Discrimination with many variables

P. J. Brown[*], T. Fearn[†] and M.S. Haque[‡]

Abstract

Many statistical methods for discriminant analysis do not adapt well or easily to situations where the number of variables is large, possibly even exceeding the number of cases in the training set. We explore a variety of methods for providing robust identification of future samples in this situation. We develop a range of flexible Bayesian methods, and primarily a new hierarchical covariance compromise method, akin to regularized discriminant analysis.

Although the methods are much more widely applicable, the motivating problem was that of discriminating between groups of samples on the basis of their near infrared spectra. Here the ability of the Bayesian methods to take account of continuity of the spectra may be beneficial. The spectra may consist of absorbances or reflectances at as many as 1000 wavelengths, and yet there may be only tens or hundreds of training samples where both sample spectrum and group identity are known. Such problems arise in the food and pharmaceutical industries, for example the authentication of food (eg. detecting the adulteration of orange juice) and the identification of pharmaceutical drug ingredients. Our illustrating example concerns the discrimination of 39 microbiological taxa and 8 aggregate genera. Simulations also illustrate the effectiveness of the hierarchical Bayes covariance method.

We discuss a number of scoring rules both local and global for judging the fit of data to the Bayesian models, and adopt a cross- classificatory approach for estimating hyperparameters.

KEYWORDS: Discrimination; Bayesian Methods; Smoothing; Hierarchical covariances; Spectroscopy; Gaussian processes; Cross-Validation; Scoring Rules.

[*]University of Kent at Canterbury, Institute of Mathematics and Statistics, Cornwallis Building, Canterbury, CT2 7NF, UK

[†]University College London, UK

[‡]Queen Elizabeth Psychiatric Hospital, Birmingham