

Warszawa, 8 czerwca 2022

Recenzja rozprawy doktorskiej

pt. „Structure Learning and Parameter
Estimation for Graphical Models via Penalized
Maximum Likelihood Methods”

Maryii Shpak

Przedmiot rozprawy

Rozprawa jest poświęcona rozpoznawaniu struktury modeli graficznych na podstawie danych. Problem polega na identyfikacji krawędzi grafu, który opisuje strukturę zależności pomiędzy zmiennymi lub procesami losowymi.

Wyniki rozprawy dotyczą dwóch klas modeli graficznych: klasycznych sieci bayesowskich (BN) oraz sieci bayesowskich z czasem ciągłym (CTBN). Są to w obu przypadkach modele oparte na grafach skierowanych. Tematyka rozprawy jest ważna i aktualna. Modele graficzne znajdują coraz więcej zastosowań i ich teoria jest wciąż rozwijana. Z drugiej strony, zadania wyboru modelu zajmują centralne miejsce we współczesnej statystyce. Rozpoznawanie struktury grafu jest szczególnie ważne ze względu na rosnącą złożoność i rozmiar modeli graficznych, które opisują skomplikowane zjawiska i wykorzystują ogromne bazy danych. W takiej sytuacji konieczna jest konstrukcja „oszczędnych” modeli, zawierających tylko istotne zależności między zmiennymi.

Literatura dotycząca rozpoznawania struktury sieci bayesowskich (statycznych BN) jest ogromna. Autorce udało się otrzymać nowe wyniki w tym dość dobrze zbadanym obszarze. Jeśli chodzi o analogiczne wyniki dla CTBN, to ten przypadek był dotąd słabo zbadany. Brak było algorytmów rozpoznawania struktury CTBN o własnościach solidnie uzasadnionych teoretycznie. Poświęcona CTBN-om część rozprawy jest nowatorska.

Zawartość rozprawy i ocena wyników

Rozprawa składa się z siedmiu rozdziałów. Rozdział 1 jest wprowadzeniem, Rozdział 2 zawiera wstępne informacje o rozpatrywanych modelach. Rozdział 3 omawia wnioskowanie i estymację parametrów modelu graficznego przy znanej strukturze. Ten rozdział ma tylko pośredni związek z zasadniczym tematem rozprawy. Najważniejsze wyniki pojawiają się w Rozdziale 4 (rozpoznawanie struktury BN), w Rozdziale 5 (rozpoznawanie struktury CTBN przy pełnej obserwowalności) i w Rozdziale 6 (rozpoznawanie struktury CTBN przy niepełnej obserwowalności). Rozdział 7 jest krótkim podsumowaniem całości.

Najogólniej mówiąc, algorytmy zaproponowane w rozprawie są oparte na idei LASSO, czyli minimalizacji gładkiej funkcji wiarygodności z karą typu l_1 . Ta metoda jest dobrze znana i zbadana w kontekście regresji. Jej zastosowanie do

modeli graficznych wymagało od Autorki (i od współautorów prac wspólnie z Autorką napisanych) zarówno pomysłowości w konstrukcji algorytmów, jak i pokonania trudności technicznych w dowodach.

Zacznę od omówienia i oceny wyników zawartych w Rozdziale 5. Ten fragment rozprawy jest w dużym stopniu oparty na wspólnym artykule Autorki rozprawy oraz Błażeja Miasojedowa i Wojciecha Rejchela (praca umieszczona w arXiv w 2020 r.). Jak już zaznaczyłem, wynik jest nowatorski. Zaproponowany algorytm (progowy LASSO) gwarantuje zgodną identyfikację struktury grafu CTBN (Wniosek 5.6) i zgodną estymację macierzy intensywności (Twierdzenie 5.5). Zakłada się obserwowalność trajektorii procesu w czasie dążącym do nieskończoności. Praktyczna użyteczność algorytmu jest potwierdzona symulacjami. Co prawda, założenia Tw. 5.5 i 5.6 (nierówności wiążące parametr λ i długość trajektorii T) są raczej niemożliwe do sprawdzenia, ale to jest cecha wszystkich znanych wyników tego typu. Na podkreślenie zasługuje zgrabna parametryzacja parametrów procesu (5.2), która w istocie nie zmniejsza ogólności, a pozwala traktować wpływ parametrów addytywnie. Jedynym elementem pozostawiającym pewien niedosyt jest (milczące) założenie o ergodyczności procesu. To, czy proces jest ergodyczny czy nie – zależy od struktury grafu, a ta jest właśnie przedmiotem estymacji.

Rozdział 6 uzupełnia rezultaty Rozdziału 5. Rozważa się przypadek, gdy proces jest obserwowany jedynie w dyskretnych, deterministycznych momentach czasu, z szumem losowym. Do maksymalizacji penalizowanej wiarygodności używa się stochastycznej wersji algorytmu *proximal gradient*. Gradient wiarygodności aproksymuje się metodą Monte Carlo, wykorzystując algorytm Rao i Teha (2013). Dowodzi się Twierdzenie 6.7, które zapewnia zbieżność do zbioru lokalnych ekstremów (ponieważ log-wiarygodność nie jest funkcją wklęsłą, tylko to można zagwarantować). Wykorzystuje się wyniki Majewskiego i in. (2018) o algorytmie *proximal gradient* i własności algorytmu Rao i Teha. Całość algorytmu jest bardzo skomplikowana. Sprawdzenie założeń wykorzystywanych w Rozdziale 6 twierdzeń wymagało od Autorki bardzo precyzyjnej pracy i zaawansowanych technik. Otrzymany rezultat jest elegancki i wykończony.

Na końcu kilka uwag o Rozdziale 4, poświęconym klasycznym (statycznym) sieciom bayesowskim. Autorka połączyła istniejący algorytm MCMC poszukujący podziału wierzchołków na „warstwy” z metodą penalizowanej największej wiarygodności. Kluczowe jest spostrzeżenie, że dla ustalonych „warstw” problem rozpoznawania krawędzi w zasadzie sprowadza się do selekcji cech w modelu regresji. W ten sposób zbudowała efektywny algorytm, działający równo w przypadku sieci gaussowskich jak i dla sieci zbudowanych ze zmiennych dyskretnych. W pierwszym przypadku, wykorzystując LASSO dla modeli liniowych, udowodniła eleganckie Twierdzenie 4.1 i Wniosek 4.2 (zgodność estymacji i zgodność selekcyjna, odpowiednio). W przypadku sieci dyskretnych algorytm działa podobnie (LASSO dla regresji logistycznej), ale wyników teoretycznych jeszcze nie otrzymano. Rozdział 4 jest nieco niedopracowany z punktu widzenia redakcyjnego, ale wyniki w nim zawarte są solidne i wartościowe.

Uwagi krytyczne

Moje uwagi krytyczne dotyczą tylko Rozdziałów 3 i 4. Rozdział 3 jest, moim zdaniem, słabo powiązany z najważniejszą częścią rozprawy. Ma charakter przeglądowy i nie zawiera nowych wyników. Zagadnienia wnioskowania i estymacji parametrów dla sieci bayesowskich są, oczywiście, ciekawe i ważne. Większość materiału zebranego w Rozdziale 3 nie jest jednak wykorzystana w dalszych częściach rozprawy i nie jest niezbędna do ich zrozumienia. Według mnie, spora część Rozdziału 3 jest po prostu zbędna.

Z kolei Rozdział 4 jest, odwrotnie, zdecydowanie zbyt lakoniczny. Proponowany w tym rozdziale algorytm jest w istocie drugim krokiem większego algorytmu. Pierwszym krokiem jest procedura „warstwowania” oparta na pracy (Kuipers i Moffa, 2017). Opis tego pierwszego kroku jest niewystarczający. Z tekstu Rozdziału 4 można się tylko dowiedzieć, że algorytm Kuipersa i Moffy przeszukuje przestrzeń podziałów aby znaleźć „najbardziej prawdopodobną reprezentację warstwową”, ale o jakim rozkładzie prawdopodobieństwa mowa, nie wiadomo. Trudno zrozumieć i ocenić całość algorytmu omawianego w Rozdziale 4 bez przestudiowania pracy Kuipersa i Moffy (może także i pracy Grzegorzcyka i Husmeiera, 2008). Rozdział 4 robi wrażenie spisanych w pośpiechu nowych wyników, w redakcyjnie niedopracowanej postaci. Muszę jednak dodać, że najważniejszy fragment Rozdziału 4 jest wyodrębnioną całością, którą można zrozumieć bez szerszego kontekstu. Mam na myśli dowód Stwierdzenia (Proposition) 4.3 i pomocniczych lematów. To jest wynik dotyczący algorytmu LASSO dla modeli liniowych. Autorka używa tu innych, lokalnych oznaczeń i ten fragment jest samodzielną całością. Teza Stwierdzenia 4.3 jest typowa dla wyników tego typu (licznych w literaturze). Brakuje mi trochę autorskiego komentarza o stopniu nowości tego wyniku (niezależnie od Twierdzenia 4.1 i Wniosku 4.2, które podsumowują własności „dużego” algorytmu).

Podsumowanie i wnioski

Rozprawa zawiera nowe, istotne wyniki dotyczące dobrze postawionego problemu o znaczeniu teoretycznym i praktycznym. Narzędzia matematyczne użyte przez Autorkę świadczą o opanowaniu warsztatu i orientacji w literaturze. Uwagi krytyczne, które powyżej zamieściłem w tej recenzji, dotyczą wyłącznie pewnych niedoskonałości w redakcji pracy, a nie jej zawartości merytorycznej. Według mnie, jest to rozprawa na wysokim poziomie, przekraczającym ustawowe i zwyczajowe wymagania. Niestety, wartościowe wyniki Maryii Shpak nie zostały dotąd opublikowane. Zarówno wspólny artykuł w *arXiv* jak i (nowsze zapewne) wyniki dotyczące statycznych BN-ów (Rozdział 4) powinny się znaleźć w dobrym piśmie, aby mogły być docenione przez specjalistów.

Rozprawa bez wątplenia spełnia warunki stawiane pracom doktorskim z matematyki. Wnoszę o dopuszczenie Pani Maryii Shpak do dalszych etapów przewodu doktorskiego.

Wojciech Niemiro