Abstrakt

Ilość danych generowanych przez ludzkość rośnie obecnie w tempie eksponencjalnym. Ten wzrost nie dotyczy jedynie większej liczby obserwacji. Jednocześnie jesteśmy w stanie dokonać większej liczby pomiarów. Ta ogromna ilość danych jest źródłem wielu wyzwań. Nas interesują zwłaszcza te, które pojawiają się w związku z tym, że liczba zmiennych p jest większa niż liczba obserwacji n. W takim wypadku większość klasycznych modeli statystycznych staje się nieidentyfikowalnymi lub boryka się z problemami bardzo wysokiej wariancji estymatorów. W związku z tym pojawia się potrzeba konstrukcji nie-nadzorowanych metod służących do redukcji wymiaru danych. To właśnie nich dotyczy ta praca.

Najpierw skupimy się na niezbędnym wstępie matematycznym. Zaczyna się on od metod optymalizacji wypukłej. Dalej dotyczy problemu wielokrotnego testowania, a także metod regularyzacyjnych na przykładzie modelu liniowego. W dalszej części wprowadzamy metodę Analizy Składowych Głównych (PCA) i problemu klastrowania podprzestrzennego (subspace clustering). Wstęp kończy się na podaniu podstawowych definicji związanych z Gaussowskimi modelami graficznymi.

Następnie przechodzimy do zaprezentowania trzech nowych, zaproponowanych metod. Pierwszą jest „Penalizowana, częściowo scałkowana funkcja wiarogodności" (Penalized semi-integrated likelihood - PESEL) - nowe kryterium do wyboru liczby składowych głównych w problemie PCA. Dowodzimy jego zgodności przy pewnych założeniach, a także pokazujemy w symulacjach, że ma ono bardzo dobre własności w porównaniu z innymi nowoczesnymi metodami. Drugą stworzoną metodą jest „Klastrowanie zmiennych i identyfikacja zmiennych ukrytych" (multivariate latent variables clustering, MLCC). Metoda ta służy do rozwiązywania problemu klastrowania poprzez identyfikację niskowymiarowych przestrzeni reprezentujących odpowiednie grupy zmiennych. Wyprowadzamy nową wersję zmodyfikowanego Bayesowskiego kryterium informacyjnego (mBIC) do estymacji liczby skupień i proponujemy algorytm przeszukujący reprezentacyjną grupę partycji w celu aproksymacji modelu maksymalizującego mBIC. Porównujemy działanie tego algorytmu z innymi nowoczesnymi metodami klastrowania przestrzennego za pomocą obszernych symulacji komputerowych, które ilustrują bardzo dobre własności proponowanej metody. Na końcu wprowadzamy nową metodę graficznego SLOPE (gSLOPE) do estymacji rzadkiej macierzy precyzji dla wielowymiarowego rozkładu normalnego (estymujemy graficzny model Gaussowski). Definiujemy wypukłę funkcję celu będącą funkcją wiarygodności pomniejszoną o karę opartą o posortowaną normę l_1 (sorted l_1) wektora elementów macierzy precyzji. Wprowadzamy dwie różne strategie na wybór tej kary (ciągi λ). Pokazujemy, że dla jednego z nich mamy, przy pewnych warunkach, kontrolę blokowego błędu I rodzaju. W symulacjach pokazujemy, że w pewnych wypadkach gSLOPE zachowuje się lepiej niż szeroko używana metoda glasso.