Feature selection for kernel methods in systems biology

Joint work with Jérôme Mariette, Rémi Flamary and Nathalie Vialaneix

12/02/2021

Outline

Introduction

What is a kernel?

Method

Applications

Introduction

The substantial development of high-throughput biotechnologies has rendered large-scale multi-omics datasets increasingly available.

Omics data features

- **High dimensional** data: sample number $\ll$ features number
- **Big data**: high throughput technologies (sequencing, screening, . . .)
- **Heterogeneous data**: count table, phylogenetic tree, graph, . . .

	Sample_a	Sample_b	Sample_c
OTU_1	3	3	3
OTU_2	5	2	3
OTU_3	2	0	2
OTU_4	4	3	0
OTU_5	0	2	2
OTU_6	0	2	2

Introduction

Why using kernels?

- Allow to analyse **heterogeneous** datasets
- Give access to a large number of similarity / dissimilarity measures

Issues

- Kernel methods usually suffer from a lack of interpretability
- The initial description in terms of features is lost
- Information of thousands of descriptors is often summarized in a few similarity measures, which can be influenced by a large number of irrelevant descriptors

Feature selection is a widely used strategy to address these issues
It consists in selecting the most promising features during or prior the analysis

Introduction

Objectives

- Propose a **feature selection** algorithm for kernel methods in two rarely met settings:
  - unsupervised (exploratory) learning
  - multiple output or non numerical output predictions

Outline

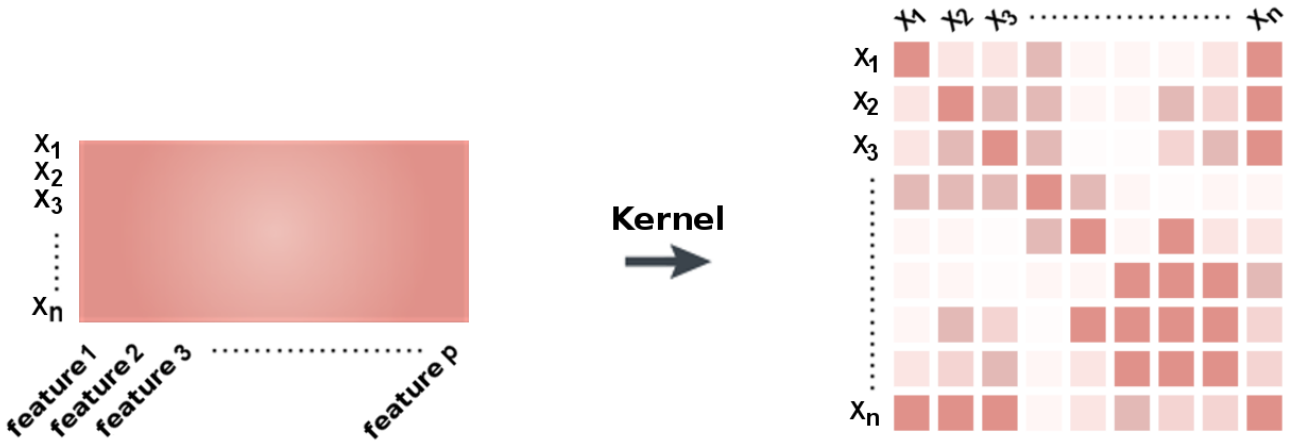
Introduction

What is a kernel?

Method

Applications

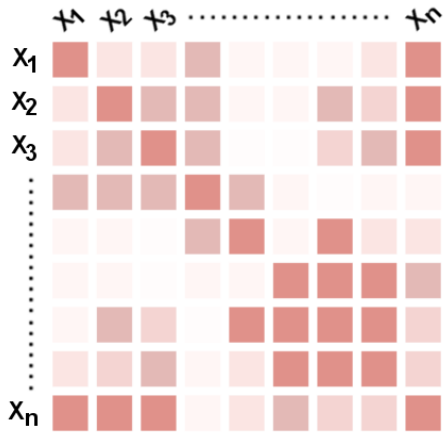
What is a kernel?

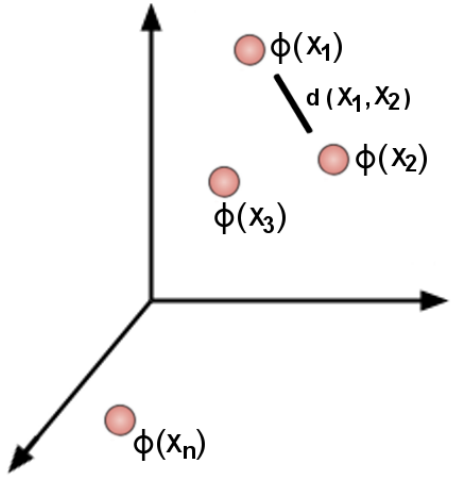


Kernels

The characteristics on the n samples $(x_i)_i$ are summarized by **pairwise similarities**.
More formally: $n \times n$ matrix K is symmetric and positive semidefinite

What is a kernel?





Representer theorem

$\exists$ a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ st:

$$K_x(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}.$$

Outline

Introduction

What is a kernel?

Method

Applications

Contribution

We propose a feature selection algorithm that explicitly takes advantage of the kernel structure.

Main idea

- The algorithm simultaneously learns weights w_j for each feature j that correspond to the feature's relevance.
- The computation of the weights are obtained simultaneously for all features, in order to better account for colinearities or redundancies between features.

Unsupervised Kernel Feature Selection

We consider a set of n observations $(\mathbf{x}_i)_{i=1,\dots,n}$, taking values in $\mathcal{X} = \mathbb{R}^p$, and described by a kernel $K_x : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$.

Principle

Select a subset of $d \ll p$ features aiming at preserving the topology structure of K_x (i.e. the relations/similarities between individuals as described by K_x) .

We introduce a vector of p weights $\mathbf{w} = (w_j)_{j=1,\dots,p}$ taking values in $\{0,1\}^p$ such that $w_j = 1$ is equivalent to select feature j.
Definition of a new kernel K_x^w:

$$K_x^w(\mathbf{x}_i, \mathbf{x}_{i'}) = K_x(\mathbf{w} \cdot \mathbf{x}_i, \mathbf{w} \cdot \mathbf{x}_{i'})$$

Unsupervised Kernel Feature Selection

Optimization problem

$$\underset{\mathbf{w}\in\{0,1\}^p}{\operatorname{argmin}} \|\mathbf{K}_x^w - \mathbf{K}_x\|_F^2$$

$$\text{s.t} \sum_{j=1}^{p} w_j \leq d$$

Continuous relaxation

$$\mathbf{w}^* := \underset{\mathbf{w}\in(\mathbb{R}^+)^p}{\operatorname{argmin}} \|\mathbf{K}_x^w - \mathbf{K}_x\|_F^2 + \lambda\|\mathbf{w}\|_1,$$

where $\lambda > 0$ is a penalization parameter.

The relaxed optimization problem is non-convex and non-smooth. We solve it using **proximal gradient descent** that is well adapted to ℓ_1 regularized problems.

Kernel output feature selection

A set of observations $(y_i)_{i=1,\dots,n}$ is associated to the $\mathbf{x}_i$. They take values in an arbitrary space, $\mathcal{Y}$, and are well described by another kernel, $K_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- **Example of outputs**: multiple numerical variables, multiple class variables, graphs, time series, nodes in a graph

Similarly to K_x, the feature map associated with K_y is denoted by $\psi : \mathcal{Y} \to \mathcal{F}_y$.

Objective

Select a subset of $d \ll p$ features that best explain the way the $(y_i)_i$ relate to each other as described by K_y.

Kernel output feature selection

Association problem between $(\mathbf{x}_i)_i$ and $(y_i)_i$

Learning a function $h : \mathbb{R}^p \to \mathcal{F}_y$ that predicts the output feature vector $\psi(y_i) \in \mathcal{F}_y$

In previous work [Brouard et al., 2016], the function h was learnt by solving:

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \|h(\mathbf{x}_i) - \psi(y_i)\|^2_{\mathcal{F}_y} + \lambda_1 \|h\|^2_{\mathcal{H}}$$

with h of the following form: $h(\mathbf{x}_i) = V\phi(\mathbf{x}_i)$.

Kernel output feature selection

Similarly to the unsupervised feature selection, we introduce weights $\mathbf{w} \in (\mathbb{R}^+)^p$ to be jointly learned with h:

$$\min_{h \in \mathcal{H}, \mathbf{w} \in (\mathbb{R}^+)^p} f(h, \mathbf{w}) + \lambda_1 \|h\|^2_{\mathcal{H}} + \lambda_2 \|\mathbf{w}\|_1,$$

where $f(h, \mathbf{w}) = \sum_{i=1}^{n} \|h(\mathbf{w} \cdot \mathbf{x}_i) - \psi(y_i)\|^2_{\mathcal{F}_y}$.

This optimization problem is solved using an iterative algorithm alterning optimization of $\mathbf{w}$ (similar to unsupervised framework) and optimization of h (using kernel trick).

Outline

Introduction

What is a kernel?

Method

Applications

Compared methods

- 2 methods based on the computation of a score:
  - the Laplacian score, denoted by **lapl** [He et al., 2005];
  - **SPEC** [Zhao and Liu, 2007];
- 3 methods based on a learning approach constrained to a sparse representation and designed for clustering:
  - **MCFS** [Cai et al., 2010];
  - **NDFS** [Li et al., 2012]
  - **UDFS** [Yang et al., 2011];
- 1 method based on neural networks:
  - **Autoencoder** [Abid et al., 2019]

Results UKFS

3 datasets

- **Carcinom**: expression of 9,182 genes obtained from 174 samples, 11 *a priori* groups
- **Glioma**: expression of 4,434 genes obtained from 50 samples, 4 *a priori* groups
- **Koren**: abundance of 973 OTUs collected from 43 samples, 3 *a priori* groups

Results UKFS

Principle

Methods ability to recover the dataset underlying classification structure using only a small number of features.

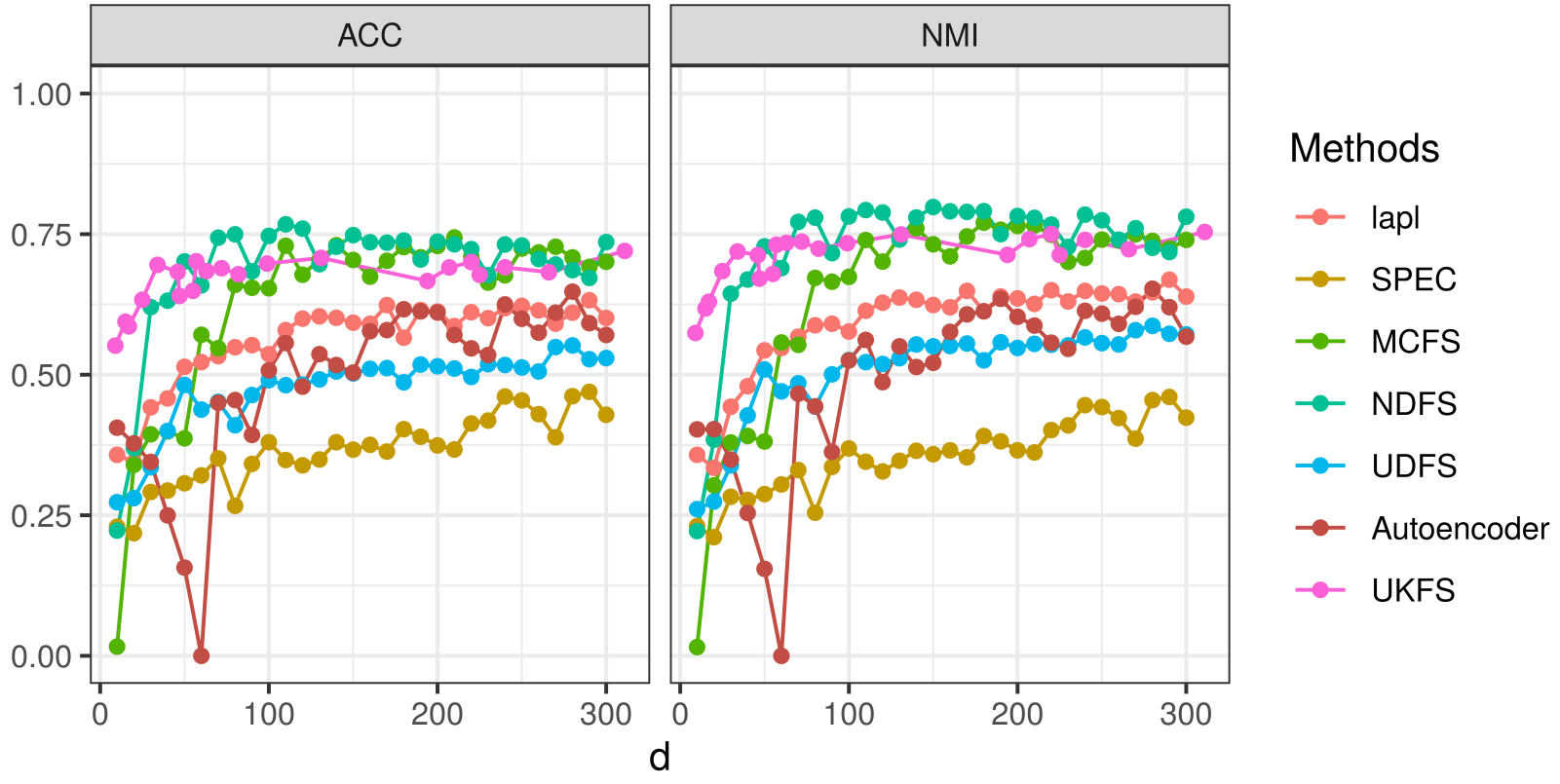
1. Select k features with increasing values of $k \in \{10, 20, ..., 290, 300\}$ (for **UKFS**, k is given by the number of selected features when increasing the regularization parameter, λ)
2. Repeat 20 times the kernel k-means algorithm
3. Evaluate the clustering relevance (NMI and ACC)

Results UKFS

	lapl	SPEC	MCFS	NDFS	UDFS	Autoencoder	UKFS
				"Carcinom" ($n=174, p=9\,182$)			
ACC (10)	0.36 (0.03)	0.23 (0.17)	0.02 (0.07)	0.22 (0.28)	0.27 (0.07)	0.41 (0.21)	**0.55** (0.04)
NMI (10)	0.36 (0.02)	0.23 (0.17)	0.02 (0.07)	0.22 (0.28)	0.26 (0.06)	0.40 (0.21)	**0.57** (0.03)
ACC (300)	0.60 (0.05)	0.43 (0.11)	0.70 (0.05)	**0.74** (0.06)	0.53 (0.05)	0.57 (0.06)	0.72 (0.07)
NMI (300)	0.64 (0.04)	0.42 (0.10)	0.74 (0.04)	**0.78** (0.03)	0.57 (0.03)	0.57 (0.05)	0.75 (0.05)
ACC AUC	164.02 (3.14)	106.52 (6.99)	184.17 (7.23)	200.88 (7.78)	138.48 (4.13)	143.13 (4.30)	**206.55** (5.62)
NMI AUC	172.50 (3.00)	103.66 (6.75)	189.96 (6.96)	212.46 (7.78)	148.78 (3.72)	145.09 (4.72)	**218.96** (3.82)
COR AUC	28.14	30.75	29.56	27.49	30.30	33.18	**24.75**
CPU time	**0.25** (0.04)	2.47 (0.39)	11.69 (5.21)	6,162 (305)	99,138 (2,913)	>4 days	326 (52)
				"Glioma" ($n=50, p=4\,434$)			
ACC (10)	**0.66** (0.04)	0.41 (0.01)	0.60 (0.01)	0.46 (0.04)	0.47 (0.03)	0.53 (0.04)	0.53 (0.06)
NMI (10)	**0.50** (0.03)	0.16 (0.01)	0.49 (0.01)	0.20 (0.04)	0.17 (0.02)	0.34 (0.04)	0.26 (0.05)
ACC (300)	0.58 (0.07)	0.49 (0.03)	**0.64** (0.04)	0.52 (0.04)	0.52 (0.06)	0.58 (0.06)	0.57 (0.07)
NMI (300)	0.47 (0.06)	0.24 (0.03)	**0.52** (0.02)	0.36 (0.07)	0.27 (0.06)	0.35 (0.05)	0.42 (0.05)
ACC AUC	166.31 (2.43)	140.72 (1.13)	172.78 (2.37)	147.77 (2.32)	147.50 (3.14)	132.76 (3.81)	**178.57** (9.43)
NMI AUC	134.79 (2.55)	68.32 (1.13)	**145.89** (1.39)	93.72 (3.83)	70.60 (2.45)	71.81 (2.63)	127.09 (9.68)
COR AUC	81.70	70.70	76.43	68.02	72.33	**45.96**	52.14
CPU time	**0.02** (0.00)	0.63 (0.02)	1.05 (0.01)	368 (21)	2,636 (93)	42 162.29 (8 721.86)	23.74 (4.03)
				"Koren" ($n=43, p=980$)			
ACC (10)	0.48 (0.09)	0.68 (0.05)	0.82 (0.10)	0.80 (0.08)	**0.94** (0.09)	0.58 (0.06)	0.84 (0.17)
NMI (10)	0.13 (0.12)	0.39 (0.06)	0.62 (0.12)	0.61 (0.11)	**0.90** (0.11)	0.33 (0.08)	0.71 (0.10)
ACC (300)	0.74 (0.18)	0.80 (0.13)	0.77 (0.16)	0.87 (0.18)	0.87 (0.15)	0.86 (0.17)	**0.89** (0.02)
NMI (300)	0.53 (0.27)	0.61 (0.19)	0.66 (0.17)	0.78 (0.21)	0.76 (0.22)	0.78 (0.21)	**0.80** (0.05)
ACC AUC	172.90 (5.65)	225.25 (6.64)	233.94 (6.71)	263.04 (4.40)	**263.48** (5.61)	239.76 (8.96)	242.39 (8.71)
NMI AUC	88.29 (8.40)	163.35 (9.46)	186.58 (7.32)	**236.38** (6.87)	234.37 (6.38)	207.48 (11.43)	216.29 (12.18)
COR AUC	48.18	52.34	49.94	48.48	48.69	**32.60**	47.77
CPU time	**0.01** (0.00)	0.07 (0.01)	1.11 (0.12)	5.88 (0.26)	9.70 (0.36)	1 650.46 (224.47)	10.69 (0.03)

High efficiency of our approach to select relevant features in a reasonable computational time with no *a priori*.

Results UKFS



Comparison of the different approaches on “Carcinom”.

Results KOKFS

Evaluation on **multiple output regression problems** (i.e. $\mathcal{Y} = \mathbb{R}^q$).

Datasets:

- **Nutrimouse**: expression of $p = 120$ genes and the concentration of $q = 21$ hepatic fatty acids for $n = 40$ mice.
- **Diogenes**: gene expression (RNA-Seq) on human adipose tissue at two different time steps of a dietary intervention for $n = 167$ individuals. ($p = q = 269$).
- **TCGA**: primary tumor samples of breast cancer for which mRNA ($q = 9884$) and miRNA ($p = 655$) expressions are both available for $n = 1194$ individuals.

Results KOKFS

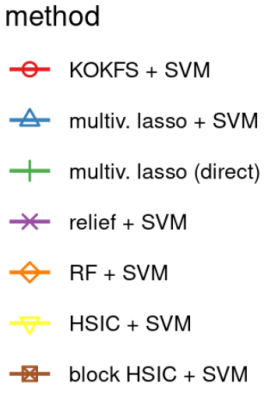
We used Gaussian input and output kernels

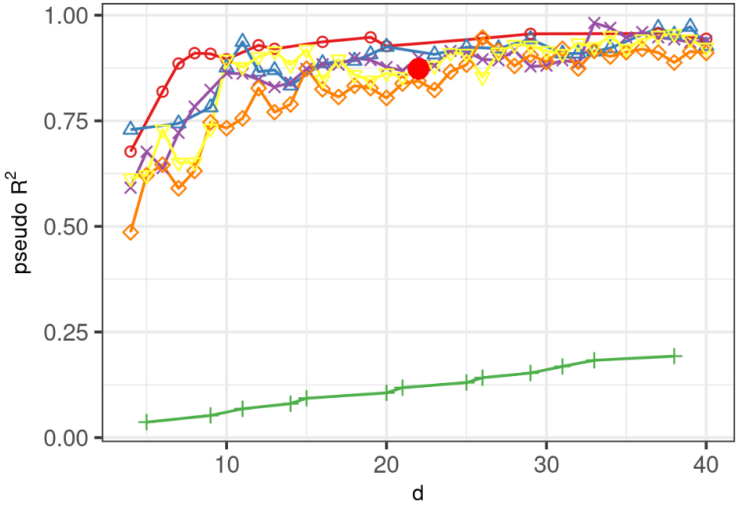
Experiments were conducted with two steps:

1. feature selection
2. performance assessment with nonparametric regressions based on the selected features

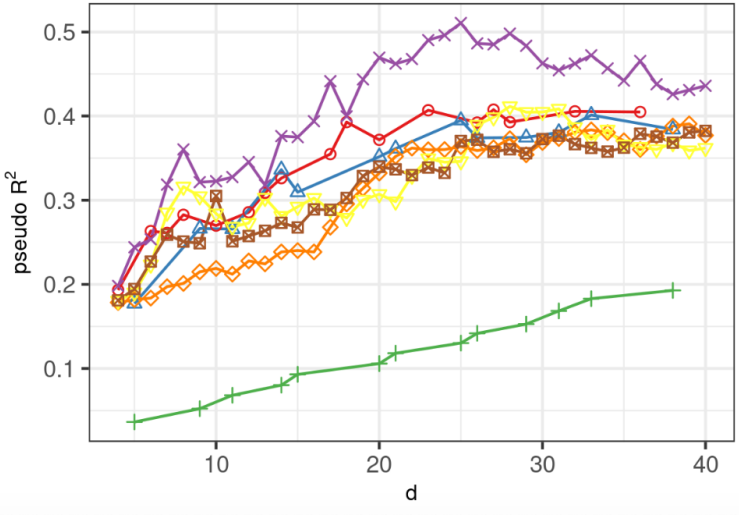
$$\text{pseudo-}R^2(y_{\cdot j}) = 1 - \frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}{Var(y_{\cdot j})}$$

Results KOKFS

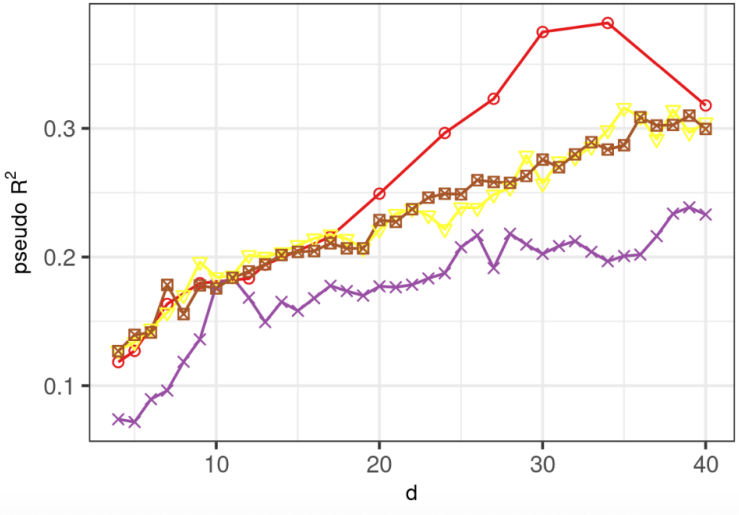




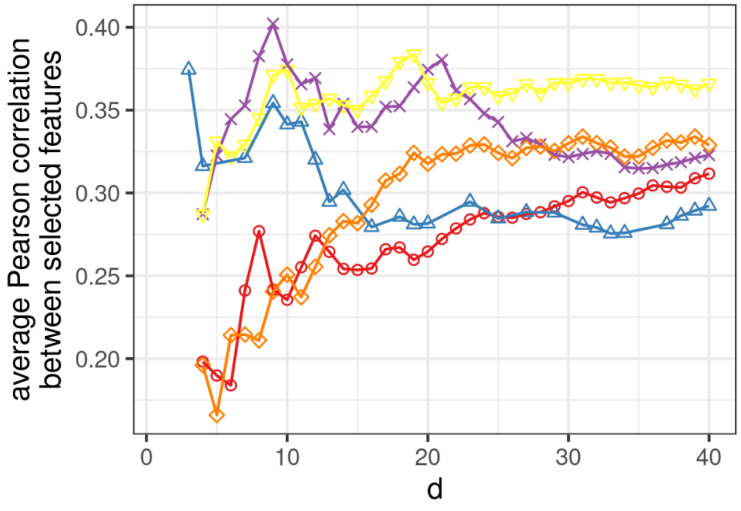
"Diogenes"



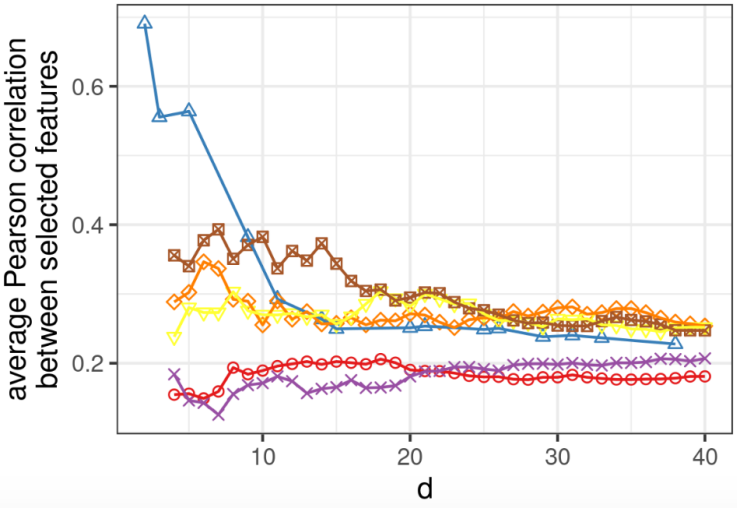
"TCGA"



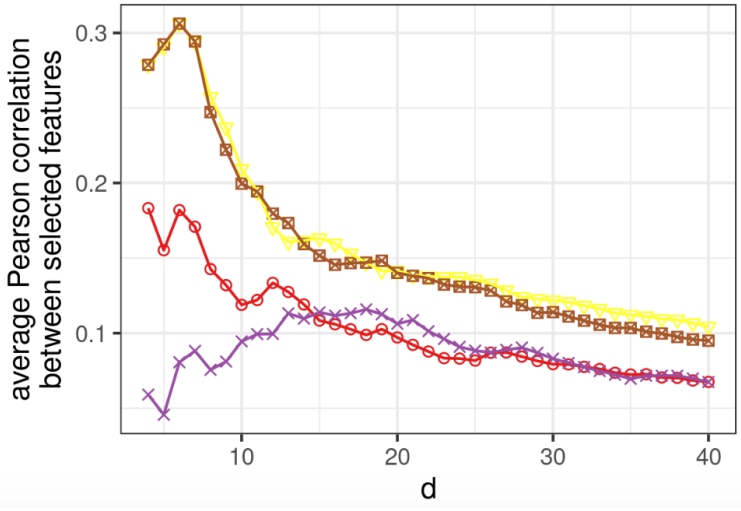
Results KOKFS: information about redundancy



"Diogenes"



"TCGA"



Conclusion & Perspectives

What did we do?

- Propose a feature selection algorithm that explicitly takes advantage of the kernel structure
- address two rarely met purposes: unsupervised learning and multiple output or non numerical output predictions;

Perspectives

- Integration of KOKFS in the **mixKernel** R package;
- Evaluation of KOKFS on a non numerical output prediction problem (Covid19 dataset)

Any questions ?

INRAE

MIA TOULOUSE

Genotoul Bioinfo

Université Nice Sophia Antipolis