Principal Discriminant Variate (PDV) Method for Classification of Multicollinear Data: Application to Diagnosis of Mastitic Cows Using Near-Infrared Spectra of Plasma Samples

JIAN-HUI JIANG,[1,2] ROUMIANA TSENKOVA [3], RU-QIN YU [2] AND YUKIHIRO OZAKI [1]*

1. Department of Chemistry, School of Science, Kwansei-Gakuin University, Uegahara, Nishinomiya 662-8501, Japan
2. College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China
3. Department of Enviromental Information and Bioproduction Engineering, Faculty of Agriculture, Kobe University, Noda-ku, Kobe 657-8501, Japan

In linear discriminant analysis there are two important properties concerning the effectiveness of discriminant function modeling. The first is the separability of the discriminant function for different classes. The separability reaches its optimum by maximizing the ratio of between-class to within-class variance. The second is the stability of the discriminant function against noises present in the measurement variables. One can optimize the stability by exploring the discriminant variates in a principal variation subspace, i. e., the directions that account for a majority of the total variation of the data. An unstable discriminant function will exhibit inflated variance in the prediction of future unclassified objects, exposed to a significantly increased risk of erroneous prediction. Therefore, an ideal discriminant function should not only separate different classes with a minimum misclassification rate for the training set, but also possess a good stability such that the prediction variance for unclassified objects can be as small as possible. In other words, an optimal classifier should find a balance between the separability and the stability. This is of special significance for multivariate spectroscopy-based classification where multicollinearity always leads to discriminant directions located in low-spread subspaces. A new regularized discriminant analysis technique, the principal discriminant variate (PDV) method, has been developed for handling effectively multicollinear data commonly encountered in multivariate spectroscopy-based classification. The motivation behind this method is to seek a sequence of discriminant directions that not only optimize the separability between different classes, but also account for a maximized variation present in the data. Three different formulations for the PDV methods are suggested, and an effective computing procedure is proposed for a PDV method. Near-infrared (NIR) spectra of blood plasma samples from mastitic and healthy cows have been used to evaluate the behavior of the PDV method in comparison with principal component analysis (PCA), discriminant partial least squares (DPLS), soft independent modeling of class analogies (SIMCA) and Fisher linear discriminant analysis (FLDA). Results obtained demonstrate that the PDV method exhibits improved stability in prediction without significant loss of separability. The NIR spectra of blood plasma samples from mastitic and healthy cows are clearly discriminated between by the PDV method. Moreover, the proposed method provides superior performance to PCA, DPLS, SIMCA and FLDA, indicating that PDV is a promising tool in discriminant analysis of spectra-characterized samples with only small compositional difference, thereby providing a useful means for spectroscopy-based clinic applications.