Learning Research*, 11:3519–3540.

[41] Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942.

[42] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

[43] Zhou, S. (2009). Thresholding procedures for high dimensional variable selection and statistical estimation. *Advances in Neural Information Processing Systems*, 22:1436–1462.

[44] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.