Wykorzystanie optymalizacji wypukłej przy wyborze zmiennych w rzadkiej regresji

Damian Brzyski

Zagadnienie wyboru istotnych zmiennych objaśniających w modelu regresji liniowej odgrywa ważną rolę w licznych problemach statystycznych. Testowanie wszystkich możliwych podzbiorów zmiennych objaśniających staje się czasowo nieefektywne już przy względnie małej liczbie danych, co czyni to zagadnienie nietrywialnym. Problem staje się szczególnie istotny w obliczu coraz powszechniejszej potrzeby analizy dużych zbiorów danych, w takich zróżnicowanych obszarach jak na przykład: badania genetyczne, analiza i przetwarzanie sygnałów czy nauczanie maszynowe.

W rozprawie doktorskiej przedstawię podejście wykorzystujące optymalizację wypukłą, prowadzące do metod, które mogą być efektywnie stosowane nawet przy bardzo licznych zbiorach danych. Rozważany będzie model regresji liniowej postaci $Y = X\beta + z$, gdzie $X \in M(n,p)$ jest macierzą danych, z zaburzeniem losowym, natomiast β szukanym wektorem, który (z założenia) ma bardzo mało niezerowych współrzędnych w stosunku do wymiaru (β jest wektorem rzadkim). Podstawowym celem będzie wskazanie indeksów odpowiadających niezerowym współrzędnym β (wskazanie nośnika rozwiązania), co będzie utożsamiane z selekcją istotnych zmiennych w modelu.

Gdy głównym celem jest predykcja nośnika rozwiązania, przy ocenie skuteczności danej metody sensowniejsze od powszechnie stosowanej oceny błędu średniokwadratowego wydają się być:

- Frakcja Fałszywych Odkryć (False Discovery Ratio, FDR), którą definiuje się jako wartość oczekiwaną frakcji fałszywie zidentyfikowanych niezerowych współrzędnych w stosunku do wszystkich niezerowych współrzędnych zidentyfikowanych przez metodę,
- Moc (Power, P), definiowaną jako frakcję prawdziwie zidentyfikowanych niezerowych współrzędnych w stosunku do wszystkich niezerowych współrzędnych prawdziwego rozwiązania.

Przy ocenie metody istotna jest kontrola FDR na niskim poziomie przy jednoczesnej maksymalizacji P. Szczególnie godnie uwagi wydają się być metody pozwalające kontrolować FDR na pewnym poziomie zdefiniowanym przez badacza, zapewniając jednocześnie możliwie wysoką wartość P.

Przy założeniu sytuacji ortogonalnej, tj. przy warunku, że $X^T X = Id$, bardzo dobrze sprawdza się opublikowana w 1995 roku procedura Benjaminiego – Hochberga. Umożliwia ona kontrolę wartości FDR przy założeniu, że rozważany błąd ma rozkład ciągły na poziomie $FDR = \frac{(p-k)q}{p}$, gdzie k jest liczbą niezerowych współrzędnych w prawdziwym rozwiązaniu. Metoda Benjaminiego – Hochberga nie daje jednak informacji o współrzędnych niezerowych estymowanego rozwiązania (podaje jedynie jego nośnik) oraz występują problemy z rozszerzeniem tej metody do sytuacji nieortogonalnej. Najnowszy trend w badaniach nad kontrolą FDR w rzadkiej regresji to wykorzystanie metod optymalizacji wypukłej i definiowanie estymatora jako argumentu, dla którego pewna funkcja wypukła osiąga minimum. Podejście to zaowocowało bardzo obiecującą metodą, SLOPE [1], definiującą estymator jako

$$\arg\min_b \ \frac{1}{2}\|y - Xb\|_2^2 + J_\lambda(b), \qquad \text{(SLOPE)}$$

gdzie $J_\lambda(\cdot)$ jest pewną zdefiniowaną rodziną norm zależnych od wektora λ.

Istnieją teoretyczne wyniki dla zaproponowanej metody dające górne oszacowanie FDR w sytuacji ortogonalnej, takiej samej postaci jak w procedurze Benjaminiego – Hochberga, jednak w przeciwieństwie do tej procedury, SLOPE można rozpatrywać natychmiastowo również w sytuacji nieortogonalnej.

Jeżeli wszystkie współrzędnie wektora λ są równe, problem SLOPE sprowadza się do znanego zagadnienia LASSO, dla którego estymator dany jest jako:

$$\arg\min_b \ \frac{1}{2}\|y - Xb\|_2^2 + \Lambda\|b\|_1. \qquad \text{(LASSO)}$$

Istnieje związana w pewnym sensie z LASSO inna metoda optymalizacyjna - Dantzig Selector [2] (DS). Zastosowanie tej metody w kontekscie identyfikowania genów mających istotny wpływ na cechy fenotypowe było testowane w [3]. W rozprawie doktorskiej pokażę związek pomiędzy LASSO oraz DS i wykorzystam go w bardziej ogólnym podejściu, czego wynikiem w kontekście SLOPE będzie nowa metoda optymalizacyjna - Ordered Dantzig Selector (ODS). Przedstawione zostaną teoretyczne wyniki dotyczące własności ODS, w tym własność kontroli FDR na zadanym poziomie w sytuacji

ortogonalnej. Zostaną również przeprowadzone symulacje mające na celu przetestowanie metod kontrolujących FDR w zagadnieniu lokalizacji genów w populacjach ludzkich za pomocą metod asocjacyjnych (Genome Wide Association Studies, GWAS). Zostaną przy tym wykorzystanie dane generowane sztucznie, jak również autentyczne dane genetyczne.

Literatura

[1] Bogdan, M., van den Berg, E., Su, W., Candès E. J., Statistical Estimation and Testing via the Ordered ℓ_1 Norm, arXiv:1310.1969v2, 2013.

[2] Candès, E. J., Tao, T., The Dantzig Selector: Statistical estimation when p is much larger than n *The Annals of Statistics*, **6**: p. 2313–2351, 2007.

[3] Brzyski D., The Dantzig Selector in localizing influential genes, preprint dostępny na www.ssdnm.mimuw.edu.pl.

Podpis Promotora

.....

Małgorzata Bogdan

Podpis doktoranta

Damian Brzyski