dr hab. Wojciech Rejchel, prof. UMK
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika

Toruń, 03.11.2023

Recenzja rozprawy doktorskiej mgr. A. Mieldzioca

„Regularyzacja i estymacja macierzy kowariancji o strukturze liniowej"

Estymacja macierzy kowariancji to jeden z podstawowych i kluczowych problemów statystycznych, nad którym pochylają się naukowcy od wielu lat. Może się wydawać, że mierząc się z tym problemem przez tak długi czas, badaczom udało się wypracować optymalne techniki estymacyjne. Nic bardziej mylnego. Chociażby ostatnie dekady wymagają od statystyków (i praktyków) umiejętności analizowania (bardzo) dużych zbiorów danych, w których liczba nieznanych parametrów może (znacznie) przekraczać liczbę obserwowanych obiektów. W tej sytuacji naturalne jest uproszczenie problemu poprzez narzucenie macierzy kowariancji pewnej struktury i dalsza praca przy tym założeniu. Obydwa te zagadnienia są aktualnie popularne i intensywnie studiowane. Ponadto są to ważne zagadnienia badawcze w ocenianej rozprawie doktorskiej.

Głównymi problemami rozważanymi w rozprawie są:
(i) zaproponowanie metody identyfikacji struktury macierzy kowariancji, gdy struktura ta jest liniowa,
(ii) wprowadzenie i analiza własności dwóch estymatorów macierzy kowariancji: pierwszy z nich jest modyfikacją standardowego rzutu na przestrzeń struktury, a drugi bazuje na maksymalizacji funkcji wiarogodności z ograniczeniami. Ważną część rozprawy stanowi analiza numeryczna, w której zbadano własności zaproponowanych metod na danych symulowanych.

Omówienie rozprawy doktorskiej

W rozdziale pierwszym wprowadzono podstawowe pojęcia, które zostały użyte w pracy. Najpierw zdefiniowano rozważany model i jego charakterystyki, a następnie strukturę liniową macierzy kowariancji wraz z przykładami. Ponadto omówiono rozkład Wisharta i jego własności (Definicja 1.5 oraz Lemat 1.3), wspomniano o trzech popularnych funkcjach straty (Frobeniusa, Steina oraz kwadratowej), a także przytoczono kilka własności operatorów macierzowych i ich pochodnych.

W rozdziale drugim opisano metodę identyfikacji liniowej struktury macierzy kowariancji, którą można krótko przytoczyć następująco: z rozważanych k struktur liniowych wybieramy tę, która jest najbliżej estymatora największej wiarogodności (NW) macierzy Σ w sensie funkcji straty Steina (ta ostatnia w rozdziale tym jest nazywana entropijną). Mówiąc dokładniej, proponuje się, aby szukać najmniejszej z wartości $\xi_1, \xi_2, \ldots, \xi_k$ dla

$$\xi_i = \min_{\Gamma \in \Psi_i^+} f(S_{ML}, \Gamma),$$

gdzie Ψ_i^+ jest zbiorem dodatnio określonych macierzy z przestrzeni liniowej Ψ_i, a $f(\cdot, \cdot)$ jest przeskalowaną funkcją straty Steina.

W podrozdziale 2.1 podano algorytm rozwiązujący powyższy problem, który jest rozszerzeniem procedury z pracy [26]. Metoda ta bazuje na podejściu Newtona i mocno eksploatuje liniowość struktury macierzy kowariancji. W podrodziale 2.2 zawarto wyniki obszernych badań numerycznych, w których analizowano skuteczność zaproponowanej procedury.

Rozdział trzeci jest, moim zdaniem, najciekawszym elementem rozprawy. Przedstawiono w nim estymator macierzy kowariancji oparty na regularyzowanej metodzie najmniejszych kwadratów. Wiadomo, że rzut ortogonalny próbkowej macierzy kowariancji na przestrzeń struktury nie musi zachowywać dodatniej określoności. Wyjątkiem jest sytuacja, gdy badana przestrzeń struktury jest kwadratowa i komutatywna, co wykazano w Twierdzeniu 3.1. Proponowane w tym rozdziale podejście jest rozszerzeniem metody z pracy [24]. Estymator zdefiniowany jest jako kombinacja wypukła macierzy celu i rzutu próbkowej macierzy kowariancji na liniową przestrzeń struktury, mianowicie ma on postać $\lambda T + (1 - \lambda) S_{LS}$. Problemem pozostaje wybór macierzy celu T oraz współczynnika regularyzacji λ. Jest on częściowo rozwiązany (zob. „Uwagi") w ważnym Twierdzeniu 3.2, które podaje analityczne formuły na „aproksymacyjnie" optymalne wartości T oraz λ. W Twierdzeniu 3.3, które kończy tę część rozprawy, wykazano, że badany estymator jest zgodny i asymptotycznie nieobciążony.

Opisany w rozdziale czwartym estymator wykorzystuje metodę NW. Jednakże przestrzeń potencjalnych rozwiązań ograniczona jest do tych dodatnio określonych macierzy o strukturze liniowej, których współczynnik uwarunkowania jest odpowiednio kontrolowany. W Twierdzeniu 4.1 wykazano, że przestrzeń ta jest stożkiem wypukłym. Następnie podano procedurę, która pozwala wyznaczyć ten estymator. Metoda ta jest oparta na podobnej idei jak algorytm z podrozdziału 2.1. Następnie przeprowadzono pokaźne badania numeryczne, w których porównano własności estymatora z tego rozdziału z regularyzowaną metodzie najmniejszych kwadratów (RMNK) z rozdziału poprzedniego. Ściśle mówiąc, zbadano trzy wersje estymatora NW w zależności od punktu startowego (rzut standardowego estymatora NW macierzy kowariancji na przestrzeń generowaną przez macierz jednostkową albo rzut na przestrzeń macierzy kompletnie symetrycznych, ewentualnie RMNK). Rezultaty eksperymentów nie pozwoliły wskazać zdecydowanego zwycięzcy. Metoda NW z ograniczeniami ma lepsze

własności statystyczne niż RMNK, ale jest czasochłonna numerycznie.

Uwagi

W tej części recenzji prezentuję moje główne uwagi według rozdziałów z rozprawy, których te komentarze dotyczą.

Rozdział 2:

1) W rozdziale tym poszukuje się struktury macierzy kowariancji. Rozwiązanie tak postawionego problemu jest niejednoznaczne w tym sensie, że macierz kowariancji zależy od jednostek, w jakich mierzone są cechy, a więc jej struktura również. Wobec tego dwukrotnie analizując ten sam zbiór danych, ale w drugim przypadku z jedną cechą wyrażoną w zmienionych jednostkach, dojdziemy do różnych konkluzji. Powyższy problem nie występuje, gdy poszukiwana jest struktura macierzy korelacji. Nie oznacza to jednak, że nie pojawią się inne niedogodności.

2) Zaproponowana metoda identyfikacji struktury jest niezadawalająca w tym sensie, że jeśli $\Psi_1^+ \subset \Psi_2^+$, to z góry wiadomo, że $\xi_1 \geq \xi_2$. Zatem postępując zgodnie z wprowadzonym przez Autora podejściem, należy wybrać bardziej skomplikowaną strukturę Ψ_2^+ nawet wtedy, gdy prawdziwą strukturą jest Ψ_1^+. Widać to na Rysunkach 2.1 oraz 2.2, gdy $x = 0$. Chyba najprostszym rozwiązaniem tego problemu byłoby podzielenie zbioru danych na dwie części. Na pierwszej z nich wyznaczamy estymatory dla różnych struktur, a na drugim sprawdzamy, który z nich jest najbliżej estymatora macierzy kowariancji (S bądź S_{ML}) w sensie ustalonej funkcji straty. Ewentualnie możnaby rozważyć modyfikacje dobrze znanych kryteriów informacyjnych (Akaike bądź Schwarza). Czy tego typu (albo inne) podejścia były już badane? Jeśli tak, to z jakim skutkiem.

3) Algorytm 2.1 jest jednym z głównych osiągnięć w rozprawie. Brakuje mi jego gruntownego omówienia, w szczególności części dotyczącej przeszukiwania wstecznego czy wyboru i roli parametrów a i b. Ponadto komenda „jeśli" powinna być dwukrotnie zastąpiona przez „dopóki".

4) Na Rysunkach 2.1 - 2.4 kolory są błędnie przyporządkowane do estymatorów, na przykład kolor niebieski powinien odpowiadać „CS" a nie „T_1".

Rozdział 3:

1) Na początku rozdziału zdefiniowano ambitne cele, czyli wyznaczenie estymatora ustrukturyzowanego, dodatnio określonego i dobrze uwarunkowanego. Jak wspomniałem, rezultaty tego rozdziału oceniam wysoko, jednak postawione cele nie zostały w pełni osiągnięte, co potwierdza następujący komentarz Autora na stronie 37: „w sytuacji, gdy ocena $\hat{\Sigma}_S$ jest nieokreślona lub źle uwarunkowana, możemy zawsze zwiększyć wartość parametru λ tak, aby uzyskać pożądany poziom uwarunkowania. Innym rozwiązaniem jest zastosowanie metody *Multi Target Shrinkage* opisywanej w [4, 16, 19]." Po przeczytaniu tych dwóch

zwięzłych zdań nasuwa się kilka niezwięzłych pytań:
- jak bardzo należy zwiększyć λ? Należy pamiętać, że w ten sposób tracimy „optymalność" z Twierdzenia 3.2,
- czego dotyczy metoda *Multi Target Shrinkage*? Przydałoby się kilka zdań omówienia,
- tutaj, a także w innych miejscach rozprawy, pojawia się kluczowe pojęcie estymatora „dobrze uwarunkowanego". Co ono oznacza? Szkoda, że Autor nie spróbował skonfrontować się z tym zagadnieniem.

2) Jak wspomniałem powyżej, problem wyboru optymalnych T oraz λ jest tylko częściowo rozwiązany w Twierdzeniu 3.2. Mianowicie na początku podrozdziału 3.2 liniową przestrzeń struktury rozkłada się na sumę k-wymiarowej przestrzeni $\mathcal{V}$ i jej dopełnienia ortogonalnego. Następnie optymalnej macierzy celu T szuka się pośród elementów przestrzeni $\mathcal{V}$. Jednakże Autor nie wspomina o tym, jak wybrać k, czyli wymiar przestrzeni $\mathcal{V}$. Zadanie to jest kluczowe, gdyż wyznaczone w Twierdzeniu 3.2 optymalne wartości T i λ zależą od wybranego k.

3) dowód Twierdzenia 3.3:
- w wielu miejscach rozważa się zbieżność ciągów zmiennych losowych, a nie deterministycznych. Zbieżności te wynikają z MPWL, zatem są to zbieżności *prawie wszędzie*,
- w dowodzie części b) nie trzeba odwoływać się do pracy [17], aby pokazać $\mathbb{E}(S_{LS}) = \Sigma$. Wystarczy użyć elementarnych rachunków ze strony 34. Jeszcze sprawniej byłoby, gdyby do dowodu części b) użyto mocnej zgodności z punktu a) oraz Twierdzenia Lebesgue'a o zbieżności majoryzowanej.

Rozdział 4:

1) Na początku podrozdziału 4.1 wspomina się, że zlogarytmowana funkcja wiarogodności nie jest wklęsłą funkcją od Σ. Co prawda jest ona wklęsła od Σ^{-1}, ale ta własność nie jest pomocna w estymacji macierzy Σ o pewnej strukturze. Jednakże na stronie 43 Autor błędnie konkluduje, że „problem maksymalizacji funkcji wiarogodności jest dobrze zdefiniowany". Problem ten byłby dobrze zdefiniowany, gdy minimalizowana ujemna log-wiarogodność była wypukła na zbiorze wypukłym, ale pierwsza z tych własności nie jest spełniona. Zatem Algorytm 4.1, będący kluczowym wynikiem w tym rozdziale, jest użyty do poszukiwania minimum funkcji niewypukłej. Nie jest to postępowanie błędne, jednakże należy pamiętać o niebezpieczeństwach, które mogą się pojawić. Główne z nich to problemy związane z potencjalnym występowaniem wielu minimów lokalnych.

2) Scenariusz badań symulacyjnych w podrozdziale 4.2 jest bardzo podobny do rozważanego w rozdziale 2. Jednakże w tej części zbadano tylko przypadek $m = 5$, a pominięto $m = 20$. Dlaczego? Można się domyślać, że powodem była duża złożoność obliczeniowa estymatora NW, co wydaje się potwierdzać uwaga na końcu strony 48. Dodatkowo chciałbym się dowiedzieć, o ile to możliwe, jakie były własności statystyczne badanych estymatorów dla $m = 20$.

Konkluzja

Uważam, że rozprawa zawiera nowe i interesujące wyniki dotyczące estymacji macierzy kowariancji. Wymienione powyżej usterki i niedociągnięcia obniżają moją ocenę, ale nie zmieniają ogólnego pozytywnego wrażenia.

Przedstawione wyniki teoretyczne i numeryczne wymagały opanowania dość zaawansowanych narzędzi algebraicznych i statystycznych, a także sumienności, dokładności i rachunkowej wprawy. Ponadto zaproponowanie nowych metod niewątpliwie bazowało na dobrej znajomości badanej tematyki, w szczególności związanej z nią literatury.

Uważam, że przedstawiona rozprawa doktorska spełnia ustawowe i zwyczajowe wymagania stawiane rozprawom doktorskim w dyscyplinie matematyka. Wnoszę o dopuszczenie mgr. Adama Mieldzioca do dalszych etapów przewodu doktorskiego.

Wojciech Rejchel