Zielona Góra, 26 października 2023 r.

prof. dr hab. inż. Dariusz Uciński
Instytut Sterowania i Systemów Informatycznych
Uniwersytet Zielonogórski

RECENZJA
rozprawy doktorskiej Pani mgr Antoniny Krajewskiej
pt. *Efficient matrix completion for data recovery in data-driven IT applications*
opracowana na wniosek Rady Naukowej Instytutu Badań Systemowych
Polskiej Akademii Nauk

I. Obszar problemowy rozprawy

W wielu problemach praktycznych związanych z analizą danych, uczeniem maszynowym lub wizją komputerową pojawia się problem rekonstrukcji macierzy w oparciu o rezultat próbkowania jej elementów. Przykładem jest uzupełnianie brakujących odpowiedzi respondentów w badaniach ankietowych. Po zadaniu określonego zestawu pytań pewnej grupie osób, odpowiedzi możemy umieścić w macierzy, której wiersze odpowiadają poszczególnym uczestnikom badania, a kolumny – pytaniom. Jeśli na wiele pytań w kwestionariuszu uczestnicy nie odpowiedzieli, istotną kwestią staje się oszacowanie, jakie najprawdopodobniej byłyby te brakujące odpowiedzi. Oczywiście, tak sformułowanego zadania rekonstrukcji nie daje się rozwiązać bez dodatkowych informacji o specyfice problemu.

W wielu przypadkach taką informacją jest to, że rekonstruowana macierz jest niskiego rzędu. Charakteryzuje się tym np. znany problem Netflixa, w którym użytkownicy tej platformy (odpowiadają im wiersze macierzy) oceniają filmy dostępne na tej platformie (każdemu filmowi odpowiada jedna kolumna macierzy). Użytkownicy oceniają z reguły bardzo niewiele filmów, co przekłada się na niewielki zbiór elementów macierzy danych mających zdefiniowane wartości liczbowe. Serwis streamingowy chciałby jednak uzupełnić macierz danych tak, aby na podstawie uzupełnionych wpisów mógł rekomendować użytkownikom nowe filmy. Rozsądnym jest wówczas przyjęcie założenia, że macierz ocen filmów dokonanych przez użytkowników jest niskiego rzędu, ponieważ powszechnie uważa się, że tylko kilka czynników wpływa na gusta i preferencje danej osoby.

Niekiedy spełnienie wymogu określenia wartości liczbowych wszystkie elementów macierzy danych jest wręcz trudne do spełnienia. Przykładowo, w bioinformatyce koszt zebrania pełnej informacji o ekspresji genów dla setek genów i tysięcy pacjentów jest zazwyczaj nieakceptowalnie wysoki.

Problem uzupełniania brakujących elementów dotyczy szczególnie bardzo dużych macierzy (o tysiącach lub milionach wierszy i/lub kolumn). Jeśli rząd takiej macierzy jest istotnie mały w porównaniu z jej wymiarami, fundamentalnym staje się więc określenie, czy można ją odtworzyć na podstawie próbkowania jej elementów. Staje się to ekstremalnie trudne w przypadku macierzy rzadkich, kiedy próbkowanie może prowadzić do błędnego wniosku, że mamy do czynienia z macierzą zerową.

Samą rekonstrukcję można byłoby sformułować jako zadanie uzupełnienia brakujących elementów macierzy w taki sposób, aby uzupełniona macierz miała najmniejszy możliwy rząd, jednak takie sformułowanie ma jedynie walor teoretyczny z uwagi na NP-trudność problemu optymalizacyjnego. Stąd w praktyce dominują różnego rodzaju sformułowania z osłabionymi wymogami, np. zamiast rzędu macierzy po uzupełnieniach minimalizuje się jej normę nuklearną (czyli sumę jej wartości osobliwych) lub wprowadza sformułowania prowadzące do wykorzystania efektywnych algorytmów optymalizacji wypukłej. Odpowiednie metody uzupełniania macierzy osiągnęły już pewną dojrzałość, znajdując poczesne miejsce w bardzo dobrych podręcznikach, np. G. James, D. Witten, T. Hastie, R. Tibshirani,

An Introduction to Statistical Learning with Applications in R, 2nd End, Springer, 2021, rozdz. 12.3, lub G. Strang, *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press, 2019, rozdz. III.5. Mimo tego, nie można stwierdzić, że istniejące rozwiązania są do końca satysfakcjonujące, co nadal motywuje badania nad nowymi, bardziej efektywnymi obliczeniowo wersjami algorytmów.

Właśnie w tym kontekście recenzowana rozprawa doktorska Pani mgr Antoniny Krajewskiej, poświęcona w całości efektywnemu uzupełnianiu macierzy na potrzeby rekonstrukcji danych w zastosowaniach informatycznych opartych o dane, jest pozycją bardzo ambitną i aktualną. Zasadniczo, oryginalny pomysł Autorki ukierunkowany jest na macierze z brakującymi elementami, których liczba kolumn jest o wiele większa od liczby wierszy, i polega poszukiwaniu macierzy z uzupełnionymi elementami w postaci iloczynu dwóch macierzy. Pierwszą z nich otrzymuje się poprzez próbkowanie kolumn oryginalnej macierzy i uzupełnienie elementów macierzy otrzymanej z połączenia tych kolumn w oparciu o wybraną tradycyjną metodę uzupełniania elementów macierzy. Drugą macierz otrzymuje się w taki sposób, aby jej iloczyn z pierwszą najlepiej przybliżał istniejące elementy oryginalnej macierze w sensie kryterium sumy kwadratów reszt. Względnie mała liczba wylosowanych kolumn prowadzi do istotnej redukcji nakładów obliczeniowych zarówno podczas uzupełniania elementów macierzy, jak i wyznaczania drugiej macierzy. Z kolei do wyznaczenia drugiej macierzy można wykorzystać którąś z intensywnie rozwijanych metod ukierunkowanych na wielkoskalowe liniowe zadania najmniejszych kwadratów, m.in. w oparciu o obliczenia równoległe lub rozproszone. Autorka proponuje trzy warianty implementacji takiego schematu obliczeniowego i dowodzi jego poprawności. Co więcej, obszerną część rozprawy zajmuje praktyczna weryfikacja działania zaproponowanych metod w oparciu o nietrywialne zadania testowe.

Biorąc pod uwagę wszystkie wymienione czynniki, sformułowane na str. 16 dwie hipotezy badawcze, jak również wynikające z nich zadania szczegółowe, są jasne i dobrze określone. Sprowadzają się one do wykazania, że proponowana metoda uzupełniania brakujących elementów macierzy w oparciu o losowy wybór kolumn wyznacza rozwiązania, których jakość jest porównywalna ze znanymi metodami uzupełniania braków danych, a nowe algorytmy implementujące tę metodę są efektywne, skalowalne, oraz konkurencyjne z innymi algorytmami wykorzystującymi minimalizację normy nuklearnej. Tak zarysowaną problematykę rozprawy uważam za istotną i nadzwyczaj aktualną, o rezultatach mogących otworzyć nowy nurt badań nad metodami rekonstrukcji elementów macierzy. Fakt ten przesądza o pozytywnej ocenie wybranego tematu jako przedmiotu opiniowanej rozprawy doktorskiej.

II. Koncepcja oraz realizacja rozprawy

Rozprawa, napisana w języku angielskim i licząca 136 stron numerowanych, składa się ze wstępu, rozdziału zawierającego preliminaria matematyczne i wprowadzającego notację, siedmiu zasadniczych rozdziałów przedstawiających stan wiedzy w zakresie tematyki rozprawy oraz koncepcję proponowanych metod wraz z wynikami eksperymentów weryfikujących działanie tych metod w praktyce, oraz rozdziału podsumowującego z uwagami końcowymi. Załączony bardzo obszerny wykaz 197 pozycji cytowanej literatury świetnie odzwierciedla stan badań w zakresie tematycznym rozprawy.

Pracę rozpoczyna zwięzłe *Wprowadzenie*, na które składa się przedstawienie motywacji zagadnień rozprawy, dwie hipotezy badawcze (w tym momencie już intuicyjnie jasne), oraz krótkie przedstawienie struktury rozprawy.

Rozdział 2 zawiera definicje wielu pojęć używanych w rozprawie, m.in. rozkładu według wartości osobliwych, norm macierzowych oraz macierzy pseudoodwrotnej.

W *rozdz. 3* opisano motywacje problemu uzupełniania brakujących elementów macierzy, systemy rekomendujące oparte o wspólne filtrowanie, uzupełnianie braków w macierzach niskiego rzędu, związek próbkowania oszczędnego z uzupełnianiem braków w macierzach w oparciu o algorytmy optymalizacji wypukłej. Na tym tle przedstawiono skrótowo podejście proponowane w rozprawie wraz z analizą potencjalnych zastosowań.

Rozdział 4 zawiera przegląd najważniejszych technik uzupełniania braków w macierzach, przede wszystkim sformułowań w kategoriach programowania wypukłego. Omawia się algorytmy dokładne, związane z minimalizacją normy nuklearnej zrekonstruowanej macierzy w oparciu o programowanie półokreślone, oraz algorytmy przybliżone, minimalizujące kryterium będące ważoną sumą błędu przybliżenia macierzy przed uzupełnieniami macierzą po uzupełnieniach oraz normy nuklearnej macierzy uzupełnionej. Narzędziem do minimalizacji w tym drugim przypadku może być metoda gradientu proksymalnego. Omawia się również podejście oparte o faktoryzację Burera i Monteiry oraz schemat naprzemiennych minimalizacji, jak również szereg innych metod zaproponowanych w literaturze.

W *rozdz. 5* omówiono dwa algorytmy minimalizujące normę nuklearną: dokładną minimalizację w oparciu o programowanie półokreślone oraz metodę gradientu proksymalnego w celu przybliżonego uzupełnienia elementów macierzy. Omówiono również sposób ilościowego charakteryzowania podatności macierzy na uzupełnianie w oparciu o minimalizację normy nuklearnej lub faktoryzację macierzową niskiego rzędu, prowadzący do zdefiniowania tzw. współczynnika spójności macierzy. Wiąże się to z obserwacją, że w celu zauważalnej minimalizacji liczby obserwacji potrzebnych do odtworzenia macierzy niskiego rzędu, jej wektory osobliwe powinny być wystarczająco mocno rozproszone, tzn. nieskorelowane z wektorami bazy kanonicznej. Współczynnik spójności, zdefiniowany w pracy Candèsa i Rechta z 2009 r., w takiej sytuacji przyjmuje swoją minimalną wartość równą jedności.

Rozdział kończy rezultat Rechta dotyczący minimalizacji normy nuklearnej, będący typowym przykładem warunków określających minimalną liczebność próby i skojarzonego z nim dolnego oszacowania prawdopodobieństwa dokładnej rekonstrukcji wszystkich braków danych w macierzy. W rozprawie rozważa się przede wszystkim algorytmy randomizowane, więc takie charakteryzacje są całkiem typowe.

W *rozdz. 6* omawia się problem wyboru podzbioru kolumn macierzy, które łączy się w jedną macierz niskiego rzędu, przybliżającą macierz oryginalną (w uczeniu maszynowym jest to równoważne nienadzorowanemu algorytmowi wyboru cech). Rozważa się również algorytm CUR+, opisany w artykule Xu *i in.* z 2015 r., w którym próbkuje się zarówno wiersze, jak i kolumny, aby na ich podstawie utworzyć dwie macierze wykorzystane do budowy macierzy niskiego rzędu przybliżającej macierz oryginalną. W celu budowy takiego przybliżenia, próbkuje się oryginalną macierz, po czym do elementów próby dopasowuje się iloczyn wspomnianych dwóch macierzy oraz macierzy parametrów o małych wymiarach, minimalizując kryterium najmniejszej sumy kwadratów reszt. Pomysł zastosowany w tamtej pracy jest najsilniejszą inspiracją do podejścia zaproponowanego w rozprawie.

Rozdział 7 jest najważniejszy w rozprawie z uwagi na zaprezentowanie w nim szczegółów oryginalnego podejścia zaproponowanego przez Autorkę, wraz z podaniem dowodu jego poprawności. Ogólny schemat procedury podano w rozdz. 7.1. Macierzy z uzupełnionymi brakami poszukuje się w postaci iloczynu dwóch macierzy. Pierwszą z nich otrzymuje się poprzez próbkowanie kolumn oryginalnej macierzy (liczebność próbki określa się arbitralnie) i uzupełnienie elementów macierzy otrzymanej z połączenia tych kolumn w oparciu o wybraną tradycyjną metodę uzupełniania elementów macierzy. Drugą macierz otrzymuje się w taki sposób, aby jej iloczyn z pierwszą najlepiej przybliżał istniejące elementy oryginalnej macierze w sensie sumy kwadratów reszt. Względnie mała liczba wylosowanych kolumn prowadzi do istotnej redukcji nakładów obliczeniowych zarówno podczas uzupełniania elementów macierzy, jak i wyznaczania drugiej macierzy. Z kolei do wyznaczenia drugiej macierzy można wykorzystać, zależnie od wymiarowości problemu, albo jedno z podejść tradycyjnych, albo którąś z intensywnie rozwijanych metod ukierunkowanych na wielkoskalowe liniowe zadania najmniejszych kwadratów, m.in. w oparciu o obliczenia równoległe lub rozproszone. Trzy takie uszczegółowione wersje Autorka prezentuje zresztą dalej. Jest to minimalizacja normy nuklearnej z wyborem kolumn (Column Selected Nuclear Norm, CSNN) dla małych i średnich macierzy, oraz metoda gradientu proksymalnego z wyborem kolumn (CSPGD) wraz z jej wersją wykorzystującą procedurę Adam (CSPGD-Adam) dla problemów wielkoskalowych. Najważniejszym wynikiem pracy jest Twierdzenie 7.2.2 (str. 61), podające dolne oszacowanie prawdopodobieństwa idealnej

rekonstrukcji macierzy, jak również warunki na minimalną liczbę losowanych kolumn oraz minimalną liczebność próby znanych elementów rekonstruowanej macierzy gwarantujące osiągnięcie tego oszacowania. Dowód jest wprawdzie rozbudowany, jednak dość czytelny. Wynik jest poprawny matematycznie, chociaż nie wydaje się być jeszcze nigdzie opublikowany, a jest tego jak najbardziej wart.

Po przeczytaniu pierwszych siedmiu rozdziałów czytelnik posiada dobrą ogólną orientację w zakresie omawianych zagadnień, dotychczas stosowanych podejściach, oraz oryginalnych rozwiązaniach zaproponowanych przez Autorkę.

W kolejnych dwóch rozdziałach Autorka przedstawia wyniki eksperymentów potwierdzających efektywność proponowanych przez Nią metod. Ta część stanowi wprawdzie prawie połowę objętości pracy, jednak jest równie interesująca, jak część pierwsza. Duże uznanie budzi przede wszystkim ogromny nakład pracy włożony w przeprowadzenie wszystkich badań, tak bardzo dalekich od trywialności. Ich rezultaty potwierdziły w praktyce zasadność zaproponowanej metodologii, i to nie tylko w oparciu o dane generowane symulacyjnie (*rozdz. 8*), ale przede wszystkim w oparciu o dane rzeczywiste (*rozdz. 9*), dotyczące systemów rekomendujących, rekonstrukcji niekompletnych obrazów, oraz predykcji połączeń w sieciach.

Rozprawę kończy podsumowanie oryginalnych wyników naukowych oraz charakterystyka otwartych problemów badawczych.

Oceniając merytorycznie całą rozprawę stwierdzam, że jest ona napisana na bardzo dobrym poziomie. Zawiera jasno sformułowany i ważny problem naukowy, oraz prezentuje poprawne rozwiązanie tego problemu, które zostało uzyskane przez Autorkę samodzielnie i z zastosowaniem właściwej metodologii naukowej. Na podstawie przedstawionego skrótowo omówienia treści całej rozprawy doktorskiej należy odnotować, że jej Autorka wykazała się dobrymi umiejętnościami formułowania problemów naukowo-badawczych oraz ich efektywnego rozwiązywania z zastosowaniem zaawansowanych narzędzi algebry liniowej, technik optymalizacji, algorytmów randomizowanych i analizy probabilistycznej, uczenia maszynowego, jak również technik algorytmicznych. Już na podstawie wstępnej analizy można stwierdzić, że rozprawa stanowi dzieło wartościowe, zdecydowanie odpowiadające wymaganiom stawianym przez stosowne przepisy, w szczególności w przypadku dyscypliny informatyka techniczna i telekomunikacja.

Pod względem redakcyjnym praca napisana jest z dbałością o szczegóły. Użyte słownictwo odpowiada powszechnie stosowanemu. Jak na tak dużą objętość, zawiera niewiele błędów składu, co tym bardziej koresponduje z jej bardzo dobrym poziomem merytorycznym. Na szczególne podkreślenie zasługuje to, że pracę napisano w bardzo dobrym języku angielskim.

III. Oryginalne osiągnięcia

Chociaż problem uzupełniania brakujących elementów dużych macierzy niskiego rzędu przyciąga uwagę badaczy od dwóch dekad i dostępny jest już cały wachlarz rozmaitych technik, to jednak poszukiwanie sposobów uczynienia go jeszcze bardziej efektywnym jest nadzwyczaj trudnym i wciąż aktualnym zadaniem. Przedstawiony w pracy matematyczny opis problemu, jego analizę oraz zaproponowane metody i algorytmy obliczeniowe uważam za najważniejszy wkład Autorki w rozważaną dziedzinę. Chociaż zaproponowane algorytmy należą do klasy algorytmów zrandomizowanych, czyli wynik ich pracy może, z pewnym prawdopodobieństwem, być niepoprawny, jednak jeśli prawdopodobieństwo błędu jest wystarczająco małe, zysk związany z czasem czy pamięcią może się bardzo opłacać. Główną zaletą podejścia proponowanego w rozprawie jest wysoka efektywność obliczeniowa oraz względnie prosta możliwość jego rozwinięcia do działającego prototypu.

Przyjmując, że głównym celem rozprawy było pokazanie, że proponowana metoda uzupełniania brakujących elementów macierzy w oparciu o losowy wybór kolumn wyznacza rozwiązania, których jakość jest porównywalna ze znanymi metodami uzupełniania braków danych, a nowe algorytmy implementujące tę metodę są efektywne i skalowalne, oraz konkurencyjne z innymi algorytmami wykorzystującymi minimalizację normy nuklearnej, należy stwierdzić, że cel ten Autorka osiągnęła. Co więcej, weryfikacji rezultatów dokonano w oparciu o nietrywialne benchmarki oraz eksperymenty w warunkach rzeczywistego użytkowania.

W szczególności, za najważniejsze rezultaty rozprawy uważam następujące:

1. Zaproponowanie nowego dwuetapowego schematu uzupełniania braków w macierzach, których jeden wymiar jest dużo większy od drugiego. Jego zaletą jest prostota i efektywność obliczeń. Pierwszy etap redukuje się bowiem do losowania pewnej liczby kolumn z rekonstruowanej macierzy, połączenia ich w jedną macierz i rekonstrukcji jej elementów w oparciu o znane techniki uzupełniania wykorzystujące normę nuklearną. Drugi etap sprowadza się do rozwiązania ciągu liniowych zadań najmniejszych kwadratów. Zdecydowaną zaletą jest przy tym brak konieczności dokonywania rozkładu według wartości osobliwych, co rzutuje na dużą efektywność obliczeniową schematu.

2. Opracowanie szczegółowych procedur obliczeniowych implementujących ogólny schemat wspomniany wyżej w zależności od wielkości rekonstruowanej macierzy (CSNN, CSPGD, CSPDG-Adam). Ich zaletą jest ponownie prostota i duża efektywność obliczeniowa.

3. Nietrywialny dowód poprawności zaproponowanego algorytmu zrandomizowanego, wraz z podaniem dolnego oszacowania prawdopodobieństwa idealnej rekonstrukcji macierzy, jak również warunków na minimalną liczbę losowanych kolumn oraz minimalną liczbę znanych elementów rekonstruowanej macierzy gwarantujących osiągnięcie tego oszacowania.

4. Wykonanie szeregu eksperymentów na danych syntetycznych i rzeczywistych, potwierdzających poprawność i efektywność zaproponowanych procedur.

5. Implementacja rozważanych algorytmów w postaci otwartego kodu w języku Python, co umożliwia społeczności naukowej rzetelną ocenę i ewentualny rozwój zaproponowanych algorytmów.

W podsumowaniu, należy stwierdzić, że sformułowany cel rozprawy został osiągnięty, a jej Autorka wykazała się głęboką wiedzą i umiejętnościami niezbędnymi do samodzielnego rozwiązywania problemów naukowo-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

IV. Uwagi i komentarze

Przedstawiona do recenzji praca zawiera istotną treść naukową i wiele nowych wyników. Stanowi logiczną całość począwszy od uzasadnienia praktycznego problemu, poprzez jego formalizację, aż do rozwiązania różnorakich wersji problemu z przykładami zastosowań zaproponowanej metodologii w nietrywialnych zadaniach testowych. Praca prezentuje wysoki poziom naukowy, a Autorka biegle posługuje się aparatem matematycznym, m.in. w zakresie algebry liniowej, technik optymalizacji, algorytmów randomizowanych i analizy probabilistycznej, uczenia maszynowego, jak również technik algorytmicznych.

Lektura rozprawy skłania jednak również do sformułowania następujących komentarzy lub uwag krytycznych:

1. W zaproponowanej metodzie liczbę próbkowanych kolumn ustala się arbitralnie. Jak dobierać ją w praktyce? Twierdzenie 7.2.2 podaje warunek na minimalną liczbę kolumn, jednak

korzystanie z niego wymaga znajomości współczynnika spójności i rzędu rekonstruowanej macierzy (parametr γ określa się na podstawie zakładanego poziomu prawdopodobieństwa dokładnego odtworzenia macierzy), które przecież nie są dostępne z uwagi na braki danych w macierzy. Czy istnieje alternatywa dla najbardziej pesymistycznych oszacowań obu wielkości?

2. Losowanie kolumn odbywa się zgodnie z rozkładem równomiernym, co jednak nie uwzględnia ewentualnych dużych różnic między nimi. W algorytmach CUR czasami stosuje się również technikę *leverage score sampling* lub próbkowanie adaptacyjne. Na ile możliwym byłoby wykorzystanie ich w proponowanej metodzie?

3. Algorytm CUR+ próbkuje nie tylko kolumny, ale i wiersze macierzy. Wydaje się, że wprowadzenie próbkowania wierszy nie powinno to być przesadną komplikacją w proponowanej metodzie. Czy ograniczenie się do próbkowania kolumn wynikało tylko z uproszczenia rozważań, czy jest raczej związane z ograniczeniami proponowanej metody?

4. Technika CSNN jest dedykowana małym i średnim problemom, a CSPGD i CSPGD-Adam – problemom wielkoskalowym. Co to jednak oznacza pojęcie problemów wielkoskalowych w kontekście problemu rozważanego w rozprawie?

5. Nie jest jasne, dlaczego do rozwiązywania liniowego zadania najmniejszych kwadratów, czyli *de facto* deterministycznego zadania optymalizacji wypukłej o bardzo dobrze rozpoznanej specyfice, proponuje się wykorzystać procedurę Adam, czyli bardzo ogólną metodę pierwszego rzędu zaproponowaną do rozwiązywania problemów optymalizacji stochastycznej. Funkcja celu w zadaniu najmniejszych kwadratów rozwiązywanym w drugim kroku proponowanej metody CSMC (str. 55) jest na dodatek separowalna i kolumny macierzy $\mathbf{Z}$ można wyznaczać niezależnie od siebie, poprzez rozwiązywanie niezależnych od siebie liniowych zadań najmniejszych kwadratów o stosunkowo małej wymiarowości (o n_1 zakłada się przecież, że jest dużo mniejsze od n_2), które można rozwiązywać z zastosowaniem znanych metod dedykowanych problemowi najmniejszych kwadratów. Stwarza to wręcz idealne warunki do zrównoleglenia obliczeń, czego tu nie wykorzystano.

6. Wydaje się, że dolne oszacowanie prawdopodobieństwa idealnej rekonstrukcji macierzy podane w Twierdzeniu 7.2.2 jest wyższe niż to podane w pracy Xu *i in.* (2015). Oczywiście, trzeba byłoby jeszcze przy tym uwzględnić nieco różne warunki na minimalne liczby losowanych kolumn oraz znanych elementów rekonstruowanej macierzy, jednak takie porównanie niewątpliwie dałoby pełniejszy obraz wartości proponowanego podejścia.

7. W dowodzie Twierdzenia 7.2.2 (str. 68) rozważa się zdarzenie, że zachodzi (7.50), oraz zdarzenie, że funkcja g jest β-silnie wypukła. Oznaczmy je odpowiednio jako A i B. Z tego, że $P(A) \geq 1 - 2e^{-\gamma}$ oraz $P(B) \geq 1 - e^{-\gamma}$ Autorka wnioskuje, że $P(A \cap B) \geq (1 - 2e^{-\gamma})(1 - e^{-\gamma}) \geq 1 - 3e^{-\gamma}$, jednak trudno powiedzieć skąd miałoby wynikać, że $P(A \cap B) \geq (1 - 2e^{-\gamma})(1 - e^{-\gamma})$. Prawidłowe rozumowanie powinno być następujące:
$$P(A \cap B) = 1 - P(\overline{A \cap B}) \geq 1 - P(\bar{A}) - P(\bar{B}) = P(A) + P(B) - 1 \geq 1 - 3e^{-\gamma}.$$

8. Losowanie elementów macierzy zgodnie z rozkładem równomiernym można interpretować jako losowanie próby prostej z pewnej populacji (tu składają się na nią wszystkie elementy macierzy). Na podobnej zasadzie, zaprezentowane próbkowanie elementów macierzy, polegające na losowaniu kolumn, można zinterpretować jako losowanie zespołowe: na pierwszym etapie losuje się zespoły (tu kolumny), a potem losuje się elementy w zespołach (losuje się elementy kolumn). Taki sposób spojrzenia na problem być może cokolwiek wniesie do polepszenia oszacowań prawdopodobieństw dokładnej rekonstrukcji macierzy. Losowanie zespołowe ma swoje wady i zalety. Im większe zróżnicowanie jednostek w zespołach, a przy tym zróżnicowanie między zespołami mniejsze, tym dobór zespołowy jest bardziej efektywny. Przy dużym zróżnicowaniu między zespołami losowanie obarczone jest jednak sporym

ryzykiem trafienia na grupy, jeśli nie ekstremalne, to w każdym razie nietypowe dla badanej zbiorowości.

Ponadto, w rozprawie pojawia się szereg nieścisłości lub błędów składu:

1. Wzór (5.2), str. 39: nie wyjaśniono, że X to oznaczenie przestrzeni Hilberta; w oznaczeniu normy indeksem dolnym powinno być X, a nie 2.

2. Wzór (5.13), str. 40: symbol dopełnienia zbioru Ω jest w niewłaściwym miejscu.

3. W Twierdzeniu 5.3.1 (str. 45) przytoczonym z pracy Rechta, we fragmencie *observed with locations sampled uniformly with replacement at random*, należy usunąć *with replacement.*

4. W Algorytmie 2 na str. 57 zmienna decyzyjna z ma być wektorem d-wymiarowym, a nie macierzą.

5. W Etapie I procedury CSNN (str. 57) macierz $\mathbf{C}$ powinna mieć wymiar $n_1 \times d$, a nie $d \times n_2$.

6. Twierdzenie 7.2.3 (str. 62) ma być Twierdzeniem 9 w pracy Xu i in. (2015), jednak w tej pracy brak twierdzenia o numerze 9.

7. We wzorze (7.11) na str. 62 zamiast $\tilde{U}Y$ powinno być $R_{\Omega}(\tilde{U}Y)$.

8. We wzorze (7.12) na str. 63 zamiast $f(\mathbf{X})$ powinni być $g(\mathbf{X})$. Należy również dodać komentarz, że parametr β ma być dodatni.

9. W Uwadze 7.2.3 na str. 63 powinno się podać jak rozumieć hesjan $\nabla^2 g(\mathbf{X})$, bo nie jest to oczywiste z uwagi na to, że $\mathbf{X}$ jest macierzą. Struktura hesjanu pojawia się dopiero na str. 66. Dobrze w tym miejscu byłoby podać odwołanie np. do klasycznej monografii Magnusa i Neudeckera (*Matrix Differential Calculus with Applications in Statistics and Econometrics, 3rd Edn,* Wiley, 2007), najsolidniejszej książki o macierzowym rachunku różniczkowym.

10. W spisie literatury wiele prac zawiera jedynie nazwiska autorów, rok wydania i tytuł pracy, bez dalszych szczegółów, np. [4], [13], [15], [19], [26], [30], [38], [72], [82], [86], [90], [131], [132], [139], [152], [159], [167], [171], [174], [178], [182], [187], [194].

Powyższe uwagi krytyczne nie mają one jednak przesadnego wpływu na ogólną opinię o recenzowanej dysertacji, którą oceniam jako bardzo wartościową.

V. Podsumowanie

Uwzględniając wyżej wymienione uwagi i komentarze oraz całość rozprawy doktorskiej wraz z oryginalnymi osiągnięciami naukowo-badawczymi stwierdzam, że

1. **recenzowana rozprawa doktorska Pani mgr Antoniny Krajewskiej spełnia wszystkie wymagania Ustawy z dnia 20 lipca 2018 r. *Prawo o szkolnictwie wyższym i nauce* w odniesieniu do rozpraw doktorskich;**

2. **w związku z tym wnoszę o dopuszczenie Autorki rozprawy do dalszych, przewidzianych przepisami, etapów przewodu doktorskiego.**

Dominik Mielniczuk