Nr rejestracyjny: 2022/45/B/ST1/00545; Kierownik projektu: dr hab. Bartosz Wojciech Kołodziejek

**Selekcja modelu dla kolorowych Gaussowskich modeli grafowych - podejście Bayesowskie i częstościowe.**

Do rozwoju analizy danych wielowymiarowych konieczne jest tworzenie nowych specjalistycznych narzędzi probabilistycznych. Przykładem sytuacji, w której „klasyczne" narzędzia probabilistyczne i statystyczne zawodzą, jest badanie danych genetycznych, gdzie liczba obserwacji (osób) jest zazwyczaj dużo mniejsza niż liczba zmiennych (genów). Jest to typowa sytuacja we współczesnych zadaniach stawianych przed uczeniem maszynowym. Aby tego typu modelowanie było możliwe, konieczna jest redukcja wymiaru parametrów modelu.

Taką redukcję można uzyskać poprzez nadanie odpowiedniej struktury warunkowej niezależności rozważanych modeli, definiując w ten sposób tzw. *modele grafowe*. Modele grafowe są podstawowym narzędziem wykorzystywanym do formułowania problemów związanych z uczeniem maszynowym oraz znajdują się na przecięciu wielu dziedzin, w szczególności statystyki, informatyki, teorii prawdopodobieństwa oraz teorii grafów. Stanowią podstawę nowoczesnych metod statystycznych stosowanych w takich obszarach jak diagnostyka medyczna, rozpoznawanie obrazu i mowy, przetwarzanie języka naturalnego i wielu innych.

W modelach grafowych wierzchołki grafu reprezentują zmienne losowe, jego krawędzie natomiast odpowiadają za opis zależności pomiędzy tymi zmiennymi. Na rozkład łączny wektora losowego nakłada się strukturę warunkowej niezależności (tzw. struktury Markowskie). Struktury takie wygodnie są opisywane przez grafy oraz pozwalają istotnie zmniejszyć wymiarowość problemu - liczbę parametrów należy wtedy odnosić do rozmiaru największej kliki w grafie, a nie do wszystkich zmiennych jak w klasycznym modelu. Często jednak graf opisujący strukturę Markowską nie jest dostatecznie rzadki (rozmiar największej kliki nadal znacznie przekracza wielkość próby) i takie postępowanie nie pozwala na wiarygodne badanie macierzy kowariancji. W takiej sytuacji z pomocą przychodzą tzw. *kolorowe modele grafowe*. W kolorowych modelach grafowych, oprócz struktury Markowskiej, na macierze koncentracji lub korelacji cząstkowych zadawane są pewne symetrie. Trzy typy takich symetrii zostały wprowadzone przez Højsgaarda i Lauritzena w celu opisania sytuacji, w których niektóre elementy macierzy koncentracji lub korelacji cząstkowej są w przybliżeniu równe. Równości te są wygodnie reprezentowane za pomocą grafu o kolorowych wierzchołkach i krawędziach. Uwzględnienie tych dodatkowych symetrii zmniejsza liczbę parametrów niezbędnych do oszacowania. Jest to szczególnie istotne, gdy dane są wielowymiarowe, tzn. gdy liczba zmiennych jest znacznie większa niż liczba obserwacji.

Głównym wyzwaniem w tej tematyce jest problem selekcji modelu czyli wyboru modelu (grafu i jego kolorowania), który najlepiej pasuje do zaobserwowanych danych. Istnieje wiele podejść do tego problemu, jednak niewiele z nich można zastosować dla wielowymiarowych danych. Co więcej, istniejące metody często nie mają teoretycznych podstaw i nie oferują statystycznych gwarancji. W naszej pracy zamierzamy zająć się dwoma nurtami związanymi z problemem selekcji modelu dla kolorowych modeli grafowych. Będziemy badali podejście Bayesowskie i częstościowe. Uzyskane wyniki przyczynią się zarówno do rozwoju metod badawczych, jak i narzędzi statystycznych.