ClustOfVar : an R package for dimension reduction via clustering of variables. Application in supervised classification and variable selection in gene expression data

Marie Chavent [1,2], Robin Genuer [1,3], Vanessa Kuentz-Simonet [4], Benoit Liquet [5] and Jérôme Saracco [1,2]

[1]University of Bordeaux, France; [2]INRIA, Talence, France; [3]INSERM U897, Bordeaux, France; [4]IRSTEA, UR ADBX, Cestas, France; [5]MRC Biostatistics Unit, Cambridge, UK.

Scientific Context

In the genomics context, a question of interest is to link a large set of p predictors, e.g. gene expression, to a categorical dependant variable, e.g. recurrent disease.

- An important goal is to tackle the problem of dimension reduction for high-dimensional supervised classification.
- In usual case studies, the sample size n is moderate with $n << p$.

The main idea of the proposed approach splits into two steps.

- First we eliminate redundancy in the set of predictors by summarizing them in few synthetic variables. This dimension reduction step is done independently from the prediction step, i.e. without using any information on the dependant variable.
- Second, we select relevant synthetic variables in order to build the classifier providing the smallest error rate.

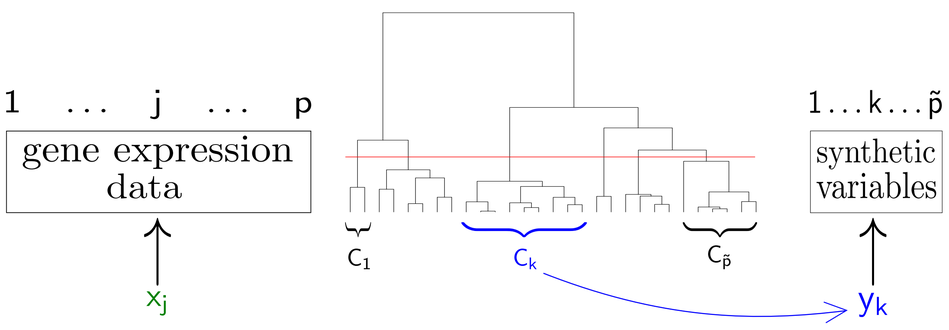
Methodology.

- Non supervised dimension reduction by clustering of variables with `ClustOfVar` [1].
  - We build simultaneously $\tilde{p}$ clusters of correlated predictors and their corresponding $\tilde{p}$ synthetic variables.
  - The p predictors are replaced by the $\tilde{p}$ synthetic variables.
- Construction of the classifier with a variable selection step.
  - The $\tilde{p}$ synthetic variables are used to construct a classifier (e.g. LDA, random forest,...).
  - A reduced model with $p^* < \tilde{p}$ synthetic variables is obtained (e.g. stepwise selection with Wilks test for LDA, or `VSURF` [2] for random forests).

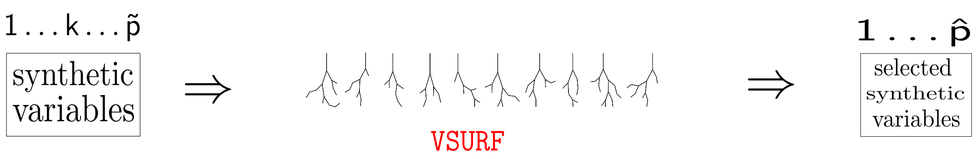
Methods

- Package `ClustOfVar`

$$\max \sum_{k=1}^{\tilde{p}} \sum_{x_j \in C_k} \text{cor}^2(x_j, y_k)$$



- Package `VSURF` (in construction)



  - Random forests: aggregation of a collection of randomized tree-based predictors
  - `VSURF`: data-driven procedure to automatically select the most important variables

References

1 M. Chavent, B. Liquet, V. Kuentz and J. Saracco. ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software*, **Vol. 50**, pp. 1-16, 2012.

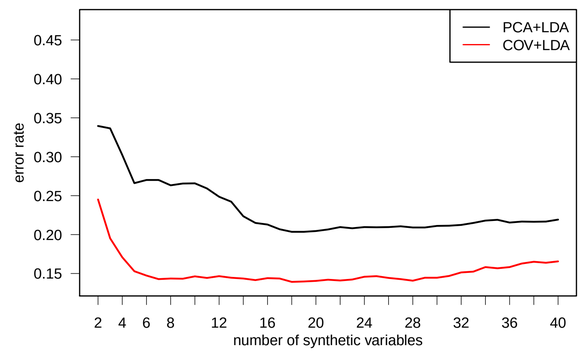
2 R. Genuer, J.-M. Poggi and C. Tuleau-Malot. Variable Selection using Random Forests. *Pattern Recognition Letters*, **Vol. 31**, pp.2225-2236, 2010.

Illustration on Gene expression data

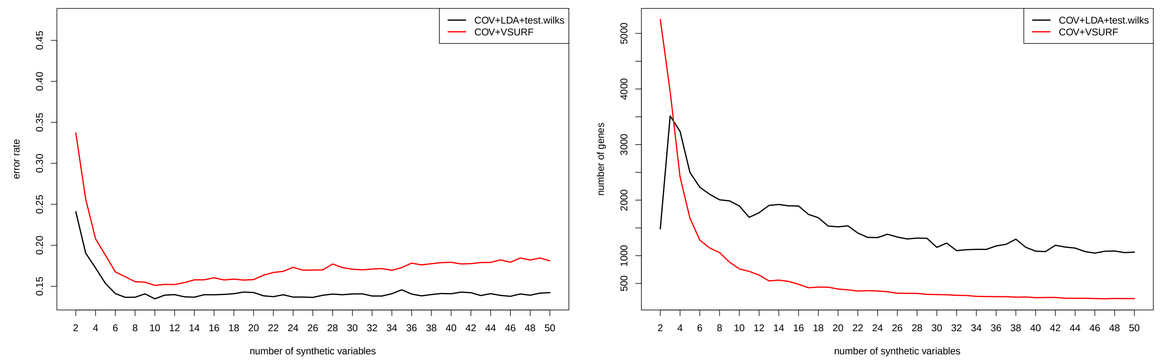
Prostate data.

- $n = 79$ patients all treated by radical prostatectomy : 37 where classified as recurrent and 42 as non-recurrent primary prostate tumor.
- Gene expression analysis was carried out using Affymetrics U133A human gene array and a prefiltered dataset contains $p = 7684$ genes.
- Data used in Stephenson et al. (2005), Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, **104**, 290-298.

A) Comparison of ClustOfVar (COV) and Principal Component Analysis (PCA) for dimension reduction before applying Linear Discriminant Analysis (LDA).

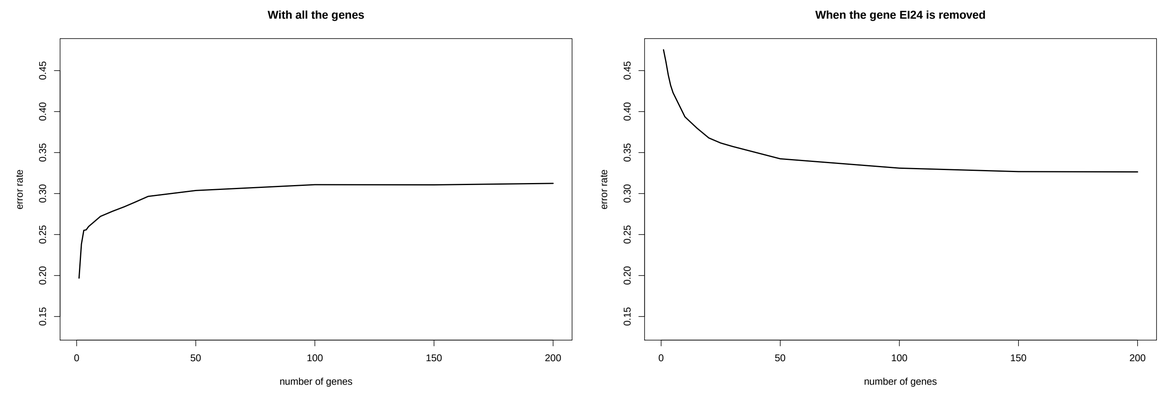


B) COV's synthetic variables selection in LDA and in random forests with VSURF.



- For k synthetic variables, $k^* < k$ are selected to provide the best error rates and we deduce the corresponding set of genes.
- Similar results are obtained when the highly discriminative gene etoposide induced 2.4 mRNA (EI24) is removed. It is not the case with SPLSDA.

C) Sparse PLS Discriminant Analysis (SPLSDA).



Highlights.

A) COV outperforms PCA in dimension reduction step.

B) COV+VSURF output : error rate of 16%, 4 synthetic variables selected among 13, 516 genes (93.5% of genes discarded) .

- \+ Small error rate.
- \+ Few synthetic variables.
- \+ Selection of groups of correlated genes.
- \- Number of selected genes usually large.

C) COV+VSURF is robust compared to SPLSDA.

COV+VSURF works for a mixture of numerical and categorical data : gene expression + clinical variables.

contact:marie.chavent@u-bordeaux2.fr