# 1ères Rencontres R

2-3 juil. 2012
France

# Table des matières

## HiDimDA: An R package for Supervised Classification of High-Dimensional Data

**A. Pedro Duarte Silva**

Faculdade de Economia e Gestão & CEGE
Catholic University of Portugal / Porto
Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal
psilva@porto.ucp.pt

**Keywords** : Supervised Classification, Discriminant Analysis, High Dimensionality, Feature Selection.

Classical methods of supervised classification often assume the existence of a training data set with more observations than variables. However, nowadays many classification applications work with data bases where the total number of original features is larger, and often much larger, than the number of available data units. For instance, in microarray applications several thousand genes are usually collected on a few dozen individuals with known clinical conditions, in order to derive classification rules capable of supporting the diagnostic of future patients (see, e.g., Dudoit, Fridlyard and Speed (2002)). A similar pattern occurs in image recognition problems, where the information contained in hundreds of pixels is trained on a much smaller set of images belonging to well defined classes (Thomaz and Gillies (2005)).
Furthermore, in most high-dimensional classification problems the majority of the original features do not contribute to distinguish the underlying classes, and can have a large negative impact if forced into the resulting classification rules (Fan and Fan 2008). Nevertheless, the number of useful features is often still comparable to, or even larger than, the number of available training sample observations.
Therefore, effective classification methodologies for these applications require scalable methodologies of feature selection, and classification rules that can use sample information in a way that is not severely limited by the number of data units in the training sample. The most common strategy to deal with the latter problem is to adopt rules (e.g., Domingos and Pazzani (1997) Tibshirani et. al. (2003)) that treat all features independently, and ignore all sample information about their dependence structure. Recent proposals (e.g., Thomaz and Gillies (2005), Fisher and Sun (2011), Duarte Silva (2011)) try to surpass this limitation by relying on estimators of covariance matrices with good statistical properties when the number of used features is close to, or larger than, the training sample size.
In this presentation, I will describe the *HiDimDA* (High Dimensional Discriminant Analysis) R package, available on CRAN, that implements several routines and utilities for supervised $k$-group classification in high-dimensional settings. *HiDimDA* includes routines for the construction of classification rules with the above mentioned properties, methods for predicting new observations of unknown origin, as well as cross-validation and feature selection utilities. The selection routines of *HiDimDA* implement modern proposals for feature selection in high-dimensional classification problems (see e.g. Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Efron, (2004), Donoho and Jin (2008), Fan and Fan (2008)), which often rely on ideas originated form the related theory of large-scale hypothesis testing.
*HiDimDA* can be used to construct, apply and assess k-group ($k \geq 2$) classification rules for problems with several thousand variables, dealing effectively with the problems of high dimensionality, and including rules that do not ignore the dependence structure of the data.

## References

[1] Dudoit S., Fridlyard, J. and Speed TP. (2002). Comparison of discrimination methods for the classification of tumours using gene expression data. Journal of the American Statistical Association, **97**, 77-87.
[2] Thomaz, C.E. and Gillies, D.F. (2005). A maximum uncertainty lda-based approach for limited sample size problems with application to face recognition. In: 18th Brazilian Symposium on Computer Graphics and Image Processing. SIBGRAPI, 89-96.
[3] Fan J. and Fan, Y. (2008). High Dimensional Classification using Features Annealed Independence Rules. Annals of Statistics, **38**, 2605-2637.
[4] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, **29**, 103-130.
[5] Tibshirani, R., Hastie, B., Narismhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science, **18** (1), 104-117.
[6] Fisher, T.J. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Computational Statistics and Data Analysis, **55**, (5), 1909-1918.
[7] Duarte Silva, A.P. (2011). Two-group classification with high-dimensional correlated data: A factor model approach. *Computational Statistics and Data Analysis*, **55** (11), 2975-2990.
[8] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57** (1), 289-300.
[9] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29** (4), 1165-1188.
[10] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99** (465), 96-104.
[11] Donoho, D. and Jin, J. (2008). Higher criticism thresholding. Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, **105**, 14790-14795.