

Sparse graphical models and the problem of precision matrix estimation

Michał Makowski¹

¹) Uniwersytet Wrocławski
michalmakowski(at)outlook(dot)com

Zależności pomiędzy obserwacjami w dużych zbiorach danych mogą być reprezentowane za pomocą grafów, w którym wierzchołki odpowiadają zmiennym, a istnienie krawędzi pomiędzy zmiennymi odpowiada pewnej zależności. Modele graficzne zyskują na popularności, szczególnie popularne stają się sieci bayesowskie (Bayesian network) oraz losowe pola Markowa (Markov random field).

W przypadku, gdy zmienne pochodzą z wielowymiarowego rozkładu normalnego, graf reprezentuje relacje warunkowej zależności; dokładnie mówiąc zależność dwóch zmiennych losowych warunkowanej wszystkimi pozostałymi zmiennymi. W takich grafach liczba krawędzi wychodzących z danego wierzchołka może być traktowana jako miara istotności danej zmiennej. Mówimy wtedy o Gaussowskich modelach graficznych, w których to struktura grafu reprezentowana jest przy pomocy macierzy precyzji, tj. odwrotności macierzy kowariancji. Wówczas niezerowe współrzędne macierzy precyzji odpowiadają istnieniu krawędzi w grafie.

Gdy liczba obserwacji w próbie jest porównywalna lub mniejsza niż liczba zmiennych, klasyczne metody estymacji największej wiarygodności zawodzą, tzn. MLE nie istnieje lub ma dużą wariancję. Zajęliśmy się tym problemem znanym jako estymacja rzadkich macierzy precyzji. Przeanalizowaliśmy i porównaliśmy dwie metody regularyzacji, gLasso i gSLOPE, które stabilizują rozwiązanie uzyskiwane przy użyciu MLE. Pierwsza jest dobrze znana metoda w tego typu problemach (stosuje się normę l-jeden aby ściągnąć współczynniki do zera, podobnie jak w Lasso), druga to nowa metoda bazująca na sortowanej normie l-jeden.

SLOPE zostało wprowadzone przez dr Małgorzatę Bogdan i in. jako nowe podejście do estymacji w modelach liniowych, gSLOPE został wprowadzony przez Piotra Sobczyka. W pracy zajęliśmy się implementacją obydwu metod przy pomocy algorytmu ADMM. Jako, że gSLOPE nie został wcześniej zaimplementowany jest to jego implementacja to główny wynik tej pracy. Porównaliśmy obydwa podejścia przy pomocy syntetycznych symulacji, pokazały one, że w wielu przypadkach gSLOPE wykazuje lepsze właściwości niż gLasso, w szczególności jeśli chodzi o krzywe ROC.