Machine Learning with Functional Data

Dimensionality reduction

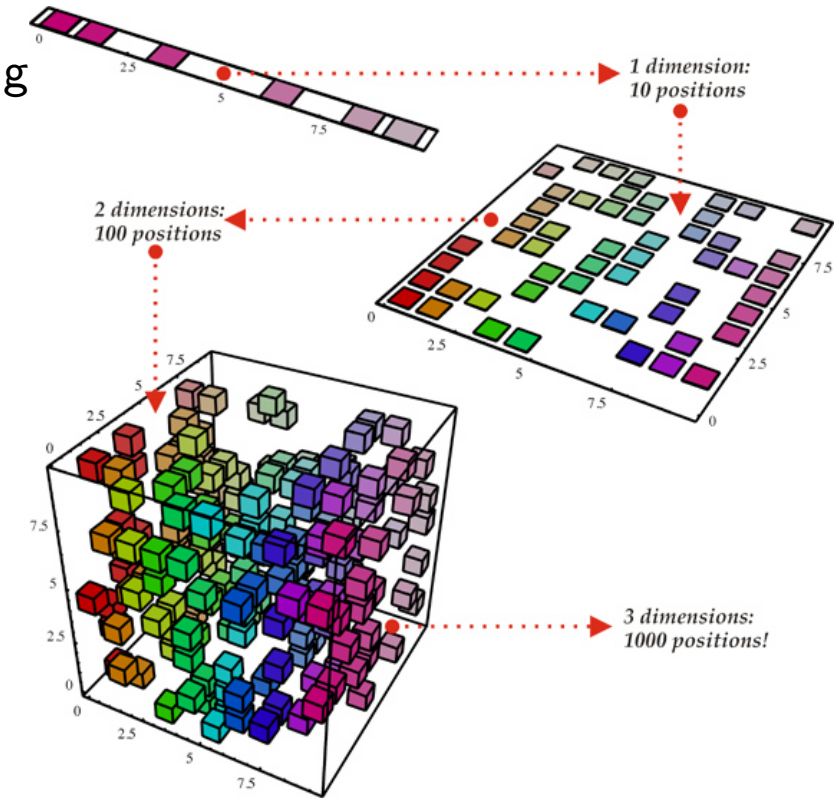
Variable generation

Sometimes, before reducing dimensionality, we expand it even more:

- Combinations of the original variables: sums, products, logical operations...
- Transformations of the original variables: logarithms, exponentials...
- Expert knowledge and exogenous variables.
- Curves: maxima, minima, mean value, derivatives...
- Images: %Red, %Green, %Blue, mean intensity...
- ...

High dimensionality issues

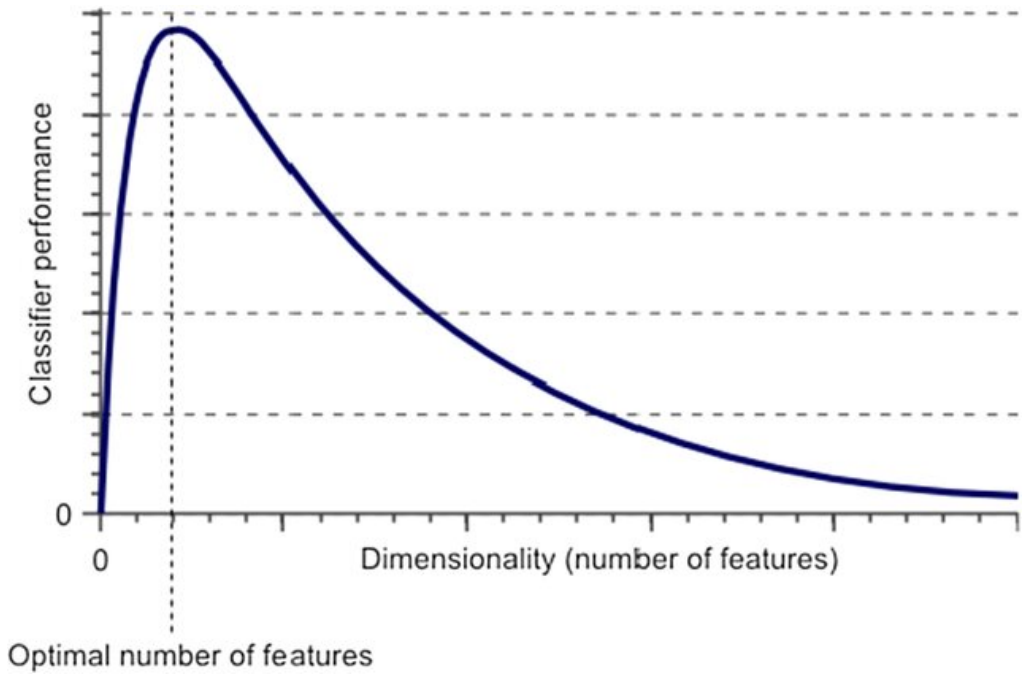
- Increased complexity in modeling and interpretation.
- Sparsity of data, leading to overfitting and poor generalization.
- Increased computation time and memory usage.
- Curse of dimensionality: increased volume of the space, leading to a decrease in the density of the data points and making it difficult to detect patterns or structure.
- Difficulty in visualization and exploration of the data.
- Risk of noise and redundancy in the data.



How do we deal with high dimensionality?

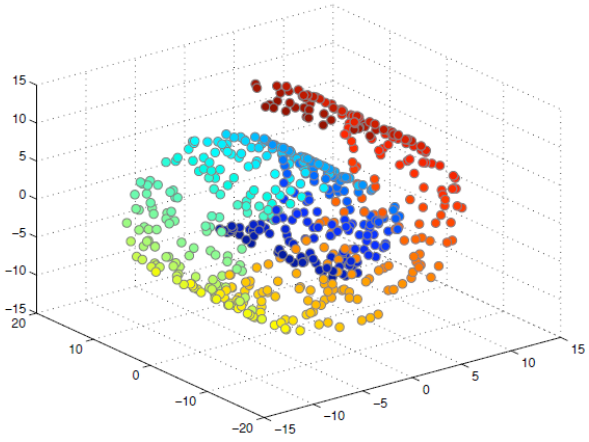
- Incorporating prior knowledge.
- Adding constraints.
- Reducing the dimension of the data.

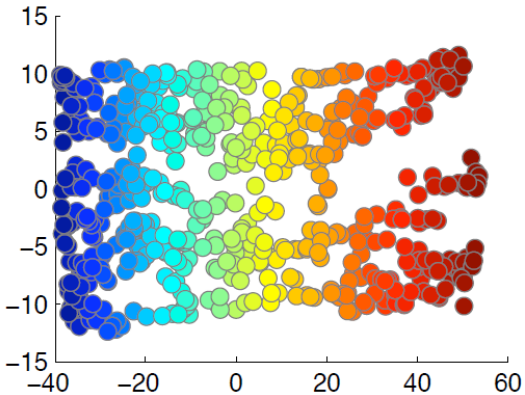
In practice, the efficiency of a classifier deteriorates beyond a certain number of variables (with a fixed number of data points). In many cases, the information loss associated with feature elimination is compensated by the better performance of the classifier in lower dimensions.



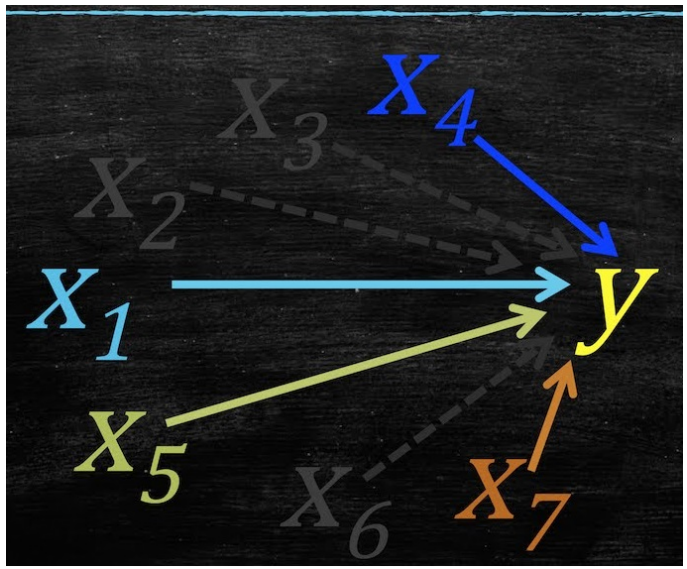
Approaches

Projection-based methods





Variable selection



Projection-Based Methods

Idea

Data can be represented in a lower-dimensional space without losing information.

Objective

To find the coordinate expression of that space to project the data.

Types

- Linear-Nonlinear: Linear methods assume that the structure of the data is in a linear subspace and apply linear transformations. Nonlinear methods do not assume this hypothesis and study nonlinear relationships.
- Global-Local: Global systems preserve the general characteristics in the lower dimension. Local systems do the same with the local characteristics.

Principal Component Analysis

Let X be a random element taking values in the sample space $\mathcal{X} = L^2[0,1]$. By analogy with the finite-dimensional case, the aim of *PCA* is to define orthonormal projection directions $\alpha_1, \ldots, \alpha_k \in L^2[0,1]$ such that the projections of X along these directions take as much variability as possible.
Thus the first principal component is given by the projection direction α_1 achieving maximum variance,

$$V(\langle \alpha_1, X \rangle) = \max\{V(\langle a, X \rangle) : \|a\| = 1\}$$

and, for $k > 1$, the k-th principal component is defined by

$$V(\langle \alpha_k, X \rangle) = \max\{V(\langle a, X \rangle) : \|a\| = 1, \langle a, a_j \rangle = 0, \text{ for } j = 1, \ldots, k-1\}$$

Then, the essential idea would be to replace in the statistical treatment the original data X_i with the corresponding k dimensional vector of projections $(\langle \alpha_1, X_i \rangle, \ldots, \langle \alpha_k, X_i \rangle)$. As in the finite-dimensional case, it can be shown that the PCA directions α_j turn out to be an **orthonormal basis of eigenvectors of the covariance operator associated with the kernel function** $\gamma(s,t) = \mathrm{Cov}(X(s), X(t))$. Also, the corresponding eigenvalues λ_j fulfill $\lambda_j = V(\langle \alpha_j, X \rangle)$.

Ramsay and Silverman (2005)

Empirical version: some smoothing must be done to exclude very rough solutions. Two possibilities are

(a) To smooth the data $X_i(t)$ and to use the "smoothed empirical" $\tilde{\gamma}_n$ associated with the smoothed data. For example the smoothing process could be done by convolution: we could define $X_{ih} = \int_0^1 K_h(t-s)X_i(s)ds$, K_h being a kernel, e.g., the Gaussian density $N(0, h^2)$. Then, the covariance operator would be estimated by

$$\tilde{\gamma}_n(s,t) := \gamma_{nh}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left((X_{ih}(s) - \bar{X}_h(s))(X_{ih}(s) - \bar{X}_h(t))\right).$$

(b) To solve a modified version of the optimization problem (7) aimed to penalize the "rough" solutions:

$$\frac{V(\langle \alpha, X\rangle)}{\|\alpha\|^2 + \delta\|\alpha''\|^2},$$

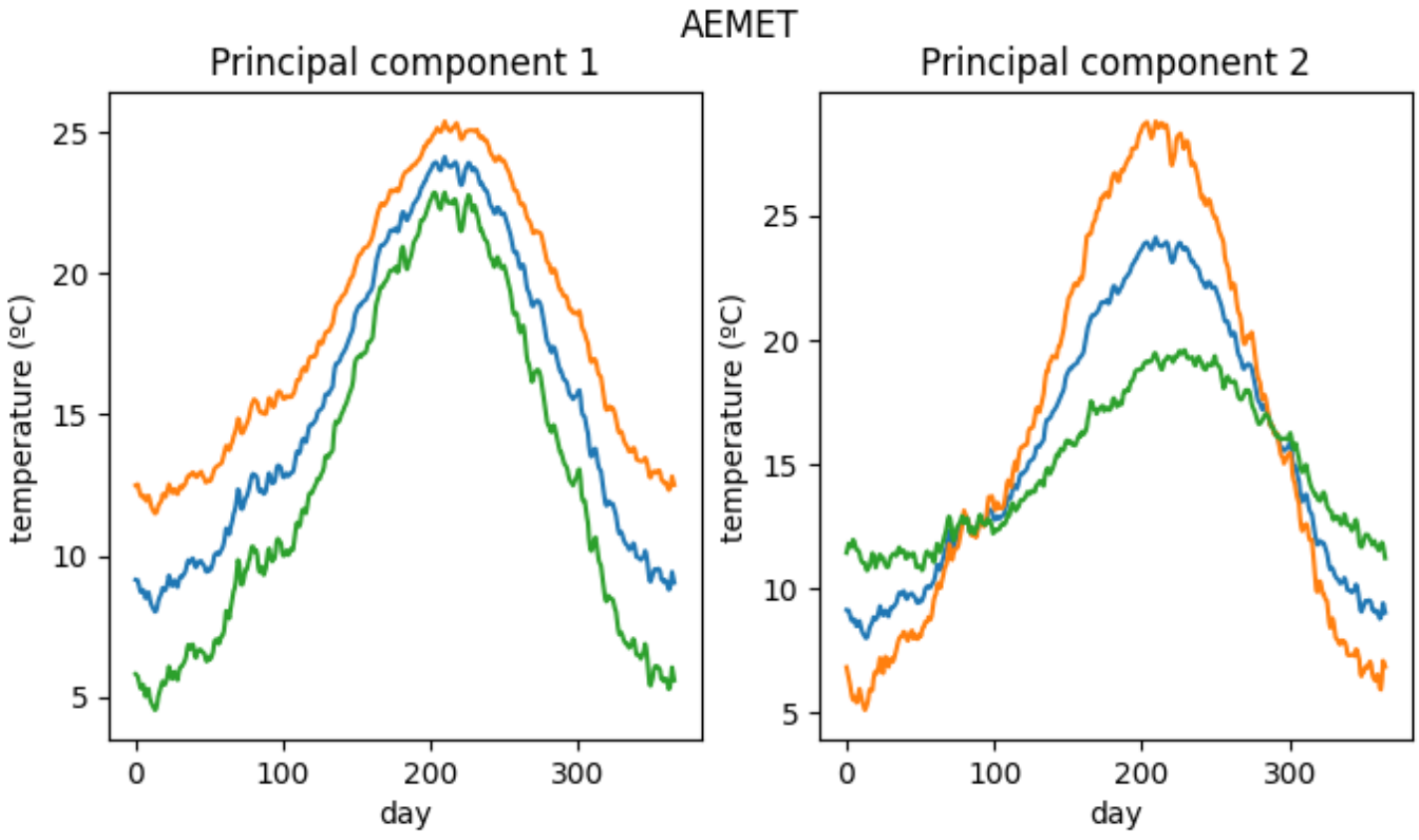
$\delta > 0$ being a roughness penalty.

Some remarks

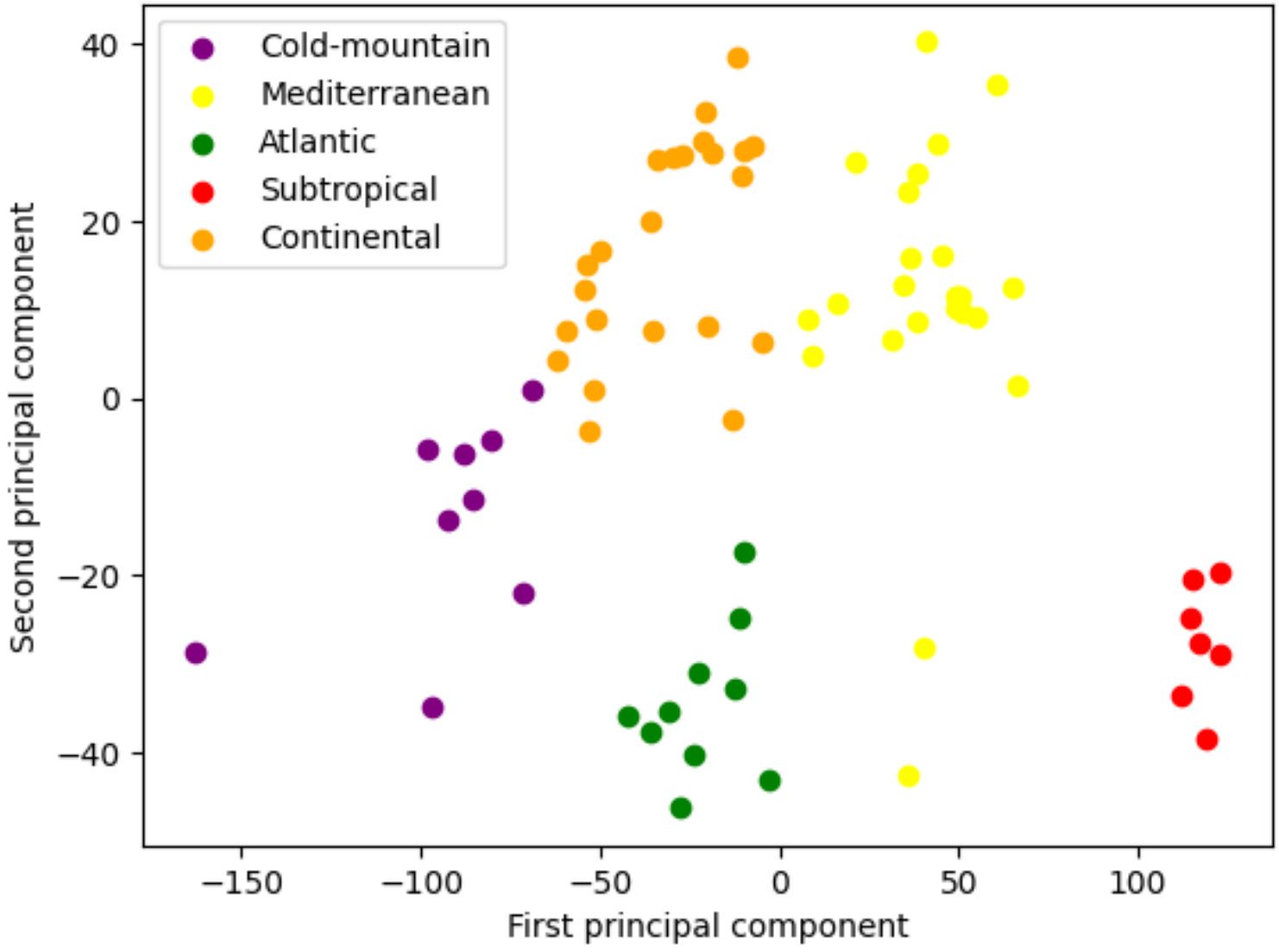
- Linear with orthogonal projections.
- Global.
- Assumes that relevant information is in the variability.
- Tries to explain as much variance as possible through projections.
- Good method of representation - Orthonormal empirical basis from data..
- May not be a good alternative if there are response variables.
- Outlier detection (carefully).

PCA: AEMET example

PCA: AEMET example



PCA: AEMET example



scikit-fda: FPCA

FPCA

```
class skfda.preprocessing.dim_reduction.FPCA(n_components=None, *,
centering=True, regularization=None, components_basis=None, _weights=None)
```

Principal component analysis. [source]

Class that implements functional principal component analysis for both basis and grid representations of the data. The parameters are shared when fitting a FDataBasis or FDataGrid, except for `components_basis`.

Parameters:

- **n_components** –

  Number of principal components to keep from functional principal component analysis.
- **centering** – Set to `False` when the functional data is already known to be centered and there is no need to center it. Otherwise, the mean of the functional data object is calculated and the data centered before fitting . Defaults to `True`.
- **regularization** – Regularization object to be applied.
- **components_basis** – The basis in which we want the principal components. We can use a different basis than the basis contained in the passed FDataBasis object. This parameter is only used when fitting a FDataBasis.

Partial least squares (PLS)

The idea behind the partial least squares method is the same as PCA: we want to select projection directions that retain the most information. The difference is that PLS takes into account the response variable and aims to preserve the covariance relationship between the variables and the class.

- Heuristic.
- Linear with orthogonal projections.
- Global.
- PLS projections maximize $Cov^2(X, Y)$.
- Preferable to PCA when there is a response variable.
- Two versions: as a solution to an eigenvalue problem and as an iterative algorithm.

Problem of eigenvalues

$$(a_{k+1}, b_{k+1}) = \underset{a \in \mathbb{R}^D, b \in \mathbb{R}^M, a^t A = 0}{argmax} \frac{Cov(a^t X, b^t Y)^2}{(a^t a)(b^t b)}$$

Where a_k is the k-th component corresponding to the k-th eigenvalue, and A is the matrix of the already chosen $k-1$ components. The matrix to be diagonalized would be:

$$H = \Sigma_{XY} \Sigma_{YX}$$

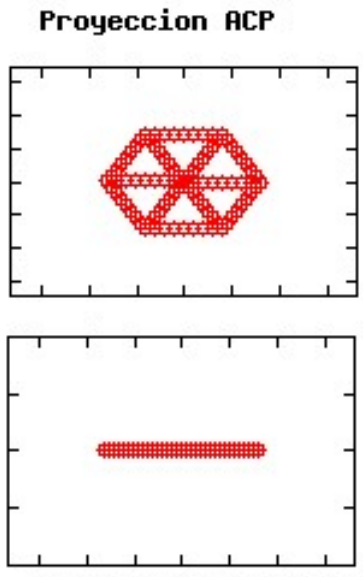
Preda et al. (2007), Delaigle and Hall (2012)

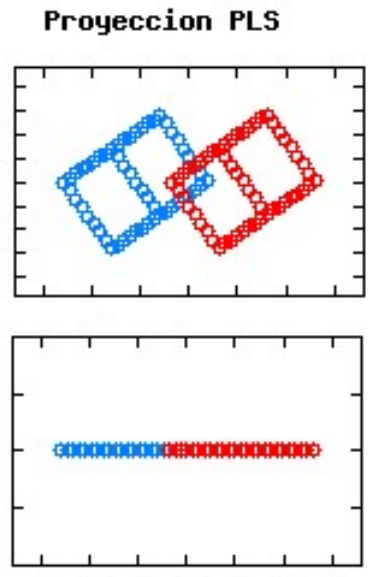
Iterative algorithm

1. Initialization: standardized $Y(0) = Y$ and $X(0) = X$.
2. For k=1 to d:
3. $w(k) = Cov[Y(k-1), X(k-1)]$
4. $w(k) = \frac{w(k)}{\|w(k)\|}$
5. $T_k = X(k-1)w(k)$
6. $v(k) = (\frac{T_k^t Y(k-1)}{T_k^t T_k})$; $b(k) = (\frac{T_k^t X(k-1)}{T_k^t T_k})$
7. $Y(k) = Y(k-1) - T_k v(k)$; $X(k) = X(k-1) - T_k b(k)$
8. If $k = 1$ $z_1 = w(1)$ else $z_k = [Id - \sum_{j=1}^{k-1} z_j b(j)]w(k)$

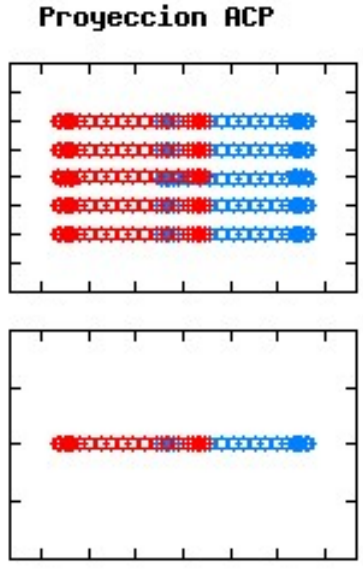
PLS vs PCA

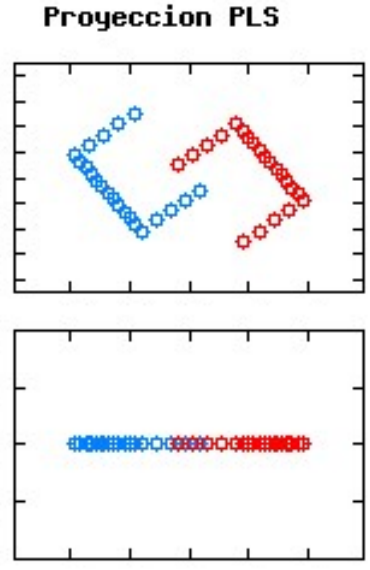

Estado inicial


Proyeccion ACP


Proyeccion PLS


Estado inicial

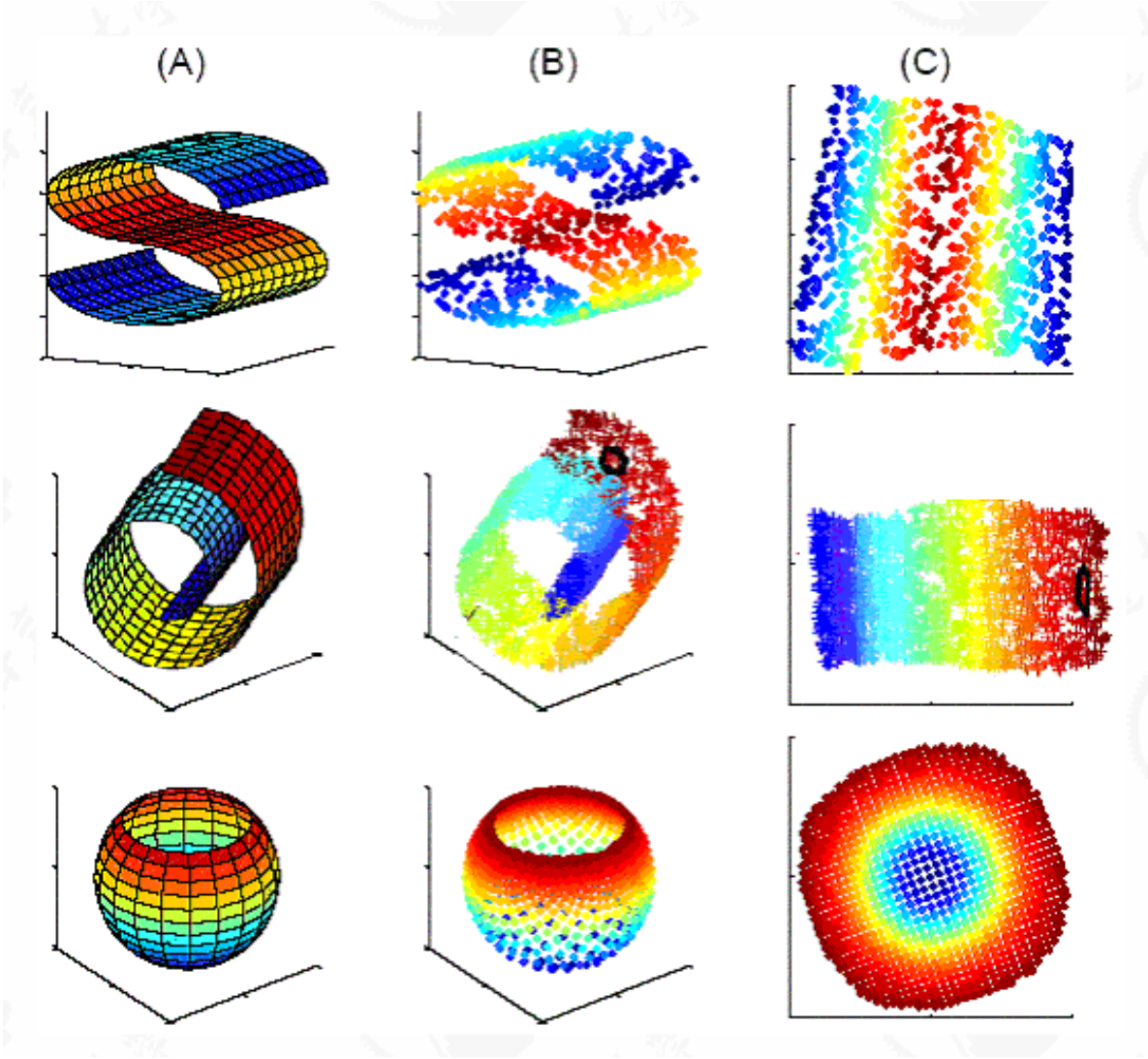

Proyeccion ACP


Proyeccion PLS

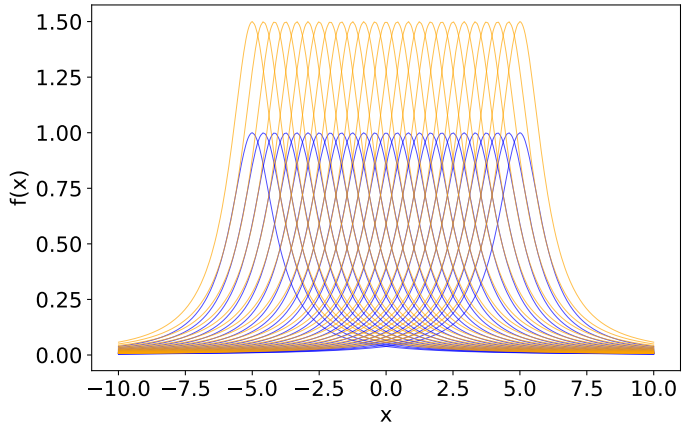
Porjection-based methods

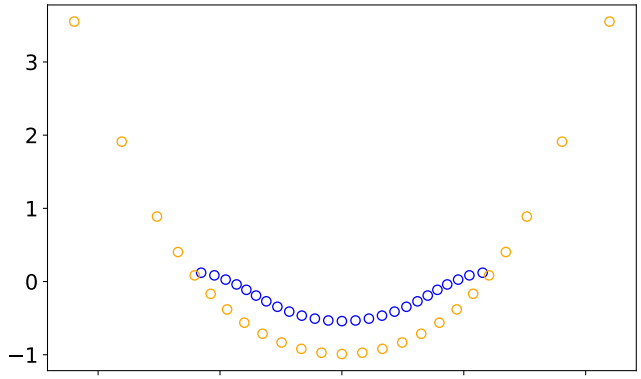
- Kernel PCA
- Factor analysis (principal factor, maximum likelihood...)
- Generalized linear discriminant analysis (GLDA)
- Projection pursuit
- Independent component analysis (ICA)
- Random projections
- Multidimensional scaling
- Singular value decomposition
- Diffusion maps
- Principal curves and submanifolds
- Isomap
- Laplacian eigenmaps
- ...

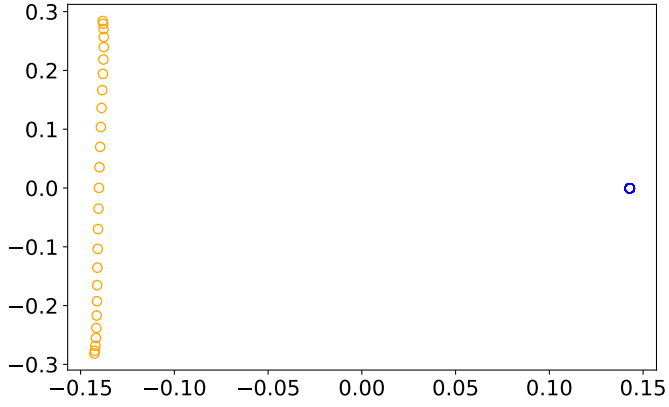
Non-linear methods

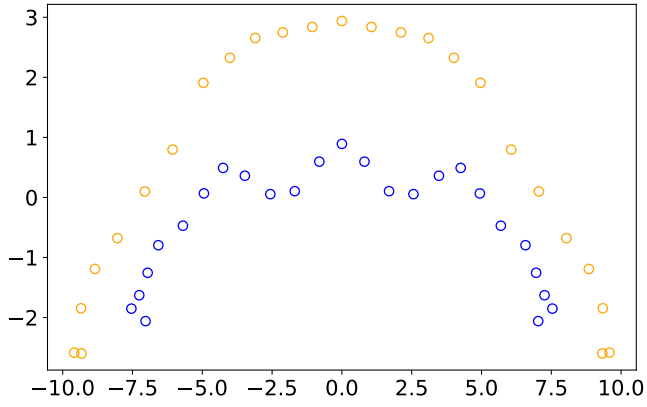


Non-linear methods





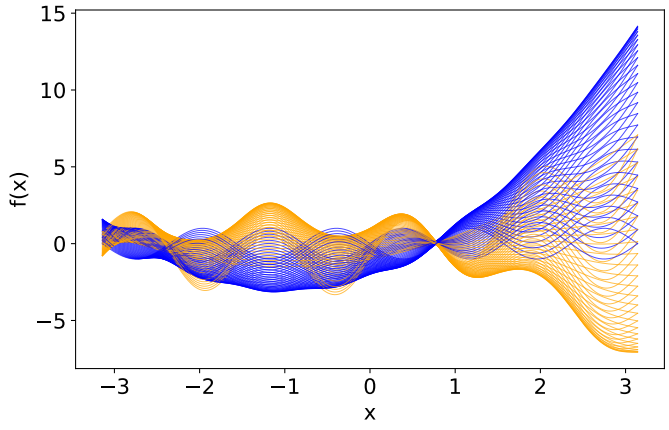


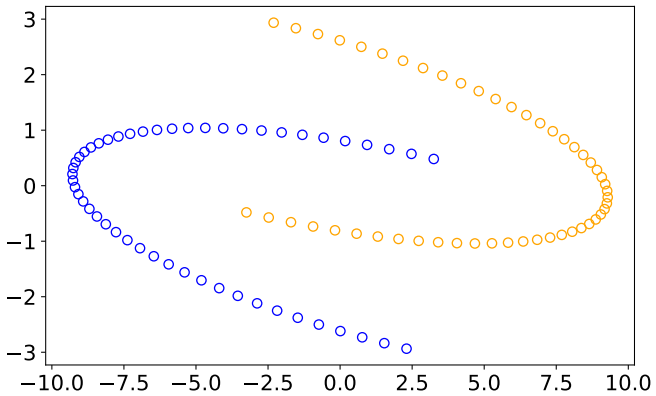


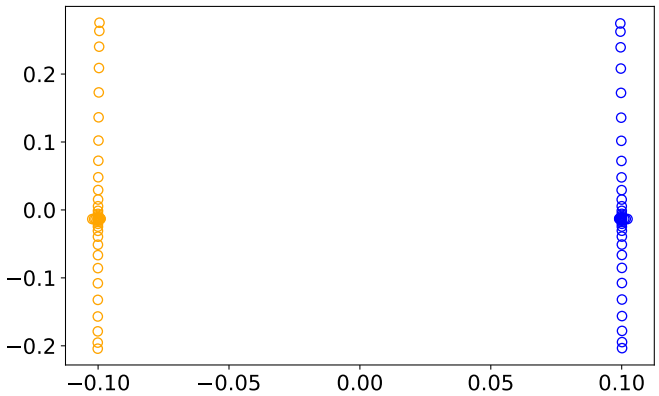
a) Cauchy densities b) FPCA c) FDM ISOMAP

Barroso, et al. (2023). Functional Diffusion Maps. Axiv

Non-linear methods







a) 'Functional' moon data b) FPCA c) FDM

Variable selection

Idea: Select the most informative subset of variables among the original variables of the problem.

Variable selection

Given a set of n data of dimension m $(X_1, \cdots, X_m)$, $t \in [0, 1]$, the goal of variable selection is to replace each observation x by a vector $(x_1, \cdots, x_d)$ with certain well-chosen variables $(d << m)$.

Once the selection is made, we can apply any classifier to the reduced data.

Objectives

1. Select a small subset of variables that can be used for classification. Non-redundant variables with high discriminant power
2. Identify the relevant variables for further research. All variables with significant information regardless of redundancy

"probably a more challenging and relevant issue [than improving prediction] is to identify sets of genes with biological relevance" Díaz-Uriarte and Álvarez de Andrés (2006).

Variable Selection

Motivation

- Variable selection is an effective dimension reduction technique in many fields.
- The reduction is made in terms of the original variables, which results in an increase in interpretability.
- It is related to the way experts proceed.

Purpose

- Eliminates redundant or irrelevant variables.
- Reduces storage and computation costs.
- Improves classifier performance and reduces overfitting risk.
- More interpretable models.

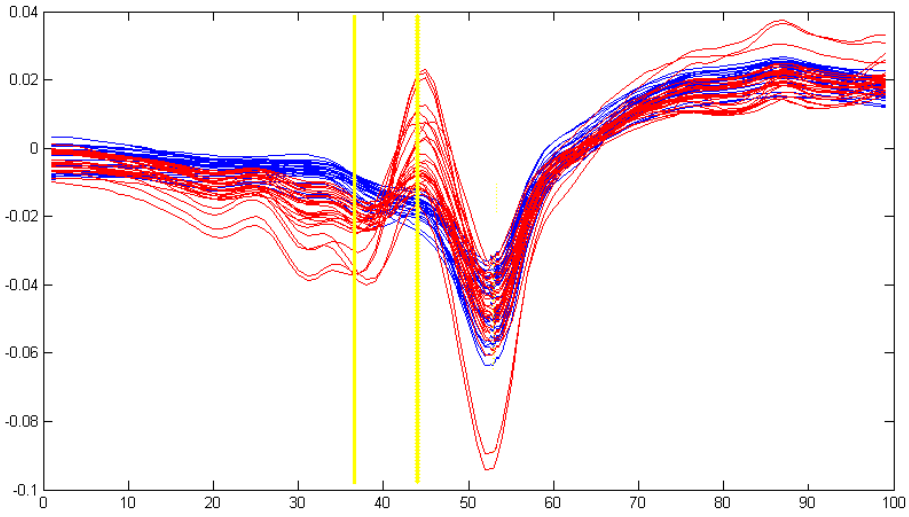
What do we mean by "variable selection" in FDA?

- **Idea** Choose the most informative subset among the original variables.
- Given a sample of functions $X_1(t), \cdots, X_n(t)$, $t \in [0,1]$ our aim is to replace every sample function X_j with a vector

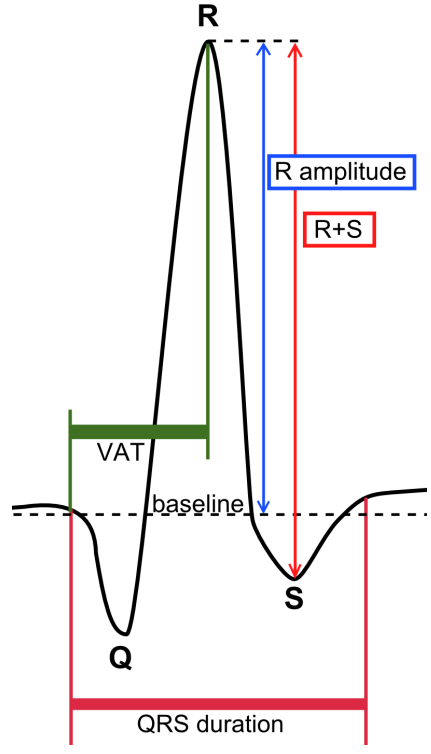
$$(X_j(t_1), \cdots, X_j(t_d)),$$

  for suitably chosen points $t_1, \cdots, t_d$.
- Then we would apply multivariate methods (regression, classification,...) to the "reduced" data.
- According to our experience, the value of d should be typically small (not much larger than 5, say).

Examples: Tecator y ECG



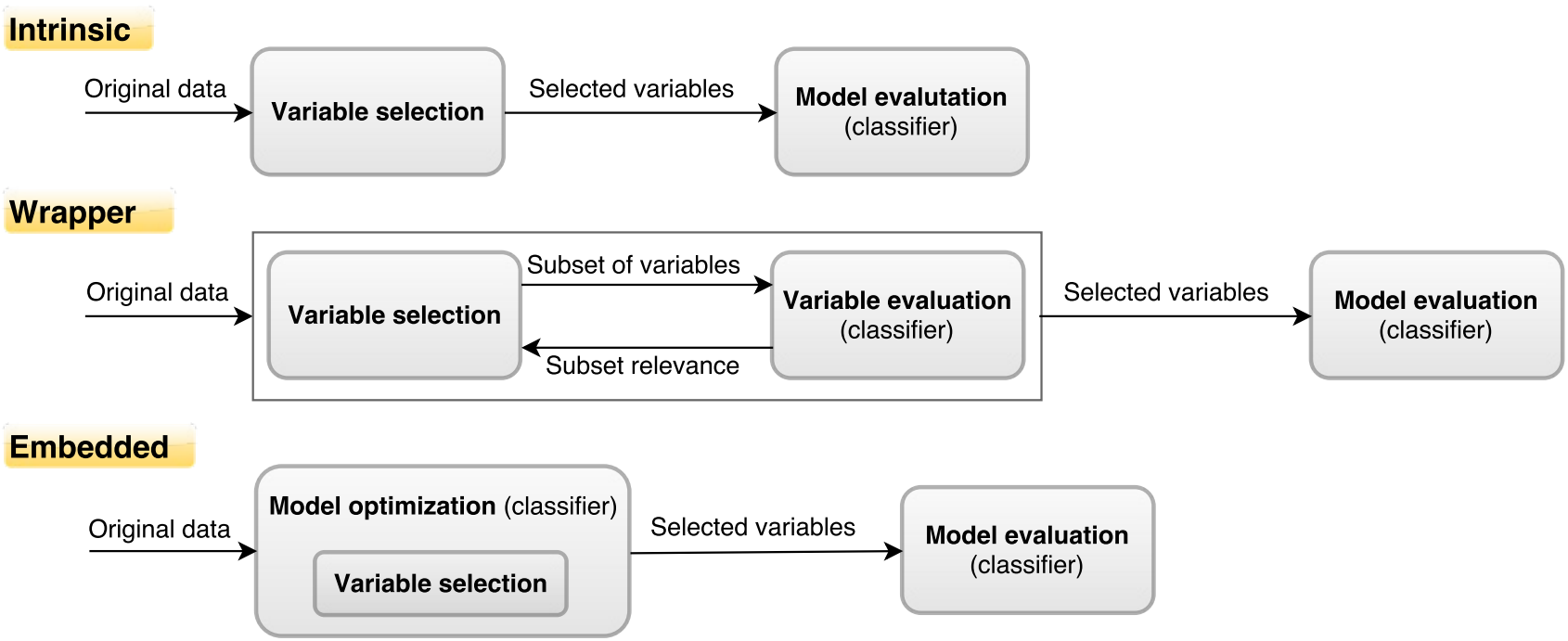
A doctor looks only a few variables of an ECG to diagnose.



The percentage of correct classification with two variables (97,23%) is even better than those (96,32% and 96,77%) obtained with the whole processes and with PLS using four components.

Relationship with the classifier

The first classification of variable selection methods is usually made according to their relationship with the learning model, whether they are independent (filter), dependent (wrapper), or inseparable (embedded).



Componentwise Delaigle et al. (2012)

It is a "wrapper" method, in the sense that it depends on the chosen classifier. Given a classifier, the method proposes a leave-one-out choice of the best variables for the considered classification problem. While this is a worthwhile natural idea, it is computationally intensive (even with the suggested computational savings).

Relation with the classifier

Filter

- Independent of the classifier, only depends on the data.
- Good generalization capability and some robustness.
- Computationally fast and efficient. Scalable.
- Ignore possible interactions with the classifier.
- Correlation-based Feature Selection, Relief, minimum Redundancy Maximum Relevance...

Wrapper

- The selection process "wraps" around the classifier. The classifier functions as a black box to score the different selections.
- Computationally very expensive (double search).
- Dependent on the classifier and not very generalizable.
- Exploit connections with the classifier and achieve high levels of accuracy (risk of overfitting).
- Any classifier, with SVM being the most prominent in the literature.

Embedded

- The selection and estimation of the model are carried out at the same time and are inseparable.
- Intermediate computational cost.
- Totally dependent on the model.
- Total connection with the model.
- LASSO, Random forest, modified SVM's...

Hybrids or two-stage

Combinations of different types that try to take advantage of the advantages of each one. The most usual is to start with a filter method and then use a wrapper strategy on the first selection.

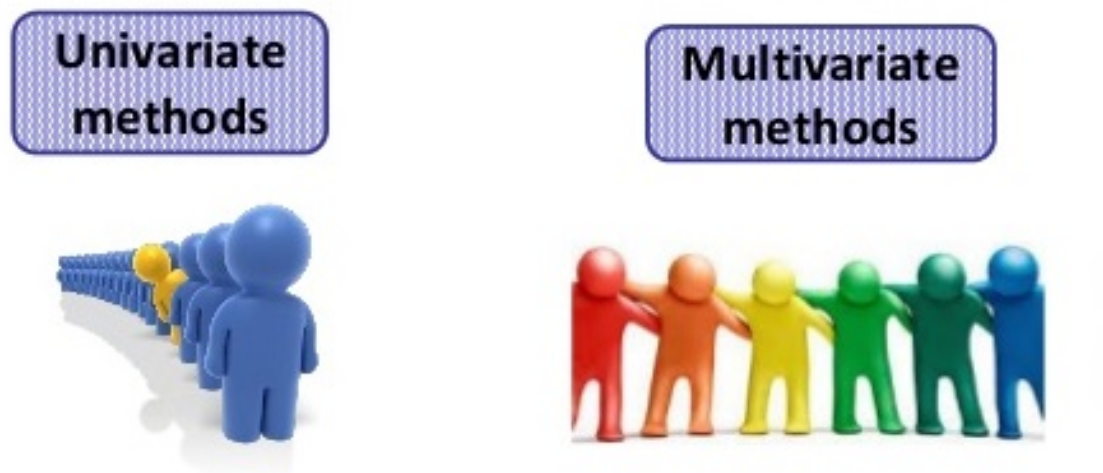
Relation between variables

Univariate

Also called ranking. Variables are evaluated individually and ordered by score.

Multivariate

Consider the interactions between variables to choose the best subset, both positive and negative (redundancy).

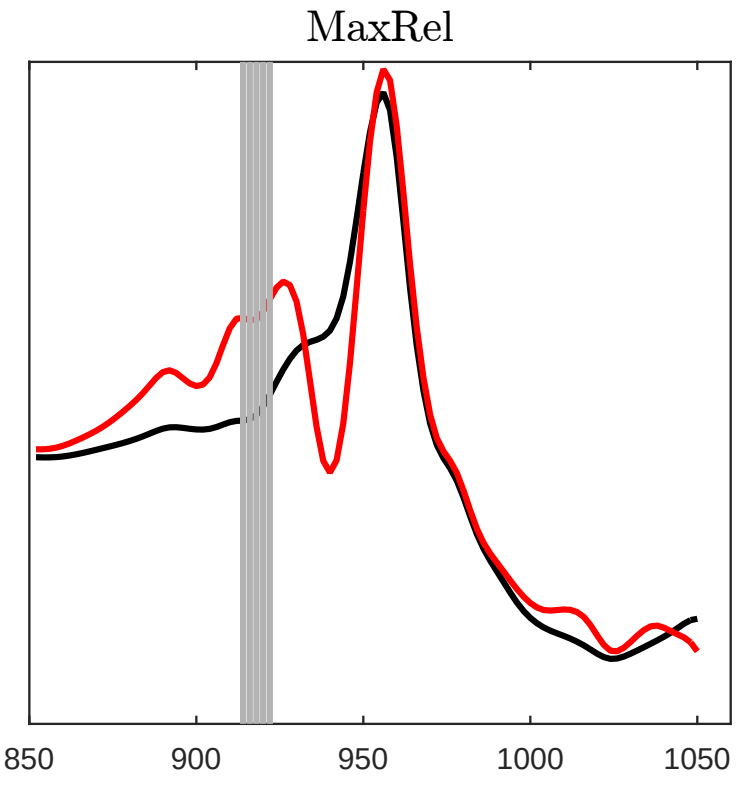


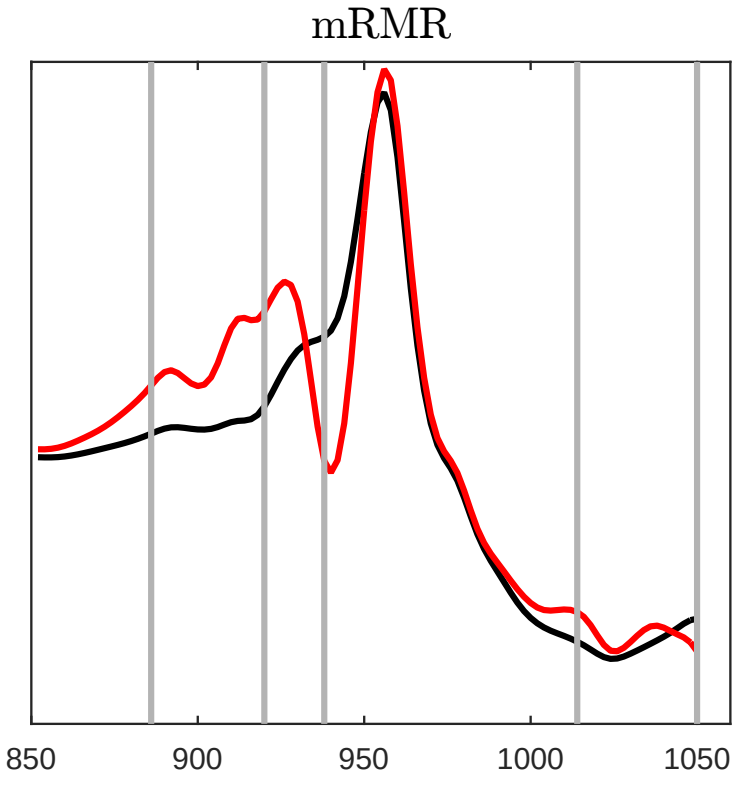
Search strategy

The search strategy defines the way in which we explore the space of all possible subsets of variables until some stopping criterion is met. This problem is NP-complete and at the end, we must resort to suboptimal and often heuristic strategies.

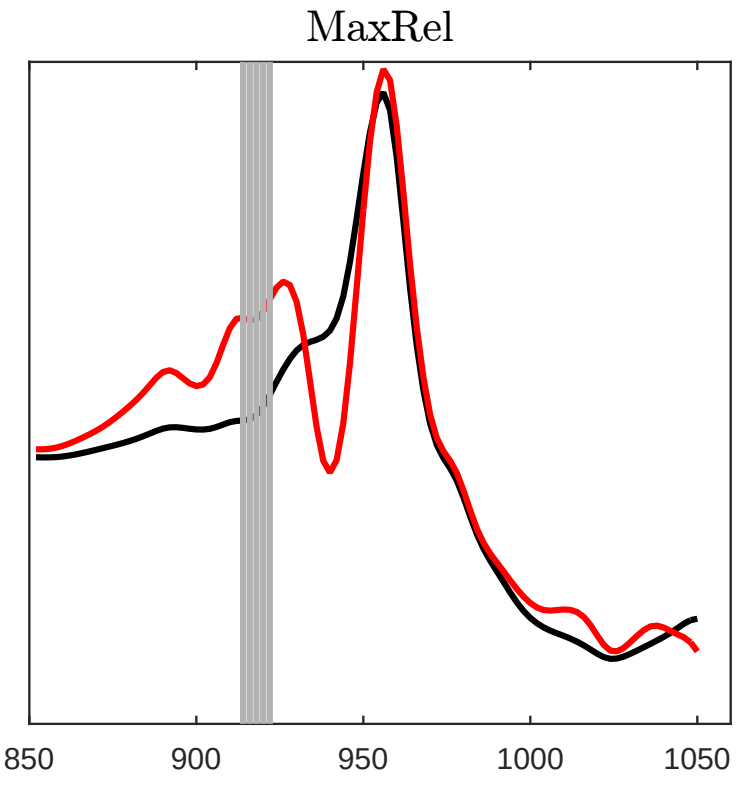
- Exhaustive or complete.
- Forward selection.
- Backward elimination.
- Random (genetic algorithms).
- Mixed.
- ...

Functional variable selection

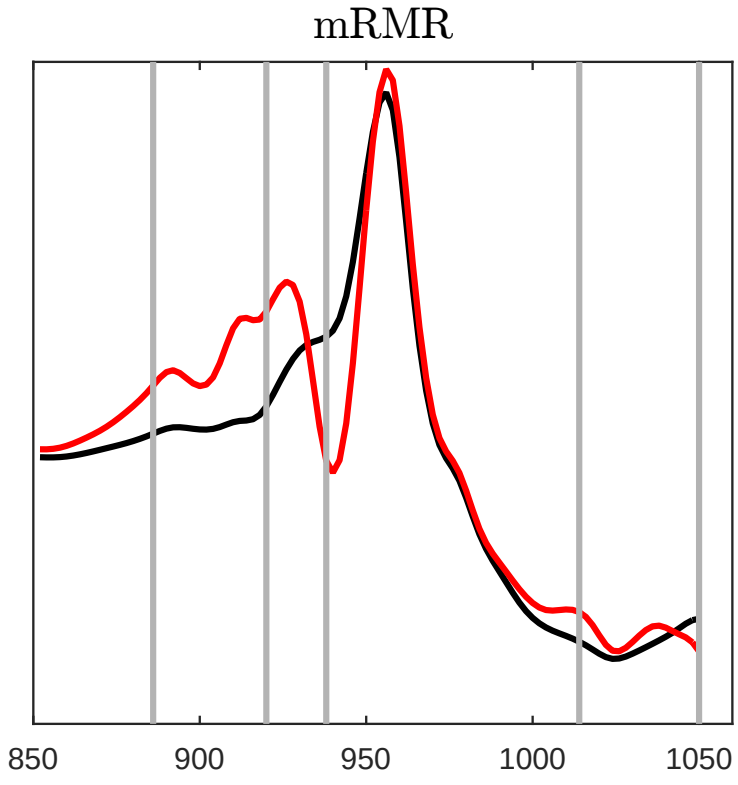




Functional variable selection



$err = 4.09\%$



$err = 1.86\%$

Relevance or quality measure

- The relevance measure is what decides whether a variable (or set of variables) is informative or not.
- There is no universal choice, it depends on the problem we are working on.
- In this context, our variables will be relevant if they are discriminant. Measures of dependence and association:
  - Gain ratio.
  - Gini index.
  - Covariance/correlation.
  - Relief.
  - Mutual information.
  - Correlation of distances.
  - ...

An example: mutual information

Mutual Information

A general measure of statistical independence between two random variables. It takes into account non-linear relationships.

Let X and Y be two random variables, with marginal probability densities $p(X)$ and $p(Y)$, and joint probability density function $p(X, Y)$. The mutual information between X and Y is defined as:

$$I(X, Y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy$$

- $I(X, Y) \geq 0$ and $I(X, Y) = 0$ if and only if X and Y are independent.
- $I(X, Y) = I(Y, X)$.

Estimating probability densities is a complex problem (especially in high dimensions). One option is to discretize the variables:

$$I(X, Y) = \sum_i \sum_j \mathbb{P}(x = i, y = j) \log \frac{\mathbb{P}(x = i, y = j)}{\mathbb{P}(x = i)\mathbb{P}(y = j)}$$

Distance correlation $\mathcal{R}$

- Distance correlation is a measure of dependence between two random vectors proposed by Székely, Rizzo and Bakirov, *Ann Stat* (2007) and Székely, Rizzo, (2009, 2012, 2013).
- For any distribution with finite first moments, $\mathcal{R}$ generalizes standard correlation in two ways:
  - $\mathcal{R}(X, Y)$ is defined for X and Y of arbitrary dimensions.
  - $\mathcal{R}(X, Y) = 0$ characterizes the independence of X and Y.
- It has an easy-to-calculate empirical estimator that does not require parameter tuning or smoothing (available in the energy R package).

Formulation of $\mathcal{R}$

The distance covariance $\mathcal{V}$ measures the distance between the joint characteristic function and the product of the marginals as follows:

$$\mathcal{V}^2(X,Y) = \| \varphi_{X,Y}(u,v) - \varphi_X(u)\varphi_Y(v) \|_w^2 .$$

In $\mathbb{R}^p \times \mathbb{R}^q$, the norm is defined as

$$\mathcal{V}^2(X,Y) = \int_{\mathbb{R}^{p+q}} | \varphi_{X,Y}(u,v) - \varphi_X(u)\varphi_Y(v) |^2 \, w(u,v) du dv,$$

where $w(u,v)$ can be any weight function for which the above integral exists. In practice, $w(u,v) = (c_p c_q |u|_p^{1+p} |v|_q^{1+q})^{-1}$ is taken, thus,

$$\mathcal{V}^2(X,Y) = \int_{\mathbb{R}^{p+q}} \frac{| \varphi_{X,Y}(u,v) - \varphi_X(u)\varphi_Y(v) |^2}{c_p c_q |u|^{1+p} |v|^{1+q}} du dv.$$

And the distance correlation is defined as:

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}} & \text{if } \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0 & \text{if } \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases}$$

Empirical version of $\mathcal{R}$

Given a random sample $(X_k, Y_k) : k = 1, \cdots, n$ of (X, Y) with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$,

$$\mathcal{V}^2(X, Y)_n = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl},$$

where $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} - \bar{a}$ and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} - \bar{b}$ with $a_{kl} =\mid X_k - X_l \mid_p$ and $b_{kl} =\mid Y_k - Y_l \mid_q$.

Now, we can define the empirical version of the distance correlation as:

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}} & \text{if } \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0 \\ 0 & \text{if } \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0 \end{cases}$$

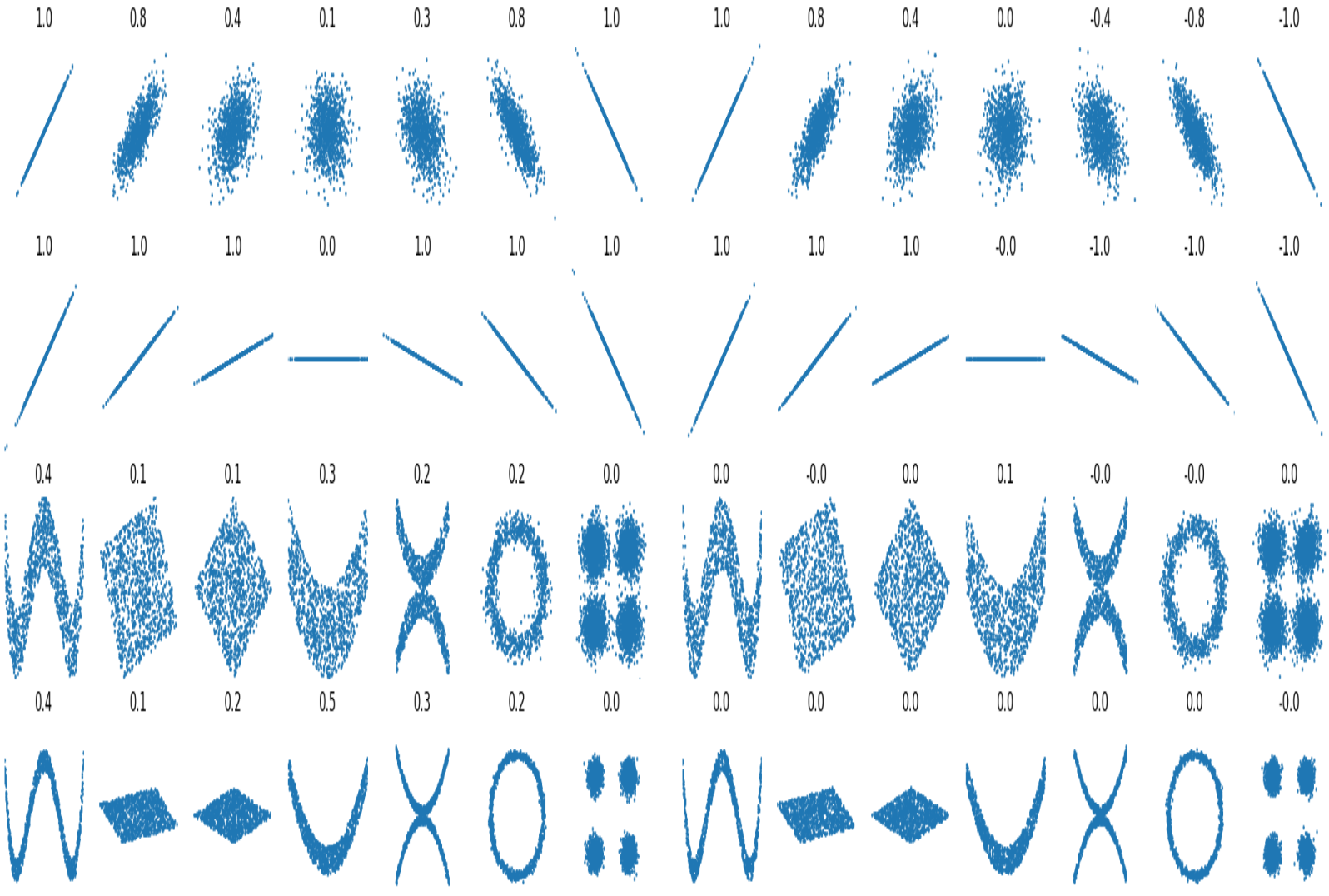
The convergence of the estimator

$$\mathcal{V}_n^2(X, Y) =\| \varphi_{X,Y}^n(u, v) - \varphi_X^n(u)\varphi_Y^n(v) \|_w^2$$

is proven, and $\lim_{n\to\infty} \mathcal{V}_n^2 = \mathcal{V}^2$ almost surely.

Ramos-Carreño, C., & Torrecilla, J. L. (2023). dcor: Distance correlation and energy statistics in Python. SoftwareX, 22, 101326.

DCor vs Cor


1.0 0.8 0.4 0.1 0.3 0.8 1.0 1.0 0.8 0.4 0.0 -0.4 -0.8 -1.0
1.0 1.0 1.0 0.0 1.0 1.0 1.0 1.0 1.0 1.0 -0.0 -1.0 -1.0 -1.0
0.4 0.1 0.1 0.3 0.2 0.2 0.0 0.0 -0.0 0.0 0.1 -0.0 -0.0 0.0
0.4 0.1 0.2 0.5 0.3 0.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -0.0

Minimum Redundancy Maximum Relevance method

mRMR is a popular feature selection procedure. It was proposed by Ding and Peng (2005), Peng et al. (2005) .

- Filter-based
- Univariate
- Forward
- Utilizes Mutual Information

The Algorithm

- Relevance measure: $I(\cdot, \cdot)$
- $Rel(X_i) = I(X_i, Y)$
- $Red(X_i, X_j) = I(X_i, X_j)$

The Algorithm

- Relevance measure: $I(\cdot, \cdot)$
- $Rel(X_i) = I(X_i, Y)$
- $Red(X_i, X_j) = I(X_i, X_j)$

Let $S = 1, \ldots, d$ be a subset of variables:

- $Rel(S) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i, Y)$
- $Red(S) = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i, X_j)$

The Algorithm

- Relevance measure: $I(\cdot, \cdot)$
- $Rel(X_i) = I(X_i, Y)$
- $Red(X_i, X_j) = I(X_i, X_j)$

Let $S = 1, \ldots, d$ be a subset of variables:

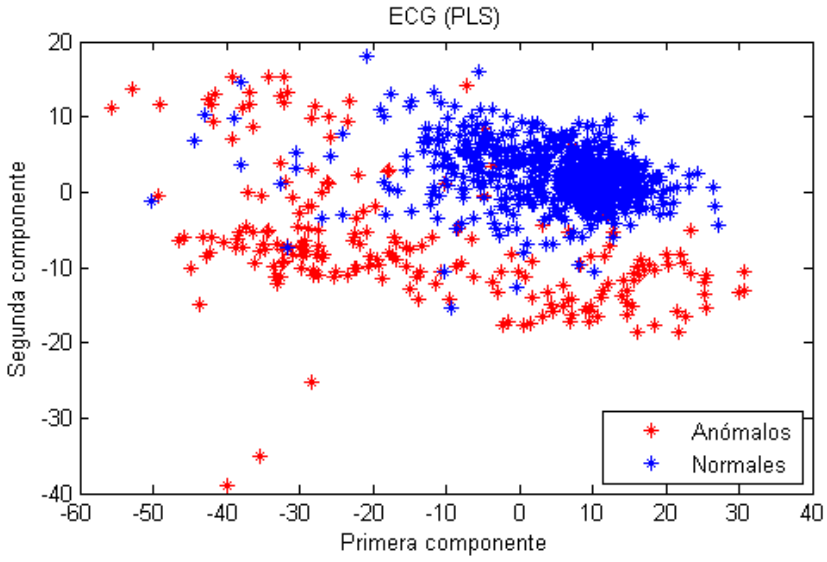
- $Rel(S) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i, Y)$
- $Red(S) = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i, X_j)$

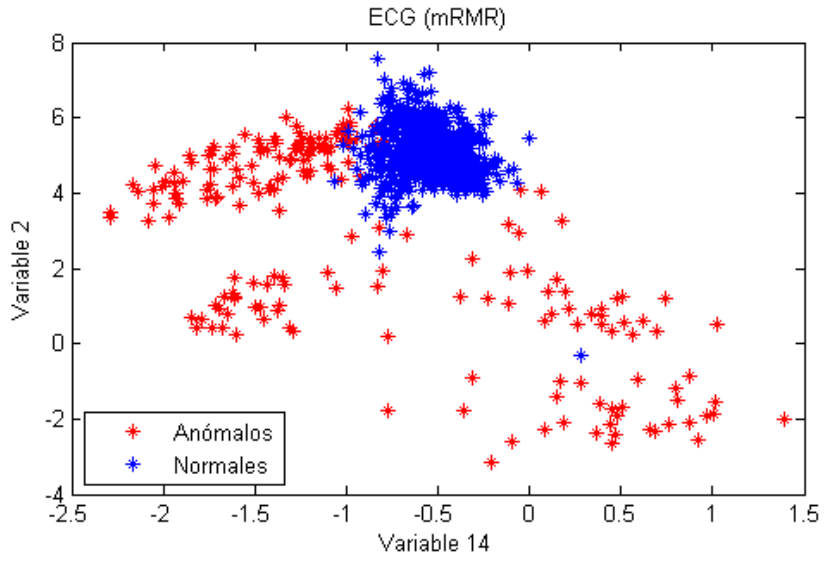
The goal is to select the subset of variables S that maximizes

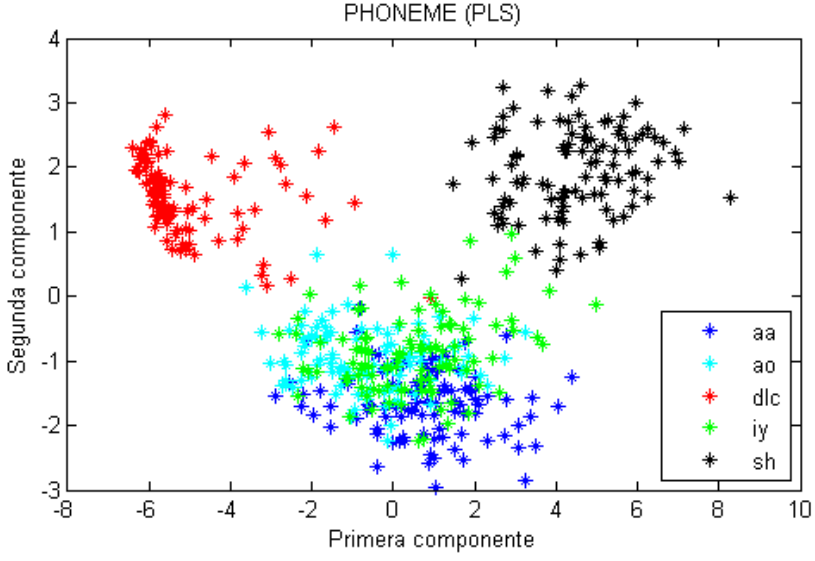
- MID: $Rel(S) - Red(S)$
- MIQ: $Rel(S)/Red(S)$

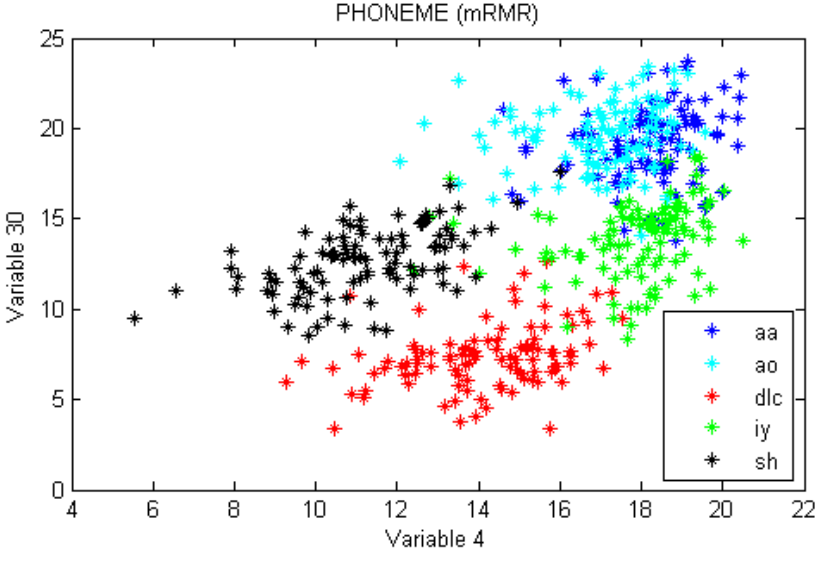
The search is done greedily, iteratively adding the variable that maximizes one of the criteria.

mRMR vs PLS


ECG (PLS)
Segunda componente
Primera componente
Anómalos
Normales

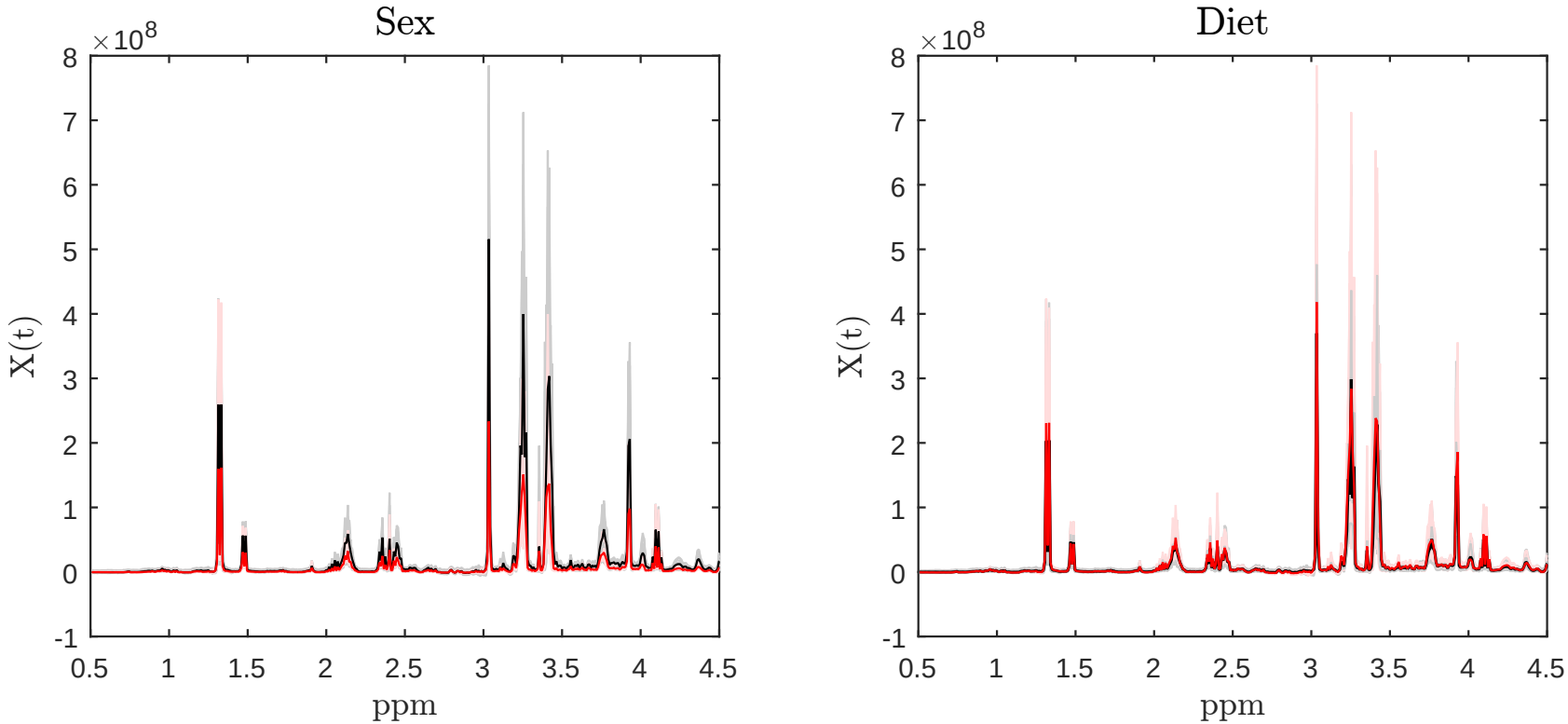

ECG (mRMR)
Variable 2
Variable 14
Anómalos
Normales


PHONEME (PLS)
Segunda componente
Primera componente
aa
ao
dlc
iy
sh


PHONEME (mRMR)
Variable 30
Variable 4
aa
ao
dlc
iy
sh

Application: NMR spectral fingerprints

The goal of the experiment was to investigate the possible relationships between metabolic phenotype and mitochondrial function with sex and the effects of a high-fat diet in mice.



Barba, *et al.* High fat diet and female sex induce metabolic changes and reduce oxidative stress in an additive maner in mice heart. Journal of Nutritional Biochemistry (2017)

Preliminary results

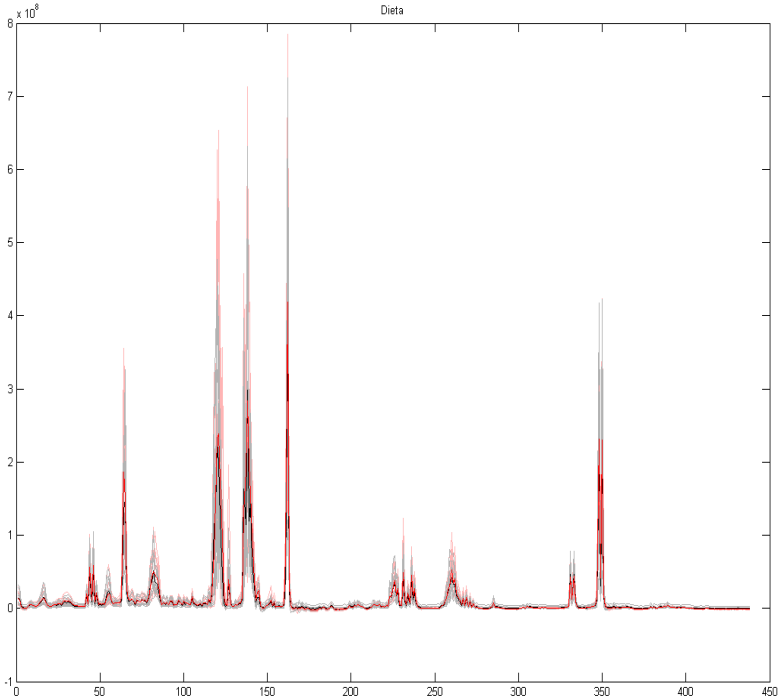
Sex (3-NN)

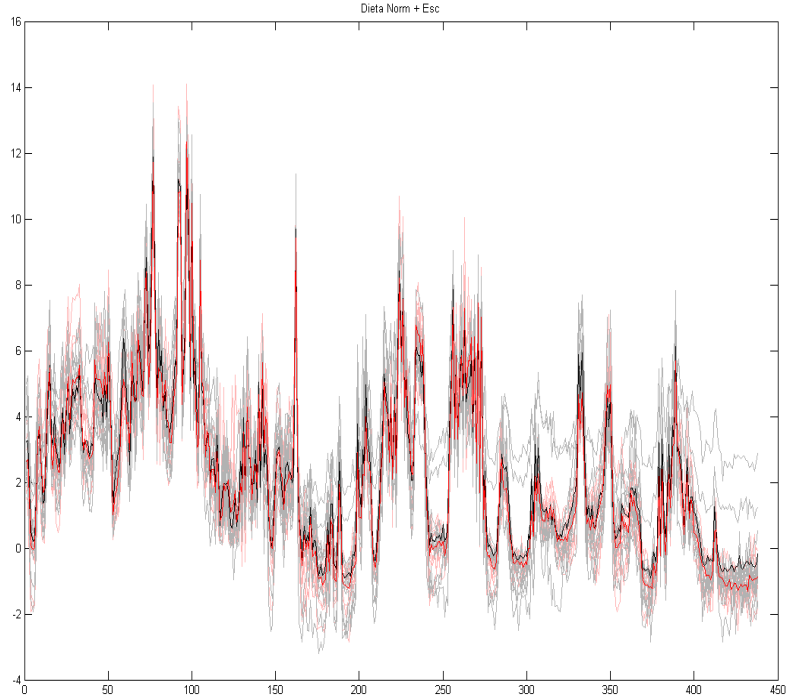
	$\hat{Male}$	$\hat{Female}$
Male	**9**	4
Female	1	**9**

Diet (3-NN)

	$\hat{HDF}$	$\hat{control}$
HDF	**5**	7
control	5	**6**

Preprocessing





- Normalization: area=1
- Standarization: mean=0, var=1
- Smoothing

Some results

Sex (raw data)

	$\hat{Male}$	$\hat{Female}$
Male	**9**	4
Female	1	**9**

Diet (raw data)

	$\hat{HDF}$	$\hat{control}$
HDF	**5**	7
control	5	**6**

Some results

Sex (raw data)

	$\hat{Male}$	$\hat{Female}$
Male	**9**	4
Female	1	**9**

Diet (raw data)

	$\hat{HDF}$	$\hat{control}$
HDF	**5**	7
control	5	**6**

Sex (norm+stand)

	$\hat{Male}$	$\hat{Female}$
Male	**11**	2
Female	6	**4**

Diet (norm+stand)

	$\hat{HDF}$	$\hat{control}$
HDF	**3**	9
control	2	**9**

Some results

Sex (3-NN)

	$\hat{Male}$	$\hat{Female}$
Male	**9**	4
Female	1	**9**

Diet (3-NN)

	$\hat{HDF}$	$\hat{control}$
HDF	**5**	7
control	5	**6**

Some results

Sex (3-NN)

	$\hat{Male}$	$\hat{Female}$
Male	**9**	4
Female	1	**9**

Diet (3-NN)

	$\hat{HDF}$	$\hat{control}$
HDF	**5**	7
control	5	**6**

Sex (mRMR+LDA)

	$\hat{Male}$	$\hat{Female}$
Male	**12**	1
Female	0	**10**

Diet (mRMR+LDA)

	$\hat{HDF}$	$\hat{control}$
HDF	**12**	0
control	2	**9**

First variables

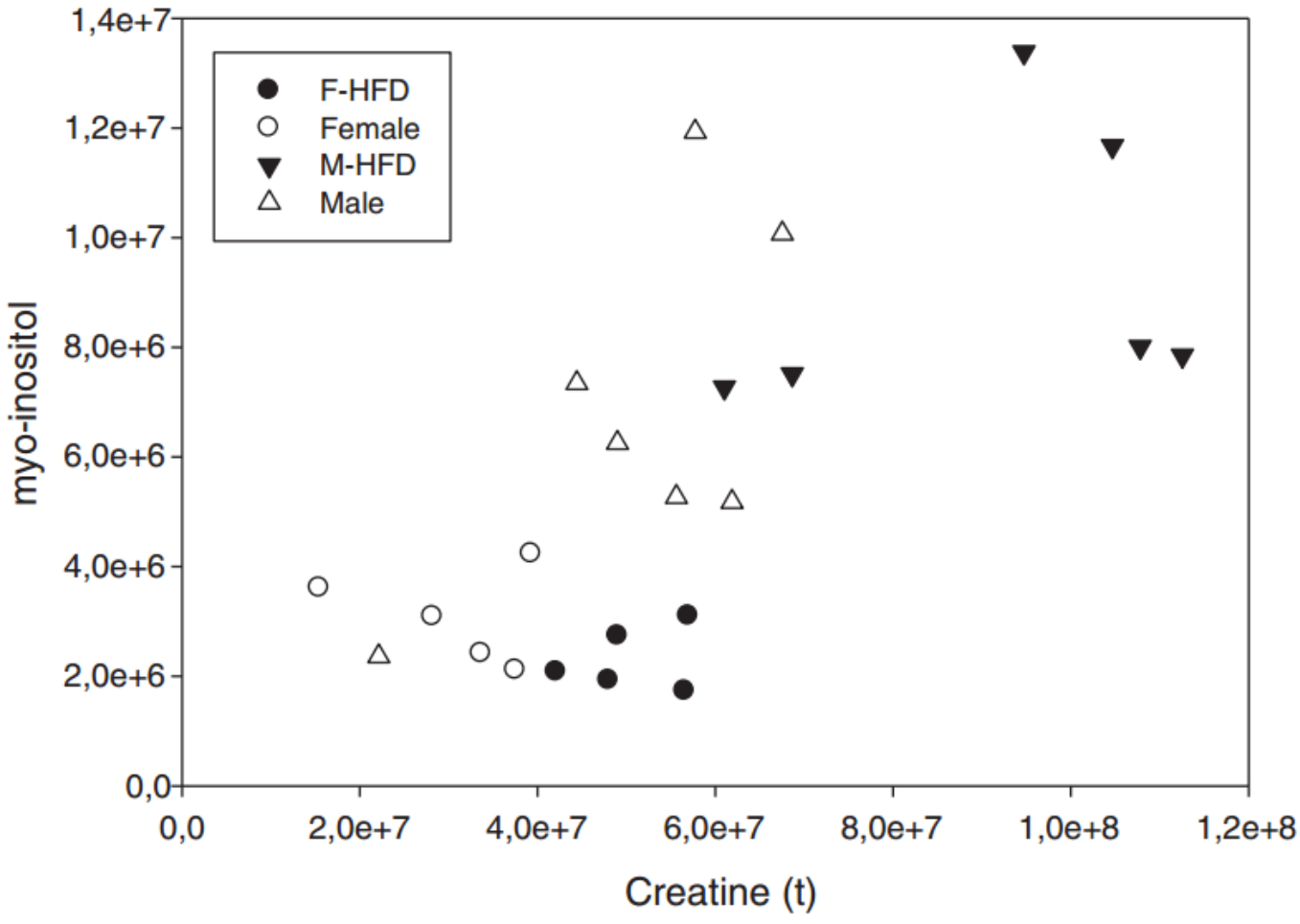


Fig. 2. Scatter plot showing the values of ^{1}H NMR peak height for creatine and myo-inositol (variables 161 and 54) chosen after mRMR + LDA analysis.

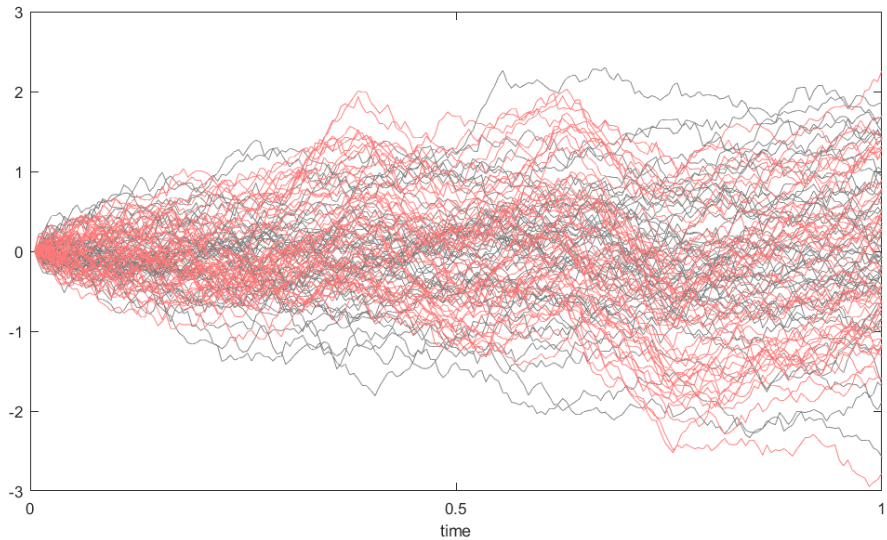
Maxima Hunting: Classification framework

We consider a homoskedastic model

$$\begin{cases} X(t)|Y=0: & Z(t) \\ X(t)|Y=1: & Z(t)+\mu(t) \end{cases}$$

- $Z(t)$ is a Gaussian processes with $\mathbb{E}[Z(t)]=0$ and covariance kernel $K(s,t)$.
- $\mu(t)=\mu_1(t)-\mu_0(t)$ is a deterministic function.

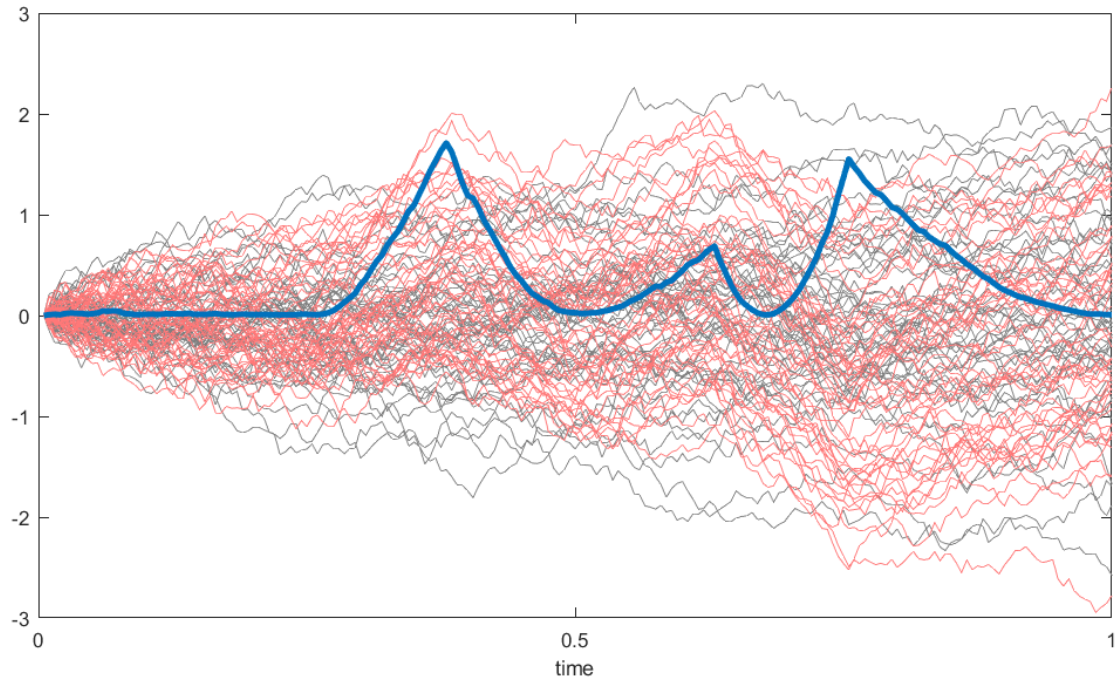
We consider a situation in which there is a clear-cut population target: the Optimal (Bayes) rule is of type $g^*(x)=g(x(t_1),\cdots,x(t_d))$ for some unknown $t_1,\cdots,t_d$. Berrendero et al. (2018)



Maxima Hunting criterion

- Choose your favourite non-negative measure of statistical dependence $I(\cdot, \cdot)$.
- Calculate the relevance function $R(t) = I(X(t), Y)$.
- Select the points $t_1, \cdots, t_d$ in according to the local maxima of the relevance function $R(X(t), Y)$.

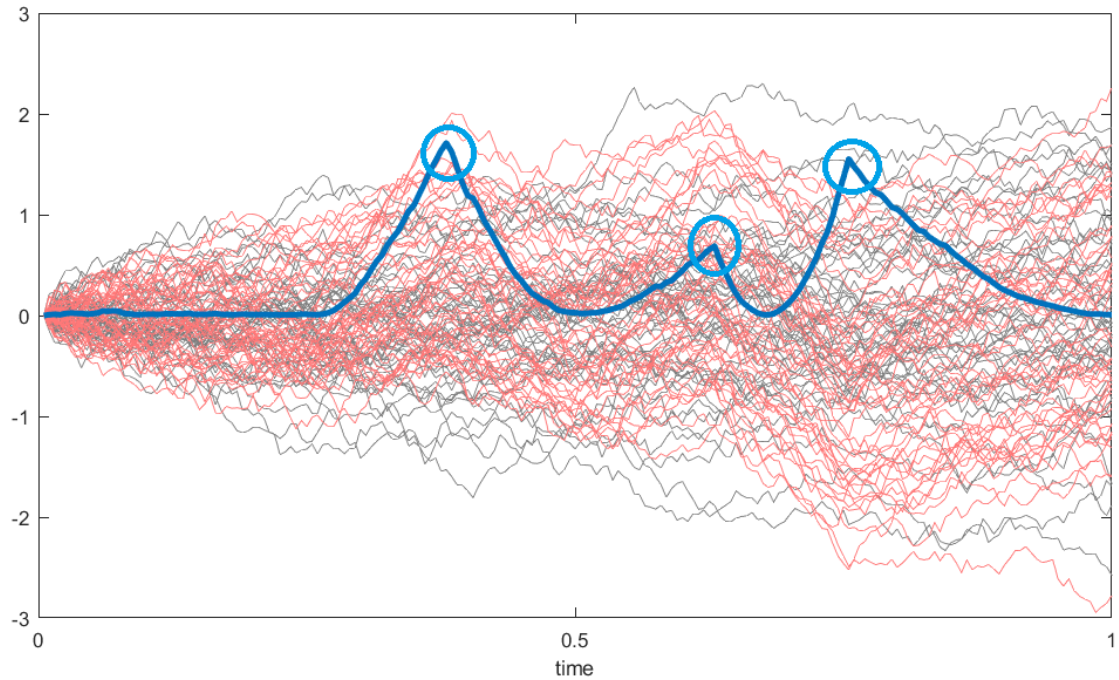
$$X(t_i) = \underset{t \in [0,1]}{argmax} \quad R(X(t), Y), \quad i = 1, 2, \ldots, d.$$



Maxima Hunting criterion

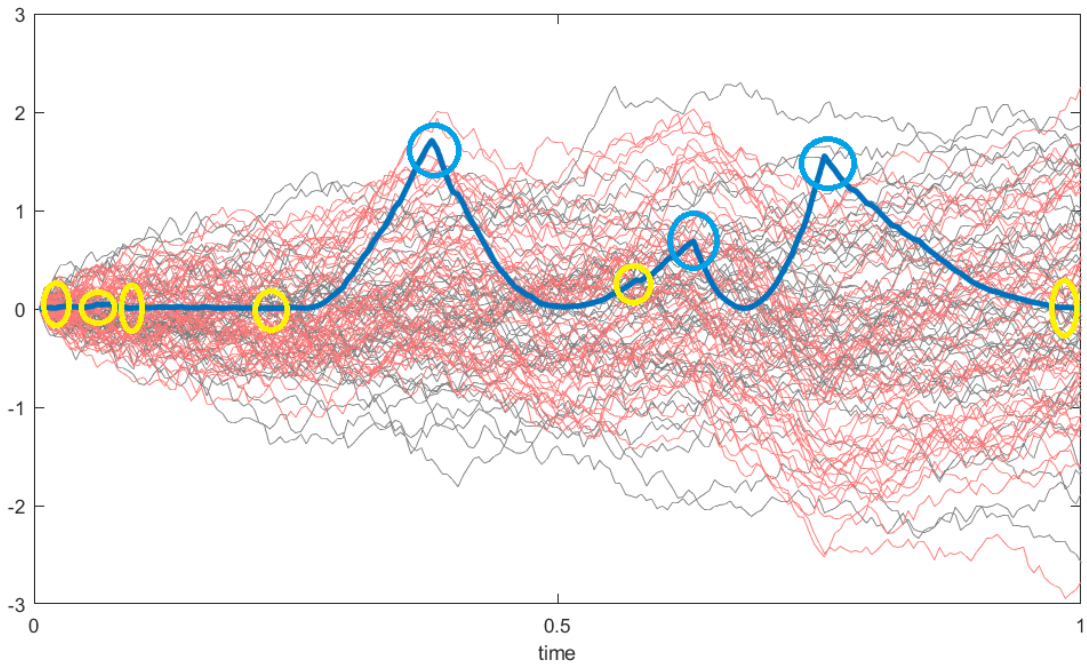
- Choose your favourite non-negative measure of statistical dependence $I(\cdot, \cdot)$.
- Calculate the relevance function $R(t) = I(X(t), Y)$.
- Select the points $t_1, \cdots, t_d$ in according to the local maxima of the relevance function $R(X(t), Y)$.

$$X(t_i) = \underset{t \in [0,1]}{argmax} \quad R(X(t), Y), \quad i = 1, 2, \ldots, d.$$



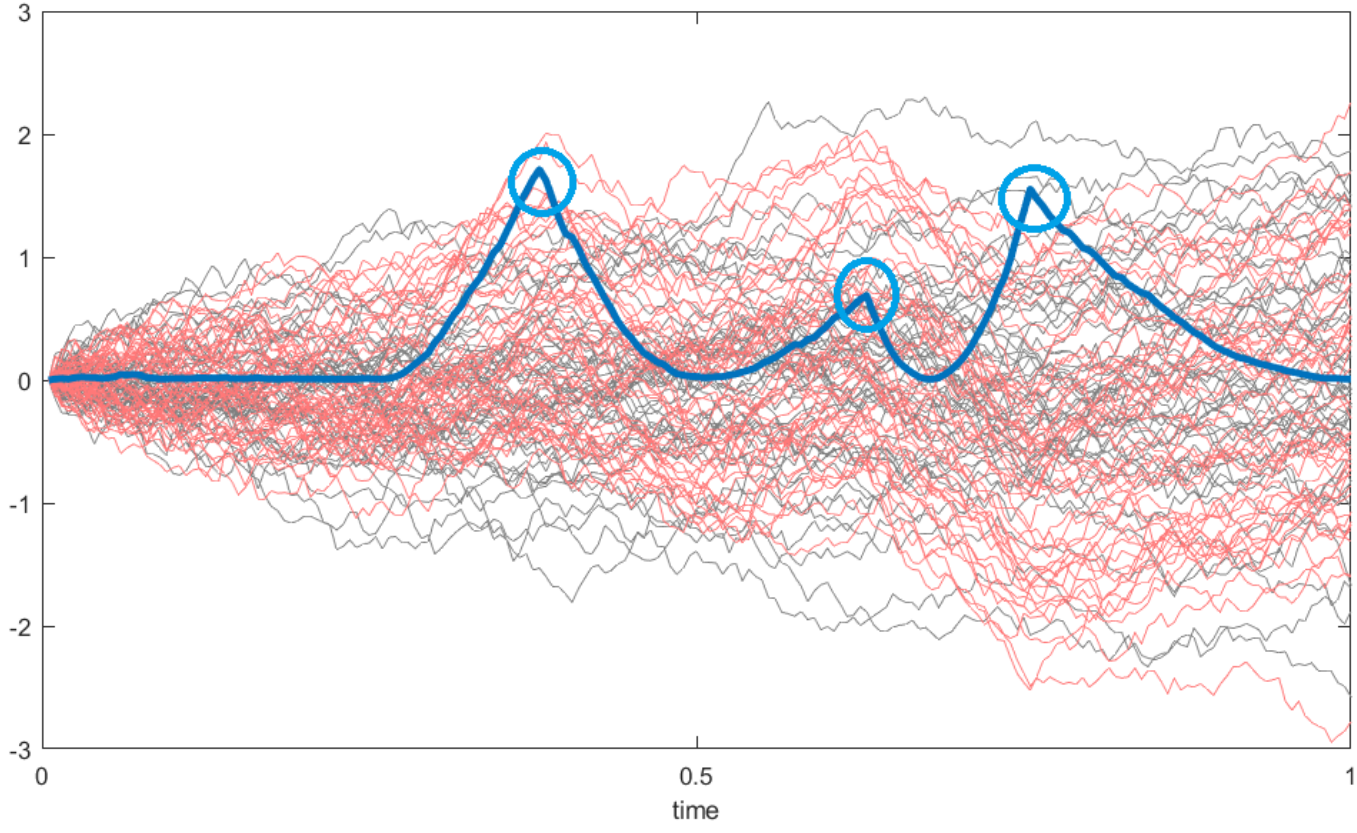
Some comments

- MH method takes care, in a natural way, of the relevance-reduncancy trade-off in the functional framework.
- It is "really functional" with a clear population target.
- There are some non-trivial computational problems to identify the local maxima of the relevance function.



Some comments

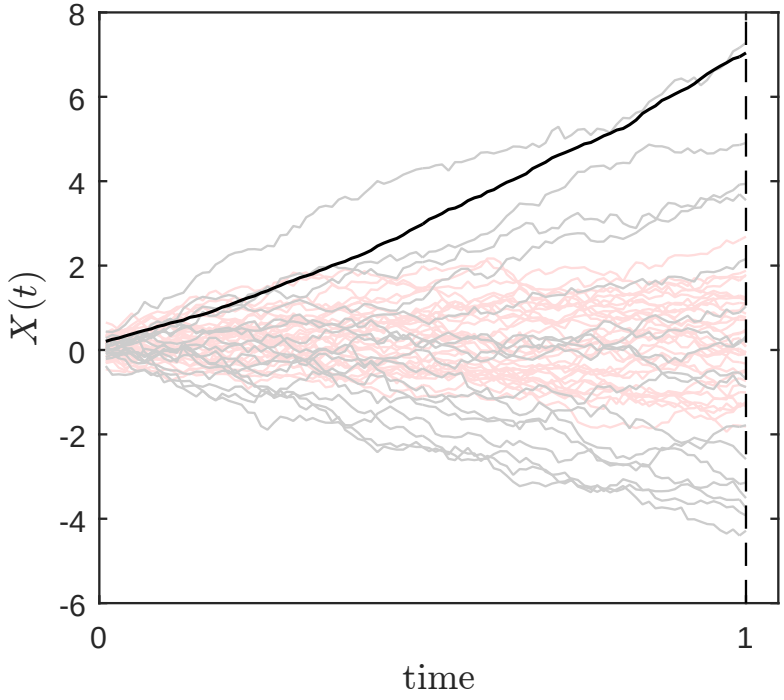
The empirical results show a remarkable good performance of MH methods in comparison with other state-of-art alternatives.

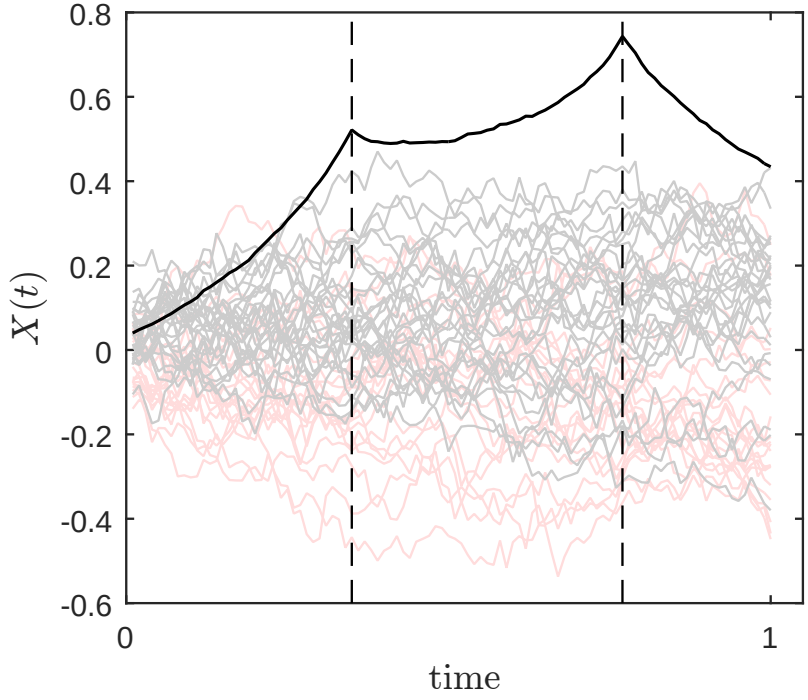


$acc_{KNN}X(\mathcal{I}) = 91.4\%$ $\qquad$ $acc_{KNN}X(t_1, t_2, t_3) = 93.5\%$

Some examples

Several non-trivial examples where the relevant information is concentrated on the maxima of $\mathcal{V}^2(X(t), Y)$.





Theoretical results

Some equivalent expressions for $\mathcal{V}^2(X(t), Y)$ in the binary case.

Theorem (uniform convergence of $\mathcal{V}_n^2$)

Let $X = X_t$, with $t \in [0,1]^d$, be a process with continuous trajectories almost surely such that $\mathbb{E}(\|X\|_\infty \log^+ \|X\|_\infty) < \infty$. Then, $\mathcal{V}_n^2(X_t, Y)$ is continuous in t and

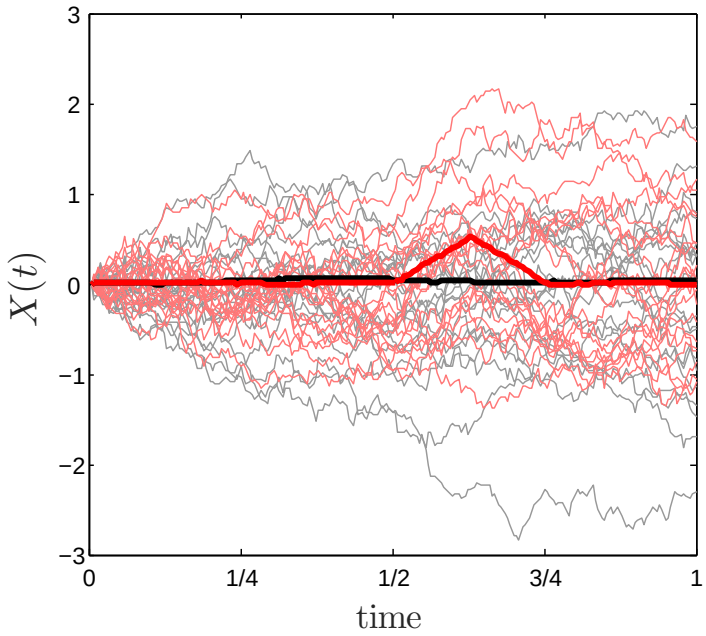
$$\sup_{t \in [0,1]^d} |\mathcal{V}_n^2(X_t, Y) - \mathcal{V}^2(X_t, Y)| \to 0 \quad \text{a.s., as } n \to \infty.$$

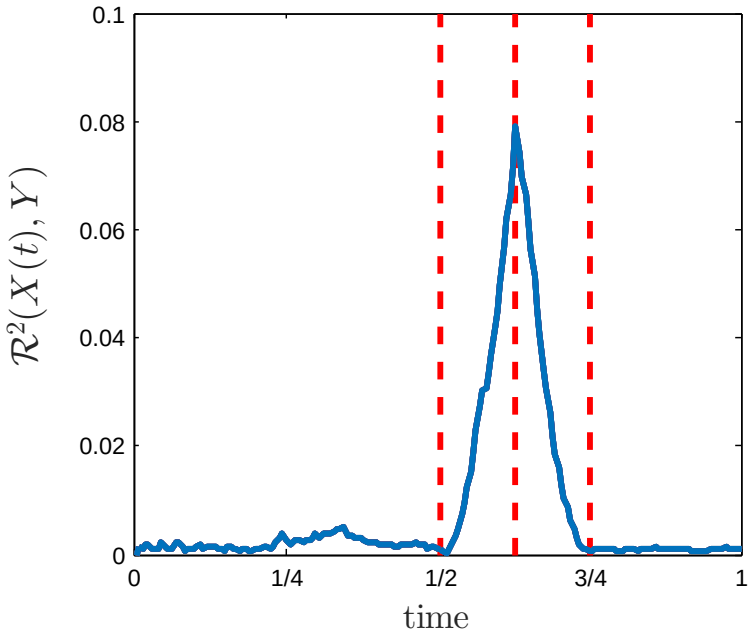
Hence, if we assume that $\mathcal{V}^2(X_t, Y)$ has exactly d local maxima at $t_1, \cdots, t_d$, then $\mathcal{V}_n^2(X_t, Y)$ has also eventually at least d maxima at $t_{1n}, \cdots, t_{dn}$ with $t_{jn} \to t_j$, as $n \to \infty$, a.s., for $j = 1, \ldots, d$.

Berrendero, Cuevas and Torrecilla. Variable selection in functional data classification: a maxima hunting proposal. *Statistica Sinica*, **26**:619-638 (2016)

Weak points

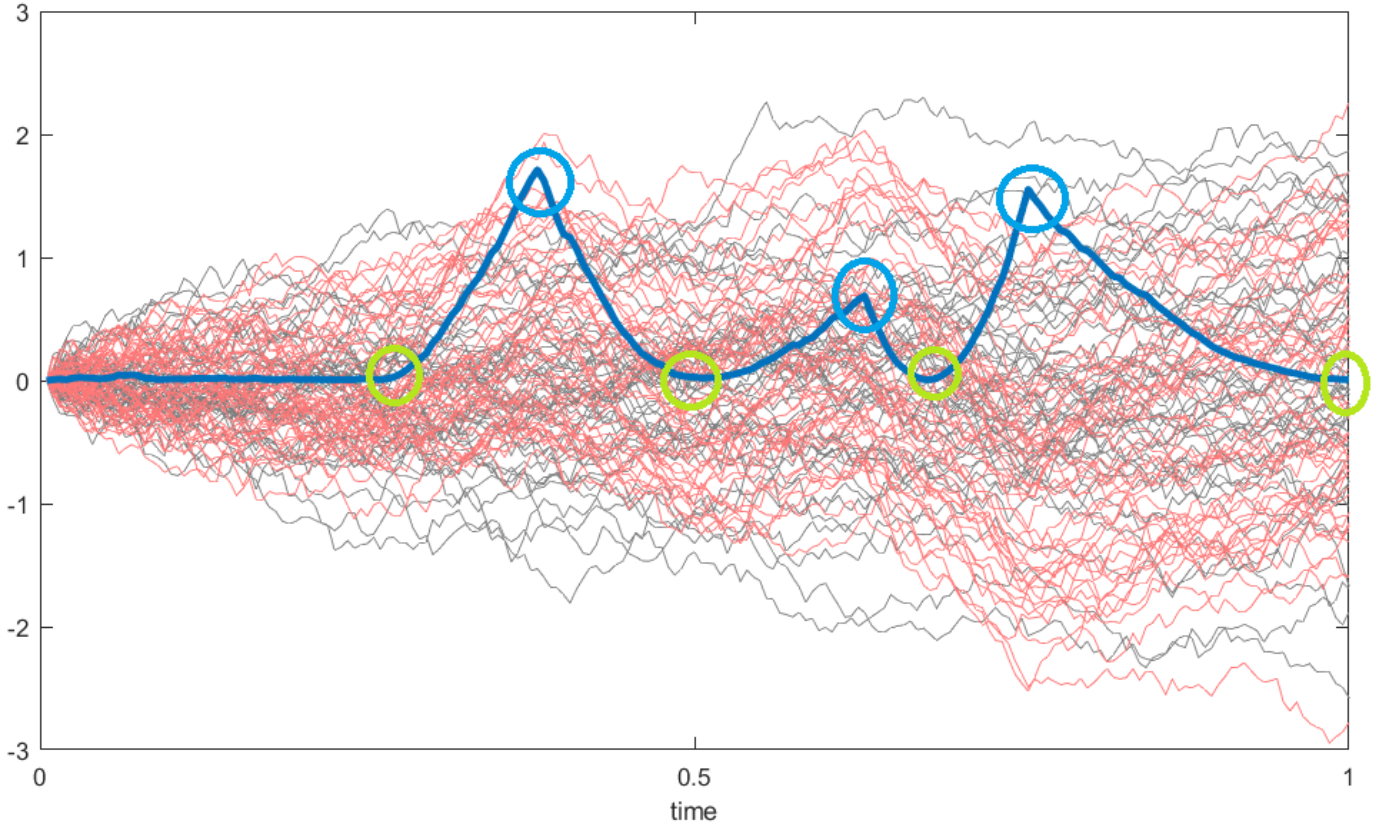
MH is easy to interpret and well motivated both empirically and theoretically. However, MH only considers the individual effect of the selected variables.





$$g^*(x) = 1 \Leftrightarrow \left(X\left(\frac{5}{8}\right) - X\left(\frac{1}{2}\right)\right) + \left(X\left(\frac{5}{8}\right) - X\left(\frac{3}{4}\right)\right) > \frac{1}{4}.$$

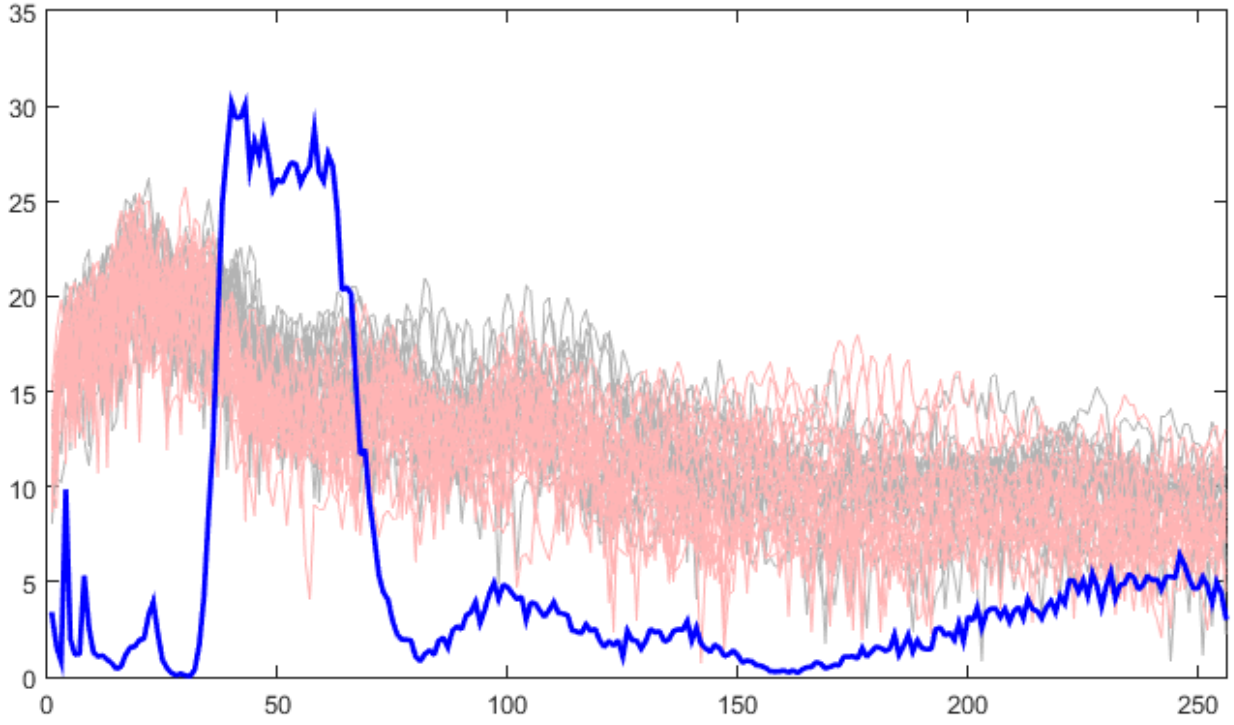
Previous example



$acc_{KNN}X(\mathcal{I}) = 91.4\%$
$acc_{KNN}X(t_1, t_2, t_3) = 93.5$
$acc_{KNN}X(t_1, \cdots, t_7) = 96.3\%$

Weak points

And local maxima are difficult to define in practice.



In the end we need some kind of smoothing parameter.

Recursive Maxima-Hunting

We propose Recursive Maxima Hunting (RMH), an extension of MH that considers interactions between variables by subtracting iteratively the information associated with a selected variable in terms of the conditional expectation.

Torrecilla and Suárez. Feature selection in functional data classification with recursive maxima hunting. *Advances in Neural Information Processing Systems*, **29**:4835-4843 (2016)

The algorithm

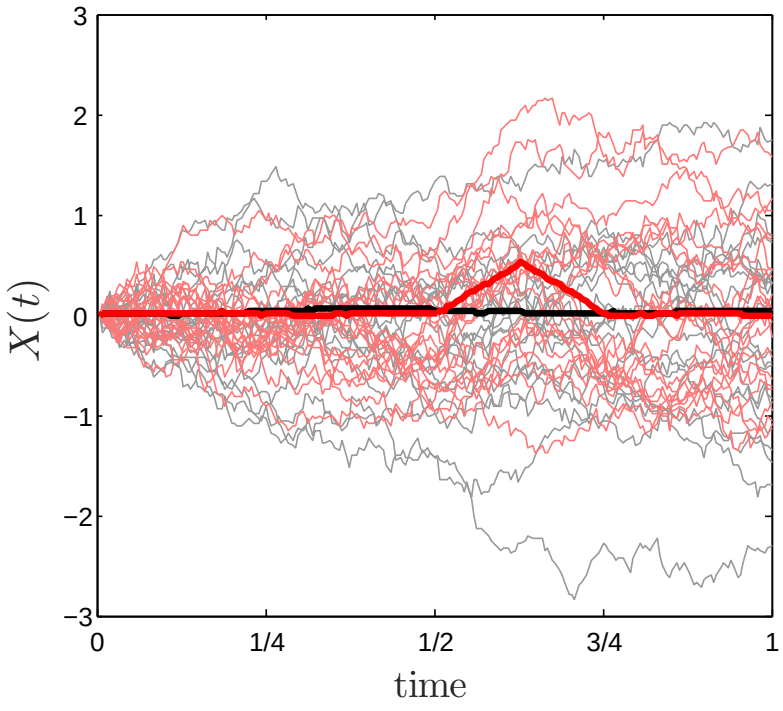
Input: $\{X_n(t), t \in [0,1]; Y_n \in \{0,1\}\}_{n=1}^{N_{train}}$,

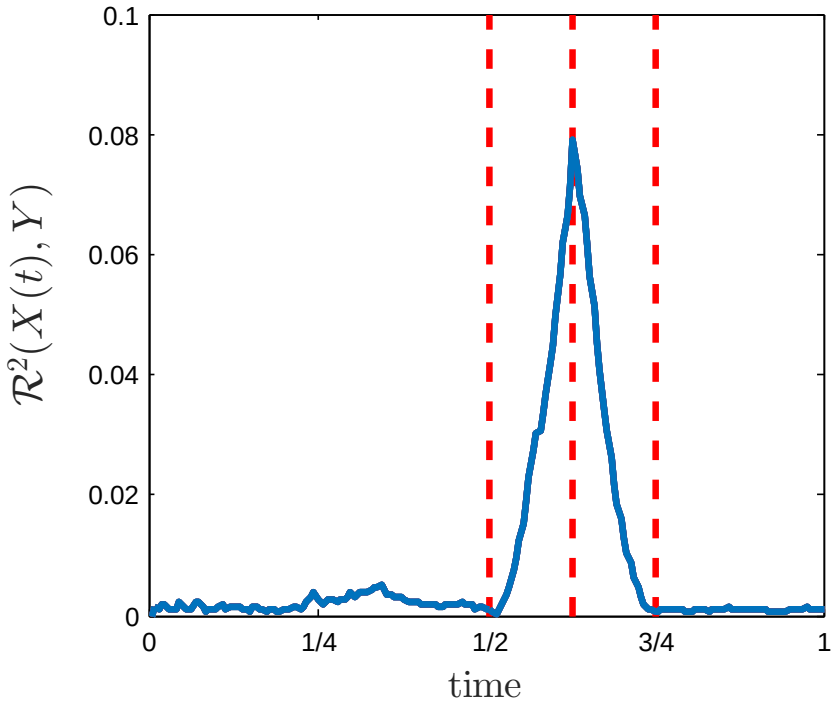
1. Select the globlal maximum $t_{max} \leftarrow \underset{t \in [0,1]}{argmax} \{I(X(t), Y)\}$
2. Until stopping criterion*
   3. Apply correction

$$X(t) \leftarrow X(t) - \mathbb{E}(Z(t) \mid Z(t_{max}) = X(t_{max})).$$

   4. Repeat the whole process excluding redundant points around the maximum,

An example





$$g^*(x) = 1 \Leftrightarrow \left[X\left(\frac{5}{8}\right) - X\left(\frac{1}{2}\right)\right] + \left[X\left(\frac{5}{8}\right) - X\left(\frac{3}{4}\right)\right] > \frac{1}{4}.$$

An example

$$\begin{cases} X(t) \mid Y = 0 : B(t) & , \quad t \in [0,1] \\ X(t) \mid Y = 1 : B(t) + m(t) & , \quad t \in [0,1] \end{cases},$$

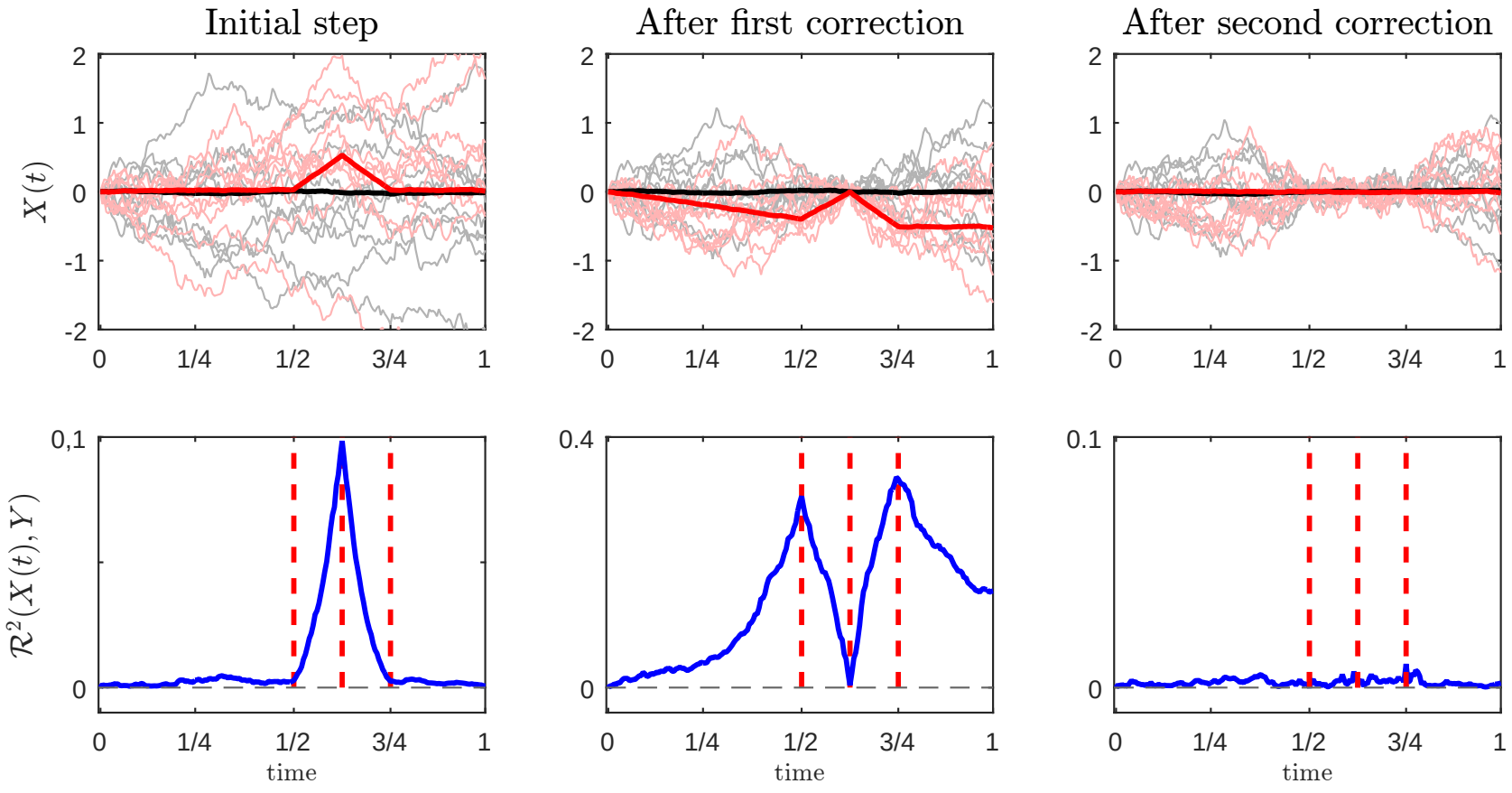
where $B(t)$ is standard Brownian motion, $m(t)$ is a deterministic trend. Assuming that the underlying process is Brownian

$$\mathbb{E}(B(t)|X(t_0)) = \frac{\min(t, t_0)}{t_0} X(t_0), \quad t \in [0,1].$$

Assuming that the underlying process is Brownian Bridge

$$\mathbb{E}(BB(t)|X(t_0)) = \frac{\min(t, t_0) - t\, t_0}{t_0(1-t_0)} X(t_0) = \begin{cases} \frac{t}{t_0} X(t_0), & t < t_0 \\ \frac{1-t}{1-t_0} X(t_0), & t > t_0. \end{cases}$$

An example


Initial step
After first correction
After second correction
$X(t)$
$\mathcal{R}^2(X(t), Y)$
time
time
time

Stopping criteria

In practice, a completely null relevance function is never achieved. Any conventional approach can be used:

- Cross-validation (loosing filter approach, partially).
- Threshold parameters (more or less arbitrary).
- ...

But we can also take advantage of the properties of the distance correlation measure. The idea is to stop selecting variables when the remaining variables are independent of the class label.

$$R = \left\{ \frac{n\mathcal{V}_n^2(X^{[i]}(t_i), Y)}{T_2(X^{[i]}(t_i), Y)} \geq \chi^2_{1-\alpha} \right\}, \text{Székely et al. (2007)}$$

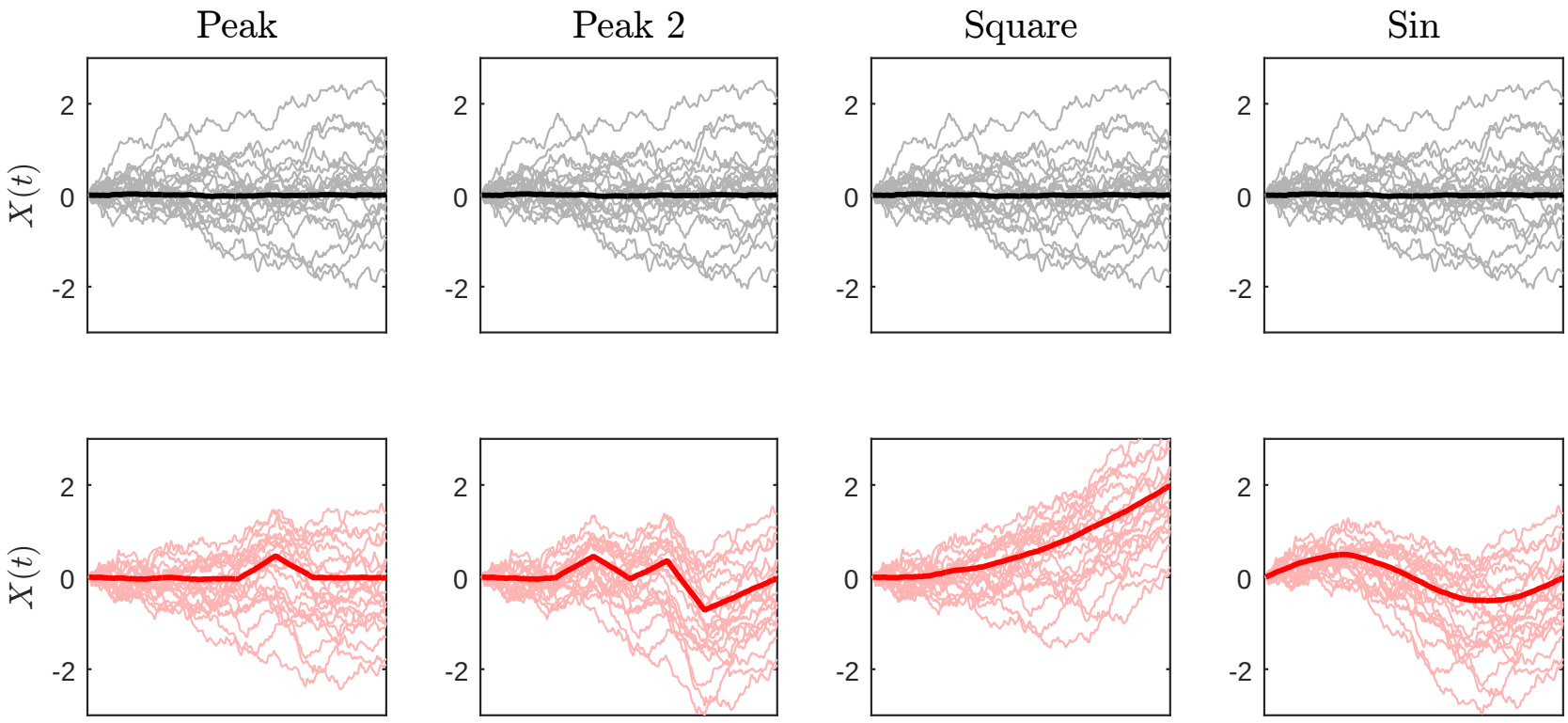
where n is sample size, $\mathcal{V}_n^2$ the empirical estimator of $\mathbb{V}^2$ and T_2 is defined for two random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ as

$$T_2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} ||X_k - X_l||_p \frac{1}{n^2} \sum_{k,l=1}^{n} ||Y_k - Y_l||_q,$$

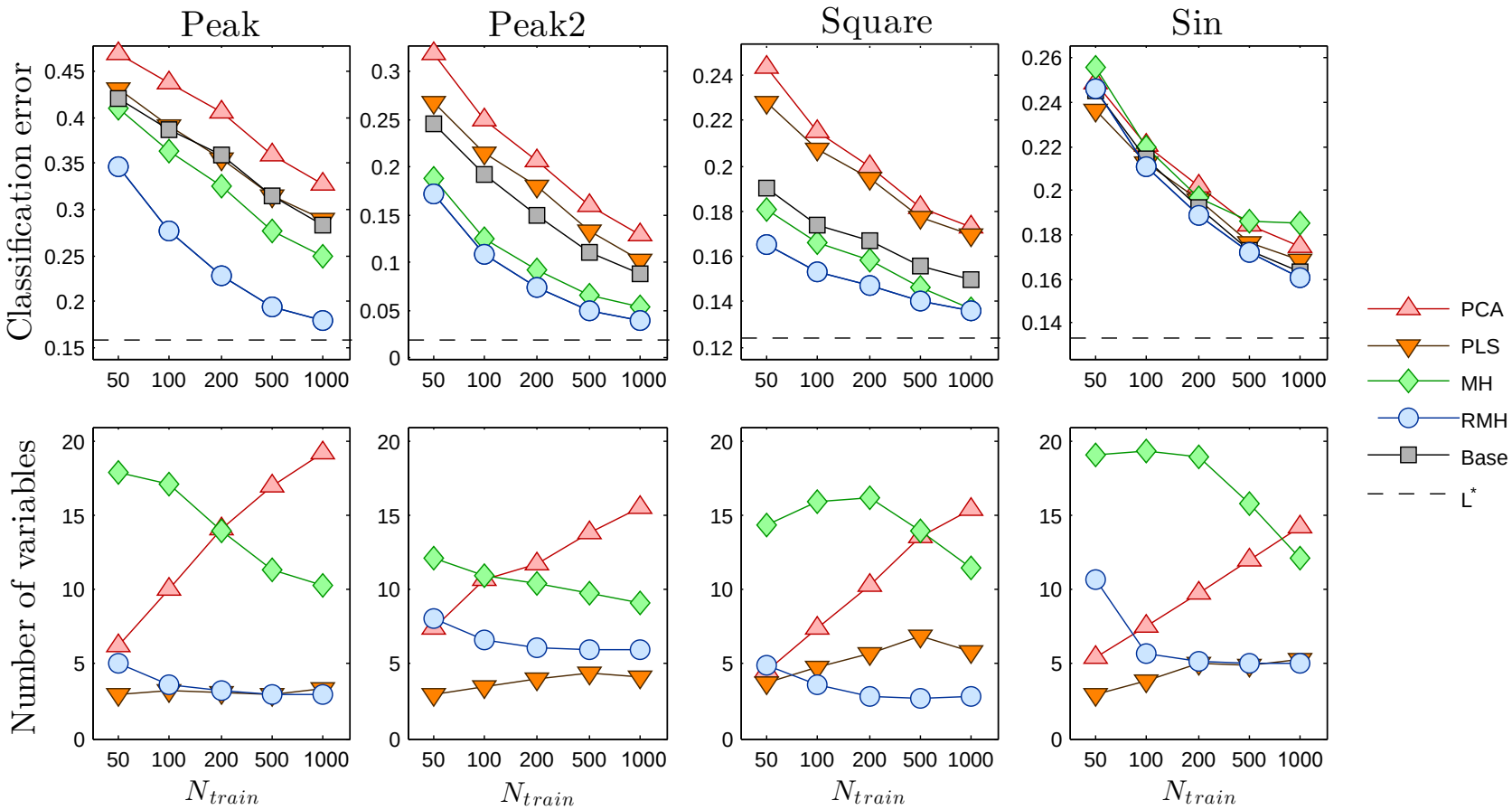
Some properties

- 'Optimality'. Under some non-trivial models RMH selects the variables which appear in the optimal classification rule.
- RMH only considers global maximum at each iteration avoiding problems related with local maxima.
- Convergence of the empirical maxima are guaranteed by the uniform convergence of $\mathcal{V}_n^2(X(t), Y)$.
- Once parameters are fixed, the number of relevant variables is given automatically.
- Corrections can be combined and only the initial kernel is needed.
- Markovianity allows "pure recursive" algorithm.
- Different kernels entail different properties.

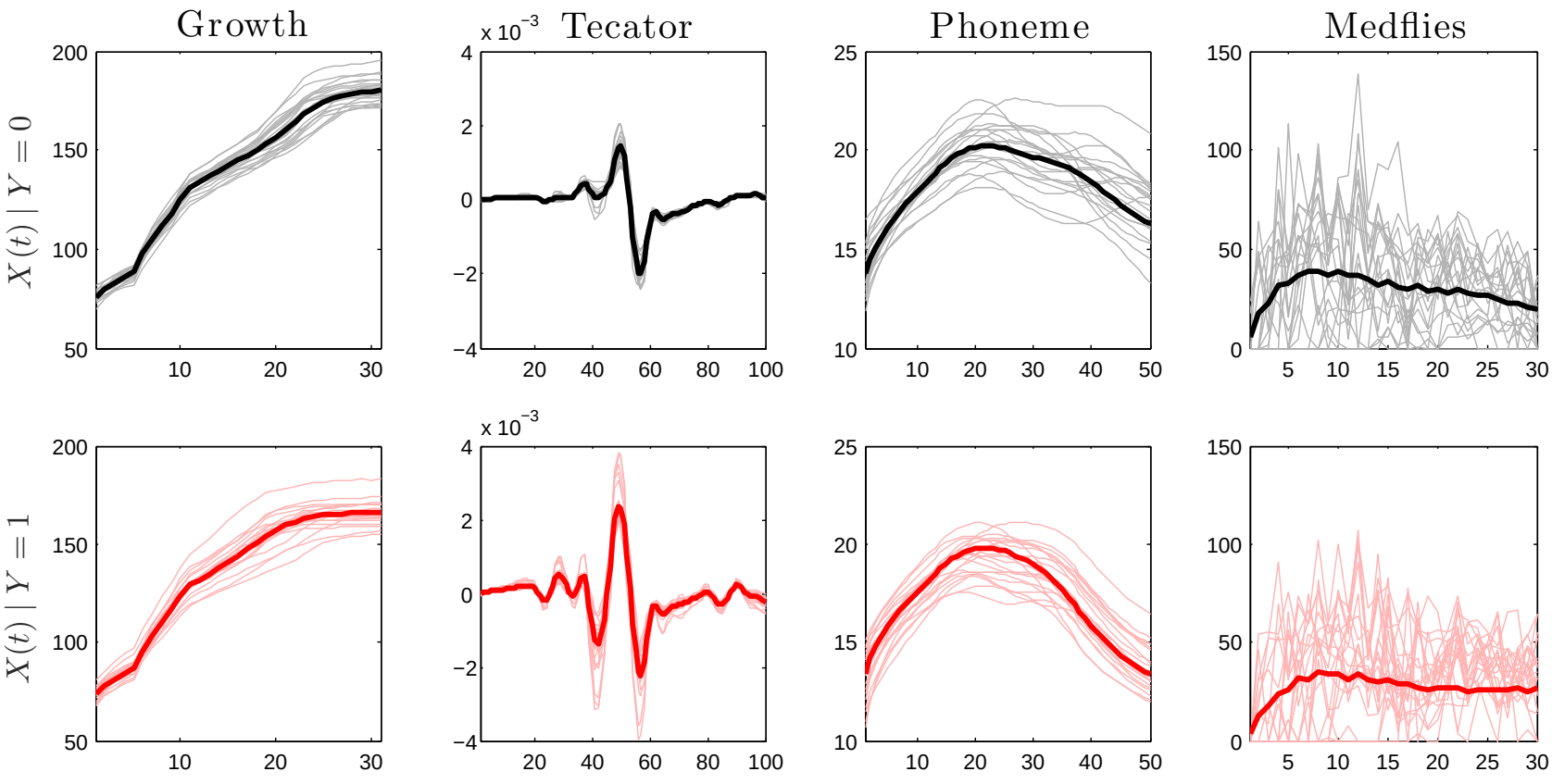
Simulations



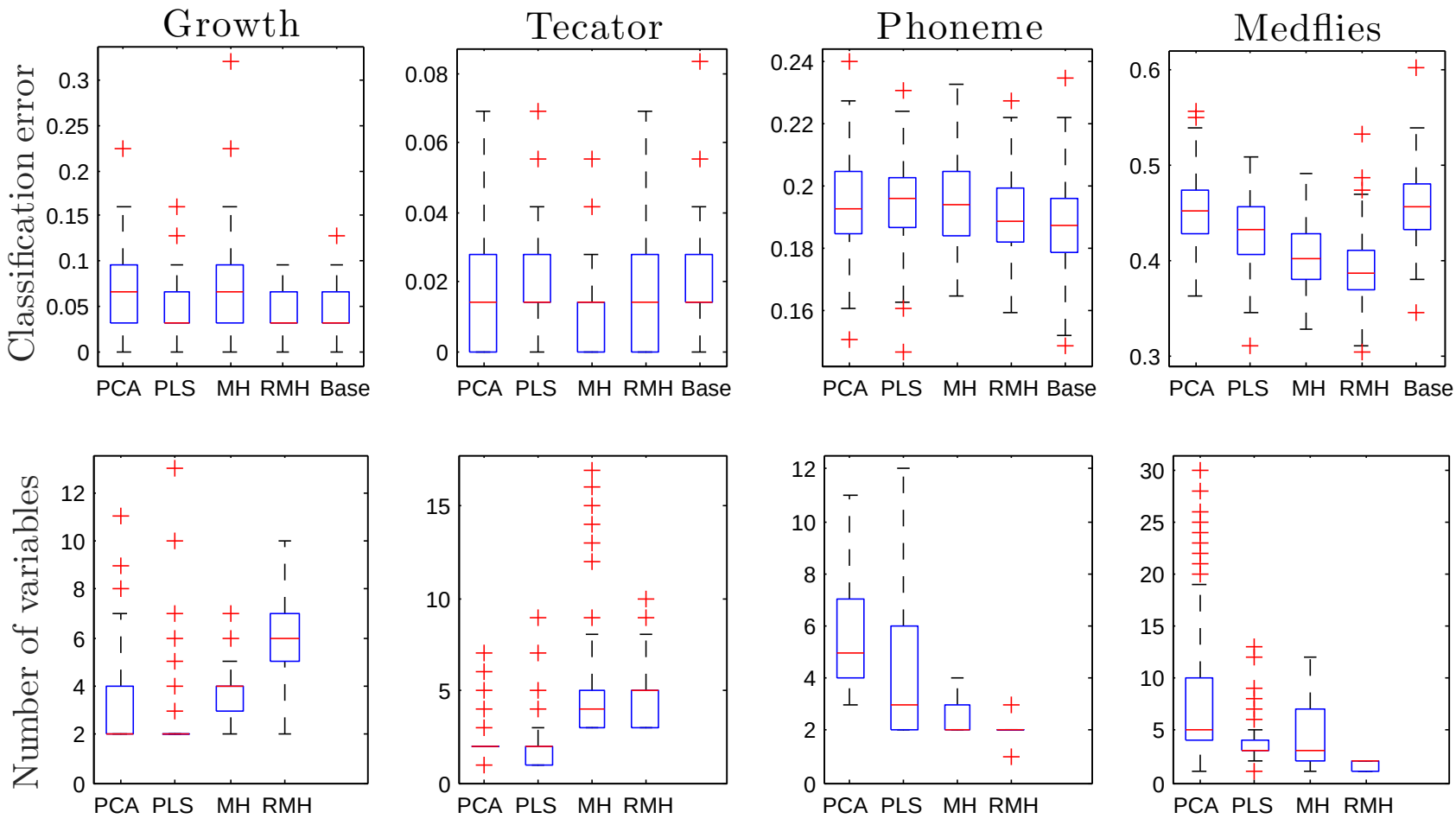
Simulations



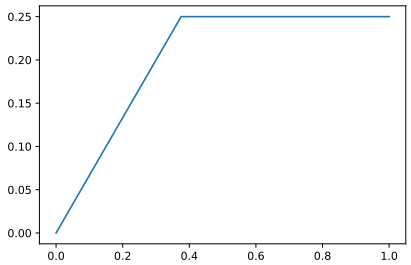
Real data



Real data

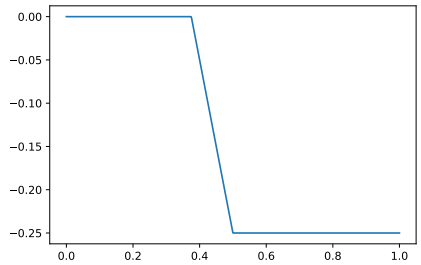

Growth
Tecator
Phoneme
Medflies
Classification error
Number of variables
PCA PLS MH RMH Base
PCA PLS MH RMH

Interpolation link



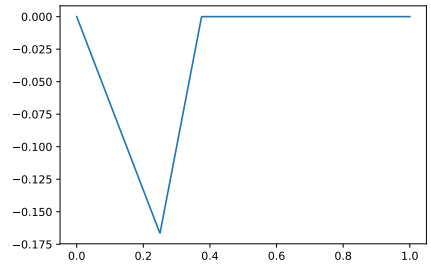
(a) First Corr.

+



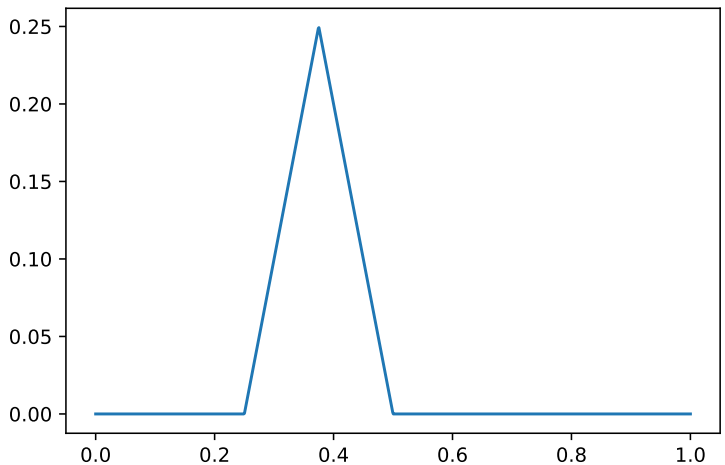
(b) Second Corr.

+

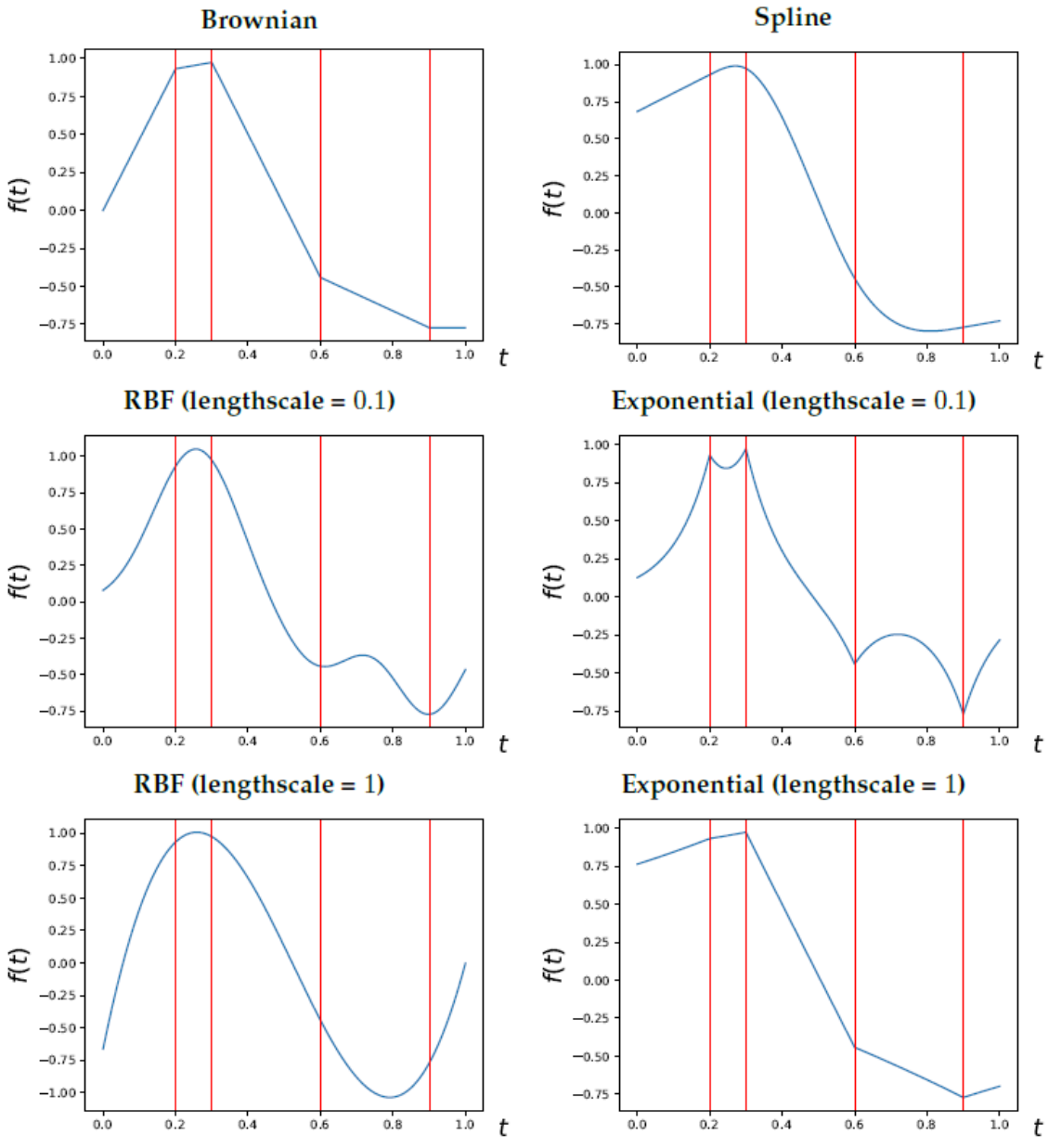


(c) Third Corr.

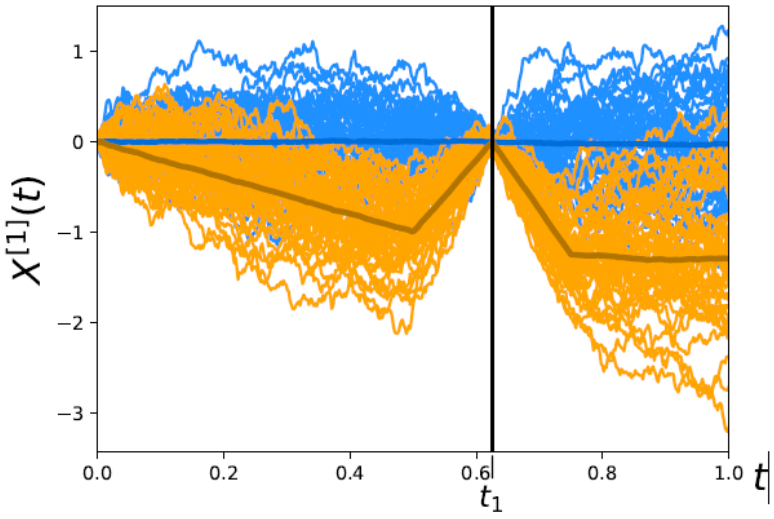
=

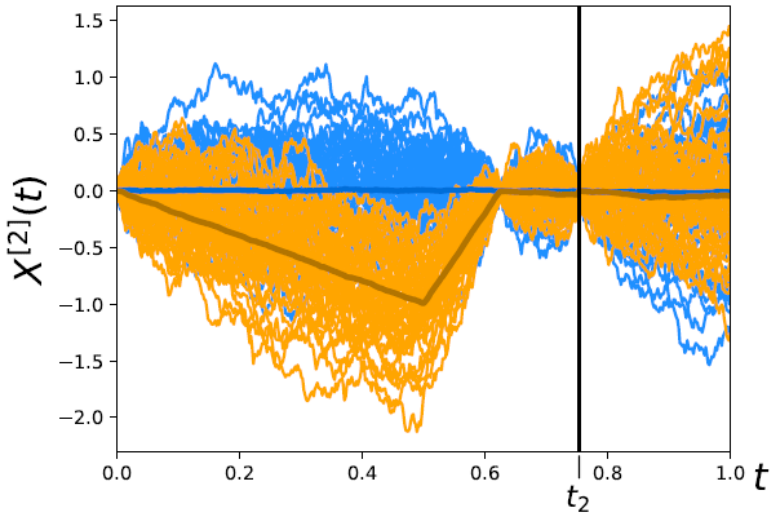


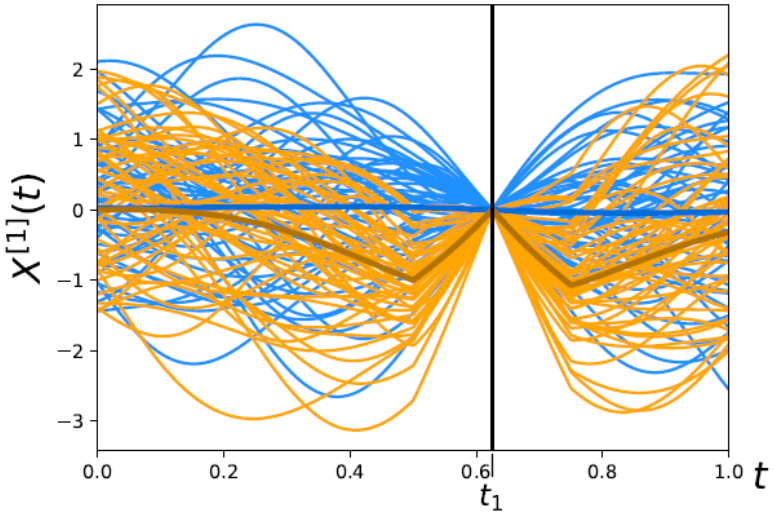
Interpolation and kernels

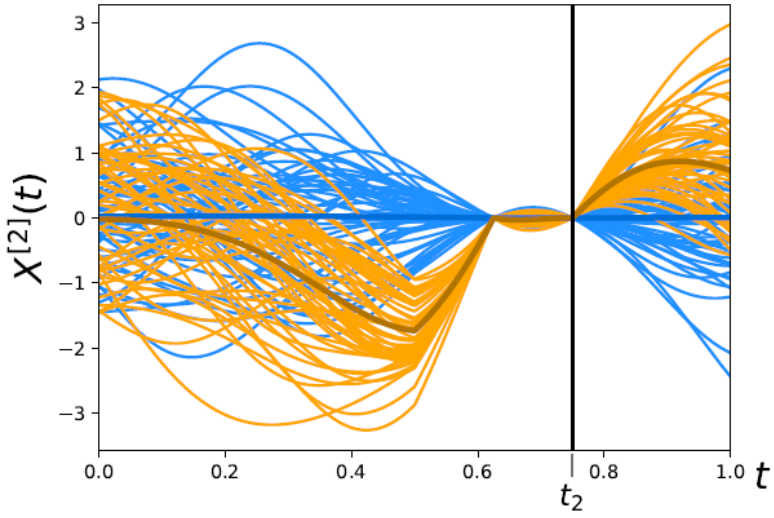


Connections with near-perfect classification









RKHS approach

"It turns out, in my opinion, that reproducing kernel Hilbert spaces are the natural setting in which to solve problems of statistical inference on time processes". Parzen, 1961

Why natural? RKHS provides an intrinsic inner product depending on the covariance structure.

Berrendero, Cuevas and Torrecilla. On the use of reproducing kernel Hilbert spaces in functional classification. J. Am. Stat. Assoc. To appear. DOI:10.1080/01621459.2017.1320287

- Explicit expressions of the Bayes rule (equivalent distributions).
- Approximate optimal rule under mutually singular distributions.
- Insight into the near "perfect classification phenomenon" (Delaigle and Hall 2012)
- Natural setting to formalize variable selection problems.

Variable selection and RKHS

Variable selection methods are quite appealing when classifying functional data since they help reduce noise and remove irrelevant information. RKHS also offers a natural setting to formalize variable selection problems.

The ability of RKHS to deal with these problems is mainly due to the fact that, by the reproducing property, the elementary functions $K(\cdot, t)$ act as Dirac's deltas.

$$\langle K(\cdot, t), X \rangle_K = X(t)$$

Sparsity assumption [SA]: there exist scalars $\alpha_1^*, \ldots, \alpha_d^*$ and points $t_1^*, \ldots, t_d^*$ in $[0, T]$ such that $m(\cdot) = \sum_{i=1}^{d} \alpha_i^* K(\cdot, t_i^*)$.

Bayes rule under the sparsity assumption

Under this assumption, the Bayes rule depends on the trajectory $x(t)$ only through the values $x(t_1^*), \ldots, x(t_d^*)$.

$$\eta^*(x) = \sum_{i=1}^{d} \alpha_i^* \left(x(t_i^*) - \frac{m_0(t_i^*) + m_1(t_i^*)}{2} \right) - \log\left(\frac{1-p}{p}\right),$$

where $(\alpha_1^*, \ldots, \alpha_d^*)^\top = K^{-1}_{t_1^*,\ldots,t_d^*} m_{t_1^*,\ldots,t_d^*}$.

$K_{i,j} = K(t_i^*, t_j^*)$

$m_{t_1^*,\ldots,t_d^*} = (m(t_1^*), \ldots, m(t_d^*))$.

This shows that under [SA], the Bayes rule coincides with the well-known Fisher linear rule based on the projections $x(t_1^*), \ldots, x(t_d^*)$.

$$L^* = \mathbb{P}\{g^*(X) \neq Y\} = 1 - \Phi\left(\frac{\| m \|_K}{2}\right)$$

$$\|m\|_K^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} \alpha_i^* \alpha_j^* K(t_i^*, t_j^*) = m_{t_1^*,\ldots,t_d^*}^\top K_{t_1^*,\ldots,t_d^*}^{-1} m_{t_1^*,\ldots,t_d^*}.$$

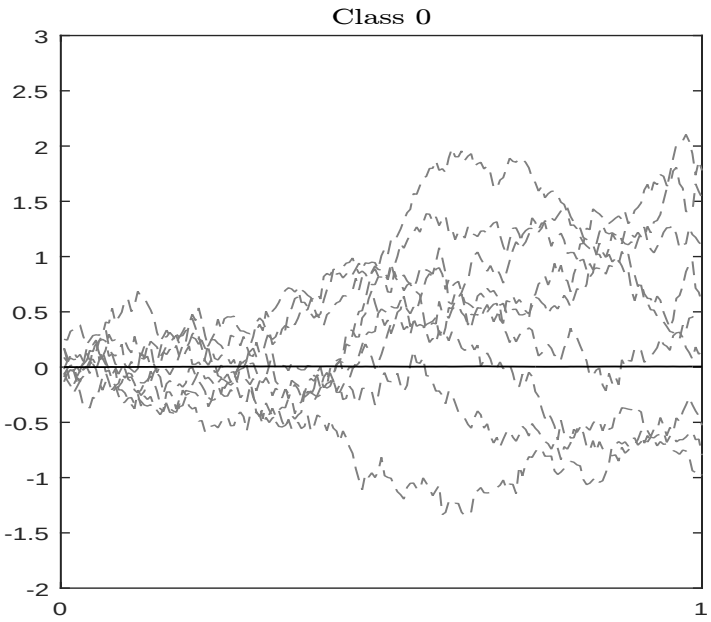
The criterion we suggest for variable selection is to choose points $\hat{t}_1, \ldots, \hat{t}_d$ maximizing

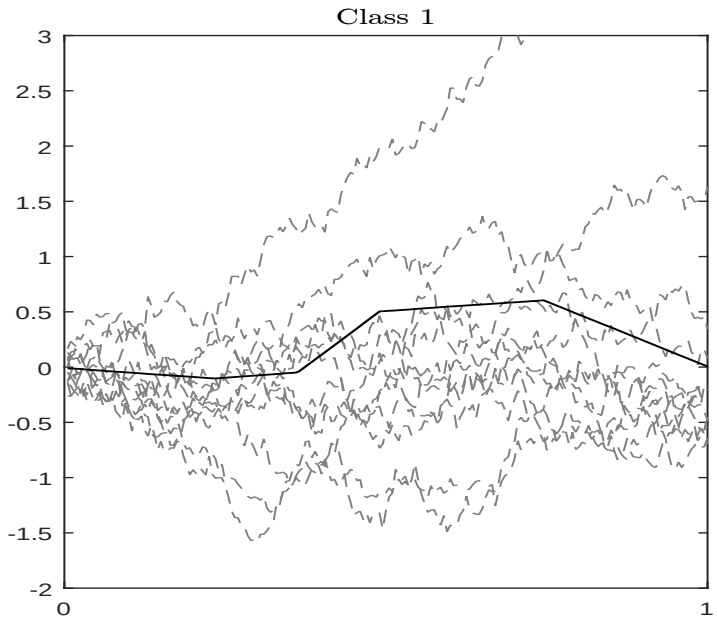
$$\hat{\psi}(t_1, \ldots, t_d) := \hat{m}_{t_1,\ldots,t_d}^\top K_{t_1,\ldots,t_d}^{-1} \hat{m}_{t_1,\ldots,t_d}.$$

The algorithm

1. Initial step: consider a large enough grid of points in $[0, T]$ and find $\hat{t}_1$ such that $\hat{\psi}(\hat{t}_1) \geq \hat{\psi}(t) \forall t \in [0, T]$. This step amounts to find the point maximizing the signal-to-noise ratio since
$$\hat{\psi}(t) = \frac{\hat{m}(t)^2}{\hat{\sigma}_t^2} = \frac{(\bar{X}_1(t) - \bar{X}_0(t))^2}{\hat{\sigma}_t^2}.$$
2. Repeat until convergence: once we have computed $\hat{t}_1, \ldots, \hat{t}_{d-1}$, find $\hat{t}_d$ such that
$$\hat{\psi}(\hat{t}_1, \ldots, \hat{t}_{d-1}, \hat{t}_d) \geq \hat{\psi}(\hat{t}_1, \ldots, \hat{t}_{d-1}, t)$$
for all t in rest of the grid.

An example



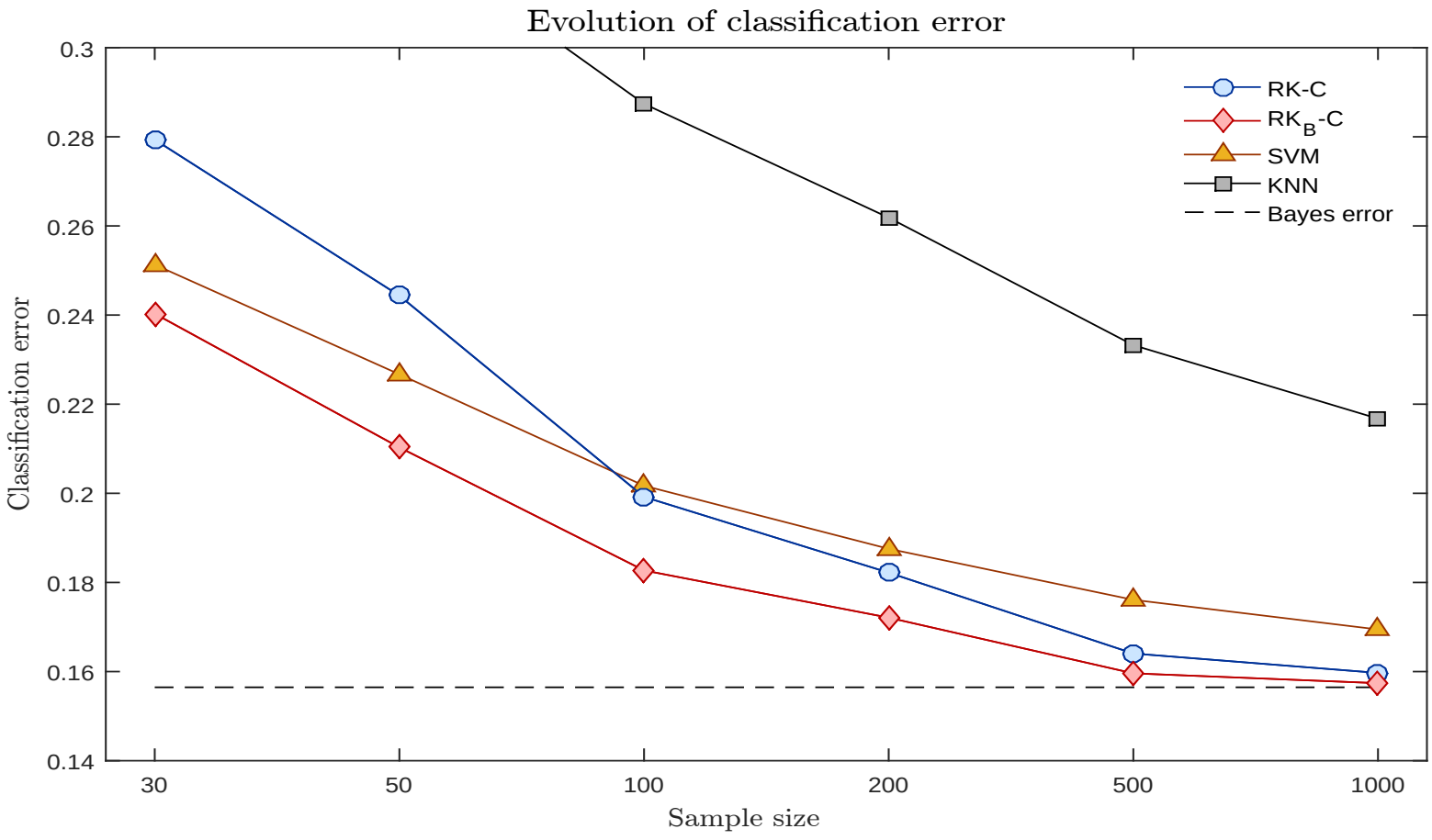


$K(s,t) = \min\{s,t\}$.

$t^* = \{0, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1\}$.

$L^* = 0.1587$.

An example (II)



An example (III)