FADA: an R package for variable selection in supervised classification of strongly dependent data

Emeline Perthame 1,*, Chloé Friguet 2, David Causeur 1

1. Agrocampus Ouest - Applied Mathematics Department, Rennes, France
2. Laboratoire de Mathématiques de Bretagne-Atlantique (LMBA), Université de Bretagne-Sud, Vannes, France

*Contact author: perthame@agrocampus-ouest.fr

Keywords: High dimension, variable selection, dependent data, supervised classification, factor analysis

Handling dependence or not in feature selection is still an open question in supervised classification issues where the number of covariates exceeds the number of observations. Some recent papers surprisingly show the superiority of naive Bayes approaches based on an obviously erroneous assumption of independence (see [2]), whereas others recommend to infer on the dependence structure in order to decorrelate the selection statistics (see [6, 1]). In the classical Linear Discriminant Analysis (LDA) framework, the present talk first highlights the impact of dependence in terms of instability of feature selection. A second objective is to revisit the above issue using a flexible factor modeling for the covariance.

Latent components of dependence are introduced in the LDA model, conditionally on which a new Bayes consistency is defined. The linear Bayes classifier derived on factor-adjusted data, namely decorrelated data obtained by subtracting the effects of latent factors, is shown to be conditionally consistent. A procedure is then proposed for the joint estimation of the expectation and variance parameters of the model, based on an iterative algorithm which alternates the estimation of the fixed parameters of the supervised factor model and the latent factors. As in [5], the estimation method adapts an EM algorithm for factor models to the present supervised classification situation. The Factor-Adjusted Discriminant Analysis (FADA) method is compared to recent regularized Diagonal Discriminant Analysis approaches (DDA), assuming independence among features, and regularized LDA procedures, both in terms of classification performance and stability of feature selection.

The talk will also focus on a demonstration of an *R* package which implements FADA. The main function of the package provides an efficient way to decorrelate data before applying a feature selection procedure. Several methods of classification are available: DDA, penalized logistic regression [4], shrinkage discriminant analysis [1] which proposes a shrunken estimation of covariance matrix through James-Stein estimator and variable selection by controlling false non-discovery rate and LDA penalized by Lasso [3].

References

- [1] Ahdesmäki, M. and K. Strimmer (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics* 4, 503–519.
- [2] Bickel, P. and E. Levina (2004). Some theory for fisher's linear discriminant function, naive bayes, and some alternatives when there are many more variables than observations. *Bernoulli* 10(6), 989–1010.
- [3] Clemmensen, L., T. Hastie, D. Witten, and B. Ersbøll (2011). Sparse discriminant analysis. *Technometrics* 53(4), 406–413.
- [4] Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.
- [5] Friguet, C., M. Kloareg, and D. Causeur (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association* 104:488, 1406–1415.
- [6] Xu, P., G. Brock, and R. S. Parrish (2009). Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics and Data Analysis* 53, 16741687.