Sparse Quadratic Discriminant Analysis
For High Dimensional Data

Quefeng Li and Jun Shao

University of Wisconsin, Madison

Abstract: Many contemporary studies involve the classification of a subject into two classes based on \( n \) observations of the \( p \) variables associated with the subject. Under the assumption that the variables are normally distributed, the well-known linear discriminant analysis (LDA) assumes a common covariance matrix over the two classes while the quadratic discriminant analysis (QDA) allows different covariance matrices. When \( p \) is much smaller than \( n \), even if they both diverge, the LDA and QDA have the smallest asymptotic misclassification rates for the cases of equal and unequal covariance matrices, respectively. However, modern statistical studies often face classification problems with the number of variables much larger than the sample size \( n \), and the classical LDA and QDA may perform poorly. In fact, we give an example in which the QDA performs as poorly as random guessing even if we know the true covariances. Under some sparsity conditions on the unknown means and covariance matrices of the two classes, we propose a sparse QDA having the smallest asymptotic misclassification rate conditional on the training data. For illustration of application, we discuss an example of classifying normal and tumor colon tissues based on a set of 2000 genes and a sample of size 62. A simulation is also conducted to check the performance of the proposed method.

Key words and phrases: Classification, High dimensionality, Normality, Smallest asymptotic misclassification rate, Sparsity estimates, Unequal covariance matrices