Linear discriminant analysis with high dimensional mixed variables

Binyan Jiang

The Hong Kong Polytechnic University
E-mail: by.jiang@polyu.edu.hk

Abstract: With the rapid development of modern measurement technologies, datasets containing both discrete and continuous variables are more and more commonly seen in different areas, and in particular, the dimensions of the discrete and continuous variables can oftentimes be very high. Though discriminant analysis for mixed variables under the traditional fixed dimension setting has been well studied since the 80's, promising approaches taking into account both the high dimensionality and the mixing nature of the data sets are still missing. In this paper, we aim to developing a simple yet useful classification rule that addresses both the high dimensionality and the mixing nature of the variables simultaneously. Our framework is built on a location model, under which we further propose a semiparametric formulation for the optimal Bayes rule. We show that the optimal classification direction and the intercept in the optimal rule can be estimated separately. Efficient direct estimation schemes are then developed to obtain consistent estimators of the discriminant components. Asymptotic results on the estimation accuracy and the misclassification rates are established, and the competitive performance of the proposed classifier is illustrated by simulation and real data studies .