Sparse and Robust Methods for Discrimination in High Dimensions

Valentin Todorov[1], Peter Filzmoser[2]

[1] United Nations Industrial Development Organization, Vienna, Austria
[2] Vienna University of Technology, Vienna, Austria

Keywords: discrimination, LDA, robust, PLS, sparse

The goal of supervised classification is to establish rules on the basis of a training sample which can be used reliably for predicting the group membership of new observations and for evaluating the performance of these rules. Unfortunately most of the well developed conventional classification methods are inapplicable or produce poor results when applied to high dimensional data. In such cases singularity problems arise if the variables are highly correlated or if less observations than variables are available in one or more groups. This situation which occurs in many areas of modern research is the normal, generic case rather than an anomalous one - this is the domain of the so called High Dimension Low Sample Size (HDLSS) statistical analysis (a term coined by Marron [1]). A possible solution is to precede the discrimination by a dimensionality reduction step like Principal Component Analysis (PCA) or Partial Least Squares (PLS) and then derive the classification functions from the scores instead of directly using the original data.

In order to cope with the high vulnerability of the classical estimates to the presence of outliers we propose to apply robust methods in both steps (dimensionality reduction and discriminations). It is also advantageous to perform variable selection simultaneously with the dimension reduction, since usually many of the large number of variables are irrelevant or redundant in the classification context. With this approach we expect to improve the predictive performance of considered methods. We will review the available sparse and/or robust methods in the setup of two phase discrimination and will evaluate them on real and simulated data sets following in part the study carried out by Pires and Branco in [2]. The methods discussed here are implemented in `R` and will be available in the Object-oriented Framework for Multivariate Analysis [3] (package **rrcov**).

References

[1] Marron, J. S. and Todd, M. and Ahn, J., Distance weighted discrimination, *Journal of American Statistical Association*, 102:1267–1271, 2007.

[2] Pires, A.M. and Branco, J.A., Projection-pursuit approach to robust linear discriminant analysis, *Journal of Multivariate Analysis*, 101:2462–2485, 2010.

[3] Todorov, V. and Filzmoser, P., An object oriented framework for robust multivariate analysis, *Journal of Statistical Software*, 32:1–47, 2009.