



UNIwersytet
Warszawski

Wydział Matematyki, Informatyki i Mechaniki

Instytut Matematyki Stosowanej i Mechaniki

dr hab. Błażej Miasojedow, prof. UW

Warszawa, 23 stycznia 2023

Recenzja rozprawy doktorskiej mgr Małgorzaty Łazęckiej

Properties of Information – Theoretic Measures of Conditional Dependence

Rozprawa dotyczy problemu testowania warunkowej niezależności zmiennych losowych w oparciu o miary pochodzące z teorii informacji. Analizowane są własności asymptotyczne nie tylko samych miar stosowanych do wykrywania warunkowej niezależności, ale też ich estymatorów.

Rozprawa liczy 129 stron i składa się z trzech rozdziałów i dwóch dodatków rozwijających materiał z prac wymienionych w załączeniu. Pierwszy rozdział wprowadza pojęcia informacji wzajemnej, warunkowej informacji wzajemnej oraz informacji interakcyjnej. Następnie wyprowadzone zostaje rozwinięcie Möbiusa, za pomocą którego można zdefiniować uogólnione kryterium $J_{\alpha,\beta}$ do wyboru zmiennych istotnych. Dla odpowiednich parametrów α, β uogólnione kryterium redukuje się do kryteriów spotykanych w literaturze takich jak JMI, SCIFE, SECMI. Rozdział kończy się twierdzeniami o własnościach asymptotycznych rozważanych kryteriów.

Drugi rozdział poświęcony jest analizie asymptotycznych własności metod replikowania stosowanych do oceny kryterium $J_{\alpha,\beta}$. Badane są dwie metody bootstrapowe oraz jedna oparta na permutacjach. Rozważane są rozkłady estymatorów kryterium $J_{\alpha,\beta}$ uzyskane trzema wspomnianymi metodami, przy założeniu warunkowej niezależności. Poza rozkładami asymptotycznymi estymatorów kryterium $J_{\alpha,\beta}$ autorka przedstawia porządek ich asymptotycznych wariancji.

W ostatnim rozdziale wyniki teoretyczne z dwóch pierwszych rozdziałów zostały zilustrowane za pomocą symulacji numerycznych.

Rozprawa dotyczy istotnego i szeroko badanego problemu szukania struktury zależności pomiędzy zmiennymi losowymi w oparciu o obserwacje. Testowanie warunkowej niezależności jest zadaniem bardzo trudnym, gdyż wymaga olbrzymiej liczby ob-

serwacji, tak aby zobaczyć wszystkie istotne poziomicę warunkowych rozkładów. Aby rozwiązać ten problem szuka się metod lub kryteriów łatwiejszych do obliczenia, a zarazem pozwalających wychwycić najważniejsze zależności. Oceniana praca rozważa kryteria oparte na teorii informacji, które są postacią względnie prostą do estymacji. Część prezentowanych wyników pochodzi z prac autorki z promotorem [4,5] lub jest istotnym rozwinięciem wyników uzyskanych w pracach [1,2,3,4,5]. Rezultaty zawarte w pracy są interesujące. Na szczególną uwagę zasuguje odpowiedź na pytanie kiedy odpowiednio skalowany rozkład graniczny jest rozkładem normalnym, a kiedy rozkładem χ^2 . Moja ocena rozprawy jest pozytywna, mimo że kierunek uogólnienia wcześniejszych wyników jest słabo umotywowany i nie widać jego potencjalnego nietrywialnego zastosowania. Ponadto mam kilka uwag, przedstawionych poniżej, które w większości dotyczą prezentacji.

- Praca dotyczy problemu estymacji struktury zależności w modelach, jednak w pracy ten problem został tylko zasygnalizowany. Brakuje dokładniejszego opisu problemu, a w szczególności przedstawienia przykładu algorytmów, w których przedstawione kryteria pełniłyby istotną rolę. Prezentacja wyników bez szerszego kontekstu utrudnia docenienie wyników rozprawy.
- Sekcja 1.1.2 dotyczy problemu wyboru cech. Problem ten nie został jednak w pracy zdefiniowany. Nie wiadomo czym jest Y , a czym są $Z_1, Z_2, \dots$. Czytelnik musi domyślić się jaki problem jest analizowany.
- Przydałaby się szersza dyskusja jak wybrać parametry α i β na podstawie wyników teoretycznych. Czy na podstawie wyników teoretycznych można zmniejszyć wariancję estymatorów jak przypuszczamy konkretny rodzaj zależności?
- W całej pracy zakłada się, że $p(x) > 0$ dla każdego x co wyklucza niektóre spotykane w praktyce zagadnienia. Autorka nie przedstawia, kiedy to założenie jest istotne. Bez tego założenia wyniki teoretyczne stałyby się trudne do sformułowania. Przydałaby się dyskusja czy i jak można przeprowadzić rozumowanie, gdy to założenie nie jest spełnione.
- W wynikach z Rozdziału 2 (Lematy 2.1.1, 2.1.2, 2.1.3, 2.1.5, 2.2.1, 2.2.2 i 2.2.4) zakłada się zbieżność warunkową pod warunkiem całej nieskończonej próby. Rozkład graniczny jednak nie zależy od konkretnej realizacji. Wyniki te, z matematycznego punktu widzenia, wydają się być niekompletne. Skoro granica jest niezależna od warunkowania to powinno się to założenie dać pominąć.
- W Rozdziale 3 dotyczącym symulacji brakuje przykładu zastosowania wyprowadzonych metod dla danych rzeczywistych. Pozwoliłoby to zilustrować odporność metody na niewielkie zaburzenia.

Podsumowując recenzowaną rozprawę, pomimo uwag wymienionych powyżej, bez wątplenia oceniam pozytywnie. Uważam, że przedstawiona rozprawa może stanowić podstawę do nadania stopnia doktora pani mgr Małgorzacie Łazęckiej.



z poważaniem

dr hab. Błażej Miasojedow, prof UW

Literatura:

1. Kubkowski, M., Łazęcka, M., & Mielniczuk, J. (2020, June). Distributions of a general reduced-order dependence measure and conditional independence testing. In International Conference on Computational Science (pp. 692-706). Springer, Cham.
2. Kubkowski, M., & Mielniczuk, J. (2021). Asymptotic distributions of empirical Interaction Information. *Methodology and Computing in Applied Probability*, 23(1), 291-315.
3. Kubkowski, M., Mielniczuk, J., & Teisseyre, P. (2021). How to Gain on Power: Novel Conditional Independence Tests Based on Short Expansion of Conditional Mutual Information. *J. Mach. Learn. Res.*, 22, 62-1.
4. Łazęcka, M., & Mielniczuk, J. (2020). Analysis of information-based nonparametric variable selection criteria. *Entropy*, 22(9), 974.
5. Łazęcka, M., & Mielniczuk, J. (2021). Multiple testing of conditional independence using information theoretic approach. *Proceedings of the Modelling Decisions for Artificial Intelligence* 81-92.

