Recenzja rozprawy doktorskiej mgr. Małgorzaty Łazęckiej "Properties of Information-Theoretic Measures of Conditional Dependence"

Omówienie rozprawy doktorskiej

Rozprawa dotyczy miar teorioinformacyjnych warunkowej zależności o rozkładach dyskretnych. W pierwszej części pracy zebrane zostały podstawowe informacje dotyczące wspomnianych miar, pokazane zostały ich związki z rozwinięciem Möbiusa dla warunkowej informacji wzajemnej oraz wyprowadzone zostały ich rozkłady asymptotyczne. Następnie zaproponowanych jest kilka metod resamplingowych i rozważany jest problem ich zgodności. Ostatnia część pracy poświęcona jest badaniom symulacyjnym.

W pierwszym rozdziale przypomniane zostały podstawowe miary (entropia,warunkowa entropia, dywergencja Kullbacka-Leiblera, informacja wzajemna). Następnie dyskutowana jest warunkowa wzajemna informacja i jej związek z dywergencją Kullbacka-Leiblera oraz zastosowanie w problemie wyboru zmiennych. Kolejne podrozdziały poświęcone są informacjom interakcyjnym oraz kryteriom opartym na nich. Najważniejszym pojęciem wprowadzonym w tej części pracy jest uogólnione kryterium selekcji zmiennych $I_{\beta,\gamma}$. Ważnym wynikiem jest jego reprezentacja w terminach entropii. Bardzo ogólna postać tego kryterium sprawia, iż w ten sposób można wyrazić wiele popularnych miar np. warunkową wzajemną informację CMI, łączną wzajemną informację JMI, itp. Rozkład asymptotyczny kryterium pokazany jest w Twierdzeniu 1.4.6. W przypadku niezdegenerowanym jest to rozkład normalny, a w zdegenerowaym jest rozkładem sumy ważonej kwadratów niezależnych zmiennych losowych o standardowym rozkładzie normalnych z wagami, które są wartościami własnymi pewnej macierzy zależnej m.in. od macierzy Hessego $I_{\beta,\gamma}$. W szczególnym przypadku, gdy tylko $\beta(1)$ i $\gamma(1)$ są niezerowe, otrzymujemy kryterium drugiego rzędu, $J_{\beta,\gamma}$, dla którego udało się (Lemat 1.4.10) przeprowadzić analizę przypadku zdegenerowanego i pokazać, że pewne wartości parametru β skutkują warunkową niezależnością zmiennych X i Y pod warunkiem Z. W konsekwencji udało się pokazać (Twierdzenie 1.4.11), że w zależności czy warunek warunkowej niezależności zmiennych jest spełniony czy nie, kryterium JMI z wyestymowanymi prawdopodobieństwami ma inny rozkład graniczny. Wynik ten pozwala na skontruowanie testu hipotezy globalnej dot. warunkowej niezależności zmiennych.

Rozdział 2 poświęcony jest metodom resamplingowym i ich zgodności w rozważanym problemie. Metoda bootstrapu CI polega na próbkowaniu z rozkładu łącznego, w którym prawopodobieństwa $p(x, y, z)$ zostały obliczone przy pomocy estymatorów największej wiarogodności w modelu zakładającym warunkowa niezależność X i Y przy danym Z. Metoda CR (warunkowej randomizacji) zakłada niezależne losowanie obserwacji X_i^* z warunkowego rozkładu $X|Z$, który jest w tym przypadku znany. Przy czym losowane są tylko X_i^*, a (Y_i, Z_i) pozostają niezmienione w próbie bootstrapowej. Kolejna metoda, bootstrap X, jest modyfikacją algorytmu CR polegająca na tym, że rozkład warunkowy $X|Z$ nie jest znany i w związku z tym prawdopodobieństwa $p(x|z_i)$ są estymowane. Ostatnia metoda jest metoda permutacyjną. Dla każdej wartości $Z = z$ permutowane są obserwacje X_i, dla których $Z = z$. Dla każdej z czterech metod pokazano asymptotyczą normalność przeskalowanego bootstrapowego estymatora $\widehat{p}^*(x, y, z)$. Dowody dla metod CI, CR oraz bootstrapu X są dosyć standardowe i polegają na wykorzystaniu nierówności Berry'ego-Esseena. Na uwagę zasługują rezultaty uzyskane dla metody permutacyjnej, pokazujące techniczną biegłość doktorantki. W rozdziale 2 analizowane są również macierze kowariancji rozkładów granicznych. Okazuje się, że macierz uzyskana w metodzie permutacyjnej jest zdominowana przez macierze uzyskane w pozostałych metodach. Macierz uzyskana w metodzie CI dominuje wszystkie pozostałe macierze. Przy czym tylko macierz metody CI jest równa macierzy rozkładu asyptotycznego oryginalnego estymatora. Bardzo interesujące, a zarazem ważne są wyniki otrzymane dla resamplingu w przypadku miary CMI. Okazuje się, że niezależnie od rozważaniej metody resamplingowej otrzymujemy ten sam rozkład graniczny, tożsamy z rozkładem asymptotycznym $2n\widehat{CMI}$, czyli każda z metod jest w tym przypadku zgodna. Natomiast w przypadku miary JMI zgodność zachodzi tylko dla metody CI.

Rozdział 3 w całości poświęcony jest wynikom badań symulacyjnych. Dobrane przykłady trafnie ilustrują otrzymane w rozdziale 2 wyniki teoretyczne oraz wskazują na zalety, jak i wady różnych kryteriów. Umiejętnie pokazana jest zarówno zgodność metod bootstrap, jak i jej brak oraz konsekwencje tego faktu. Sekcja 3.3 celnie podsumowywuje wszystkie otrzymane wyniki.

Uwagi dodatkowe:

1. Wydaje się, ze rozważania przeprowadzone w Lemacie 2.2.6., można by uzyskać prościej stosując np. tw. Poly'a.

2. Brakuje komentarza wyjaśniającego oznaczenie $\widehat{\Sigma}_{-K}$ wprowadzone na str. 52. Wyjaśnione jest dopiero na str. 53, gdy pojawia się oznaczenie Σ_{-K}.

3. W rezultatach pokazujących rozkłady graniczne metod resamplingowych mylące jest oznaczanie macierzy kowariancji zawsze jako Σ, gdy nie jest to ta sama macierz. Szczególnie, że w dalszej części pracy (sekcja 2.1.5) macierze te są już oznaczane różnie.

4. Na str. 81 zamiast "method of resampling" użyto "method of subsampling". Ponadto na stronach 85-86 użyto " B subsamples" zamiast "B resamples". Pojęcia te używane są w przypadku metody subsamplingu, która nie jest tematem tej pracy.

5. Na rysunkach 3.3-3.5 zaznaczone kwantyle (0.9, 0.95, 0.975) są podpisane jako 0.1, 0.05 oraz 0.025.

Wyniki przedstawione w pracy doktorskiej zostały opublikowane 3 artykułach.

Dwa z nich to recenzowane arykuły konferencyjne (jeden z J. Mielniczukiem, a drugi z M. Kubkowskim i J. Mielniczukiem), przy czym należy podkreślić, że konferencja ICCS znajduje się w gronie prestiżowych (ma rangę Core A). Trzeci artykuł to publikacja (z J. Mielniczukiem) w bardzo dobrym czasopiśmie *Entropy*. Kolejna praca (wspólna z M. Kubkowskim i J. Mielniczukiem) jest przygotowana do złożenia. Tematyka rozprawy jest bardzo ważna i aktualna. Wszystkie zaprezentowane w pracy dowody wyników są według mnie poprawne. Praca prezentuje wysoki poziom merytoryczny i jest dobrze umotywowana praktycznymi zastosowaniami. Rozdział 3 znakomicie dopełnia całości pracy doktorskiej i pokazuje, że doktorantka nie tylko posiada biegłość techniczną w udowadnianiu twierdzeń, ale rozumie otrzymane wyniki i umie je potwierdzić w dobrze zaprojektowanych rozległych badaniach symulacyjnych.

Wnioski

Praca doktorska mgr. Małgorzaty Łazęckiej spełnia wszystkie wymogi stawiane rozprawom doktorskim w dyscyplinie matematyka. Wnoszę o dopuszczenie p. Małgorzaty Łazęckiej do dalszych etapów przewodu doktorskiego.

Kraków, 20 listopada 2022

Anna Dudek

dr hab. Anna Dudek, prof. AGH