On the Theory and Practice of Variable Selection for Functional Data

José Luis Torrecilla

under the supervision of
José Ramón Berrendero and Antonio Cuevas

Departamento de Matemáticas
Universidad Autónoma de Madrid

Lectura de tesis
Madrid - December 3, 2015

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

What are functional data?

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{I} \subseteq \mathbb{R}$ an index set, an stochastic process is a collection of random variables $\{X(\omega, t) : \omega \in \Omega, t \in \mathcal{I}\}$ where $X(\cdot, t)$ is an $\mathcal{F}$-measurable function on Ω.

A functional data is just a realization (often called "trajectory") of a stochastic process for all $t \in [0, T]$.

Difficulties and particularities

- No obvious order structure (distribution functions), nor closeness or centrality notions (outliers, depth).
- Representation problems.
- Function spaces are "difficult to fill".
- No natural densities: no natural translation-invariant measure plays the role of Lebesgue measure in $\mathbb{R}^n$.
- Redundancy: close variables are closely related (continuity). Fails in linear models.
- High dimension: the curse of the dimensionality, overvitting, computational cost...

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Variable selection

- **Idea** Choose the most informative subset among the original variables.
- **Motivation**
  - Variable selection is a successful technique of dimension reduction in other fields.
  - This was an almost unexplored topic in FDA classification.
  - The dimension reduction is made in terms of the original variables (interpretability).
- **Goals**
  - Remove useless and redundant variables improving temporal and storage performance.
  - Improve the classification accuracy decreasing the overfitting risk.
  - Get theoretical and more interpretable models.

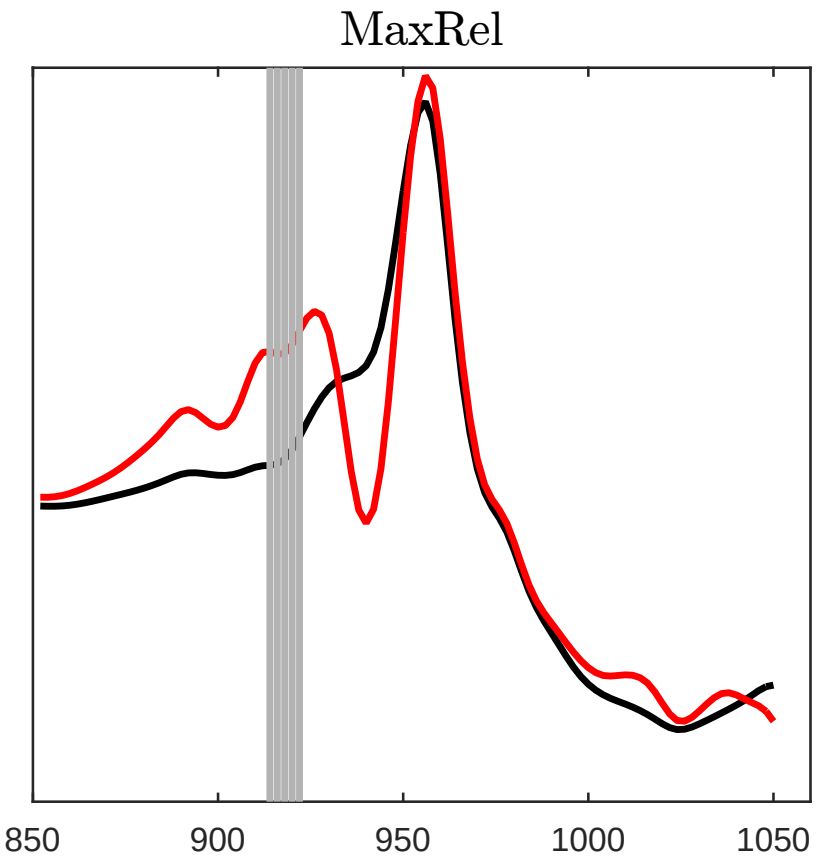
What do we mean by "variable selection" in FDA?

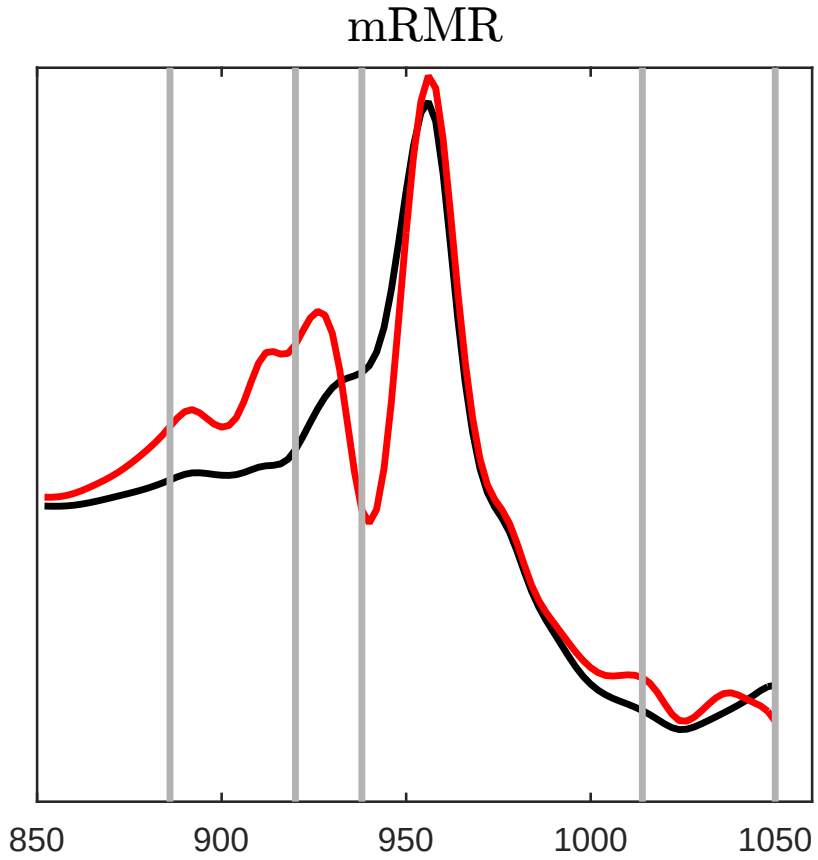
- Given a sample of functions $X_1(t), \cdots, X_n(t)$, $t \in [0,1]$ our aim is to replace every sample function X_j with a vector

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

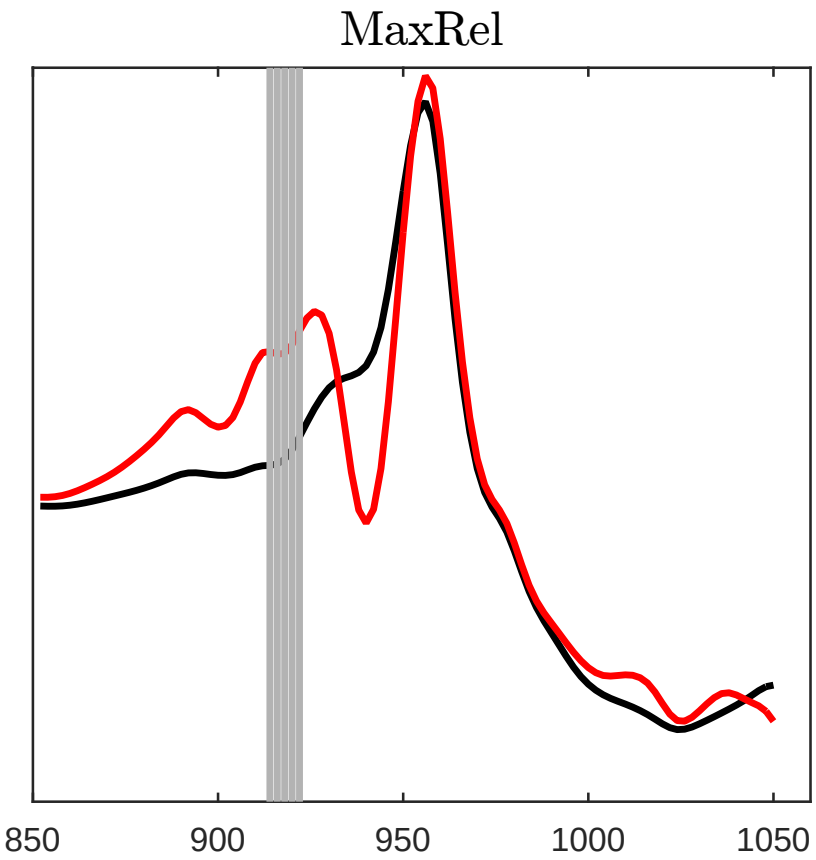
  for suitably chosen points $t_1, \cdots, t_d$.
- Then we would apply multivariate methods (regression, classification,...) to the "reduced" data.
- According to our experience, the value of d should be typically small (not much larger than 5, say).

Relevance Vs. Redundancy

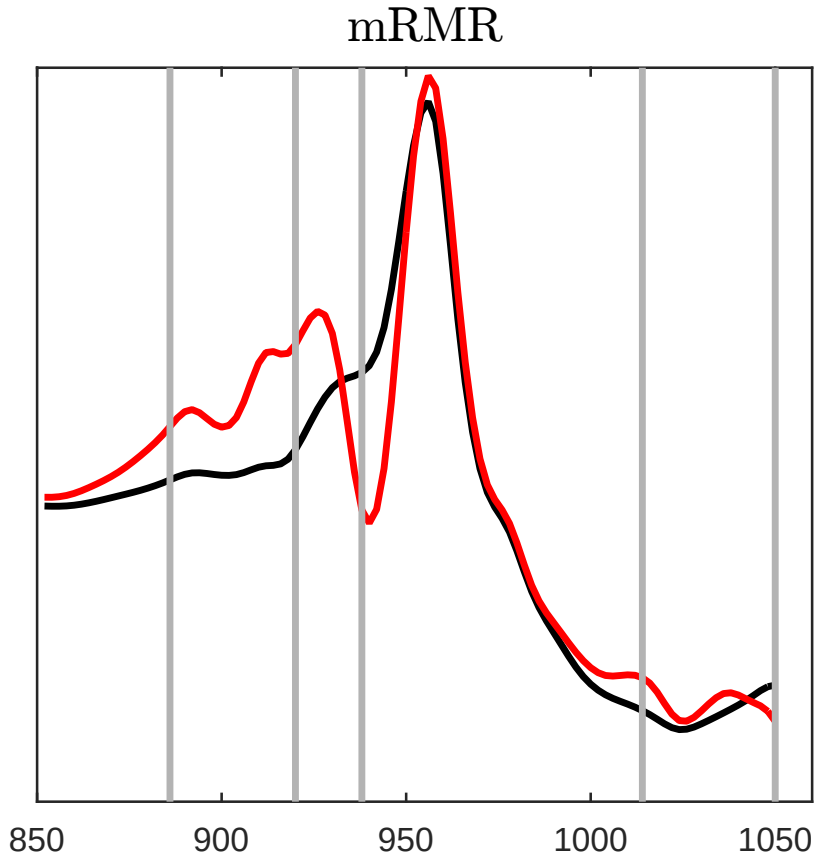




Relevance Vs. Redundancy



$err = 4.09\%$

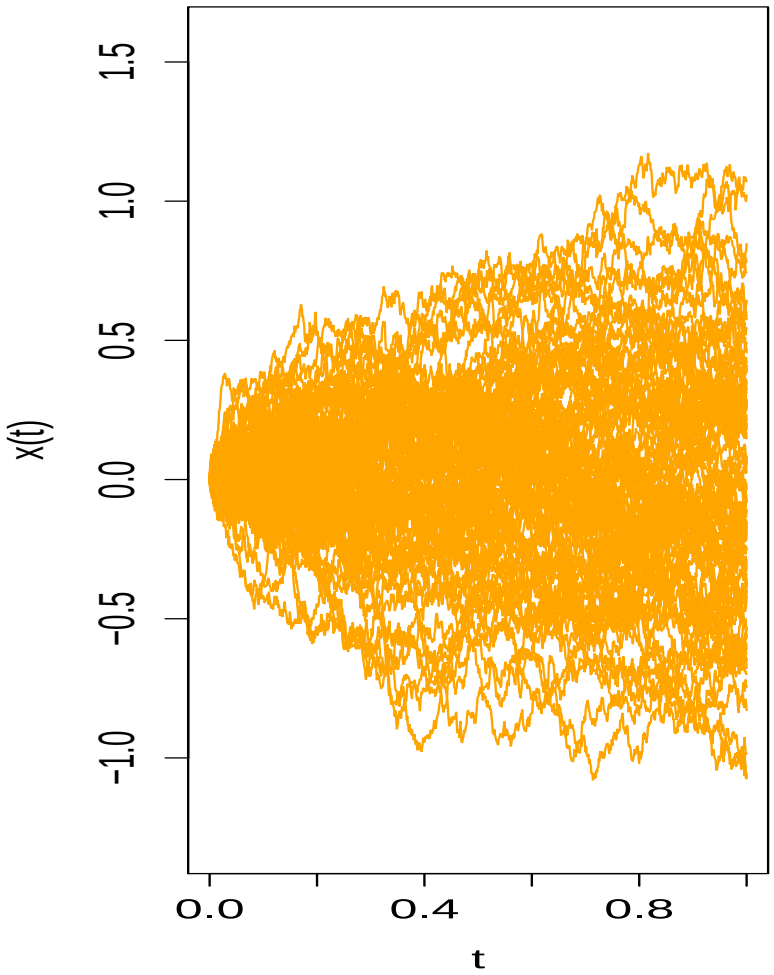


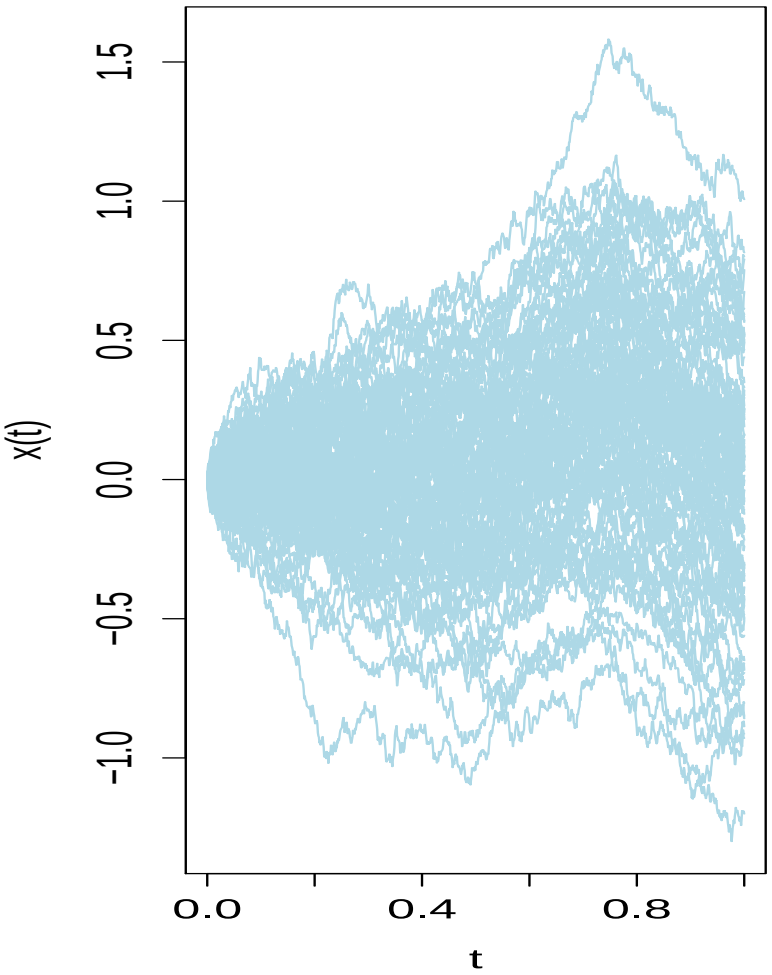
$err = 1.86\%$

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

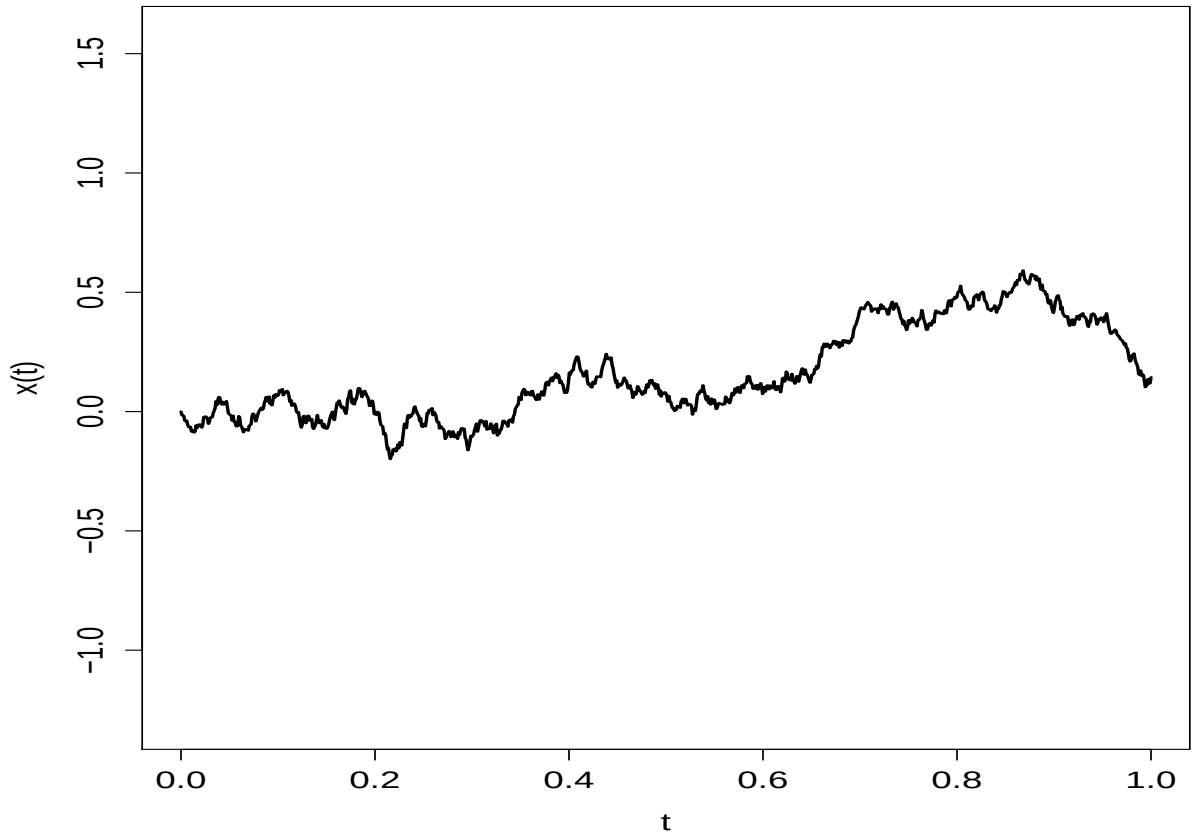
Functional classification problem





Functional classification problem (II)

Which is the class of this trajectory?



Statement of the problem

Independent observations: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
$X \in \mathcal{F}[0, T]$
$Y \in \{0, 1\}$

Statement of the problem

Independent observations: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
$X \in \mathcal{F}[0, T]$
$Y \in \{0, 1\}$

Optimal classification rule (Bayes rule)

$$g^*(X) = \mathbb{I}_{\{\eta(X) > 1/2\}},$$

where $\eta(x) = \mathbb{E}(Y|X = x)$.

Bayes Error

$$L^* = \mathbb{P}(g^*(X) \neq Y).$$

Statement of the problem

Independent observations: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
$X \in \mathcal{F}[0, T]$
$Y \in \{0, 1\}$

Optimal classification rule (Bayes rule)

$$g^*(X) = \mathbb{I}_{\{\eta(X)>1/2\}},$$

where $\eta(x) = \mathbb{E}(Y|X = x)$.

Bayes Error

$$L^* = \mathbb{P}(g^*(X) \neq Y).$$

$$g^*(X) = 1 \Leftrightarrow \frac{dP_1}{dP_0}(X) > \frac{1-p}{p}$$

See Baíllo et al., Scand. J. Stat. (2011), Theorem 1

Our general approach

- We consider the functional data as trajectories drawn from a stochastic process.
- We have tried to motivate our results and proposals in terms of this underlying stochastic process.
- This is somewhat in contrast with the mainstream research line in FDA, mostly centred in algorithmic aspects and real data analysis.

"Curiously, despite a huge research activity in the field, few attempts have been made to connect the area of functional data analysis with the theory of stochastic processes" Biau et al. 2015

Contributions

a) A mathematical contribution to the functional classification problem (RKHS)

b) Functional variable selection: a theoretical motivation and three different proposals.

c) Large and replicable simulation studies.

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

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.

- 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 (RK-VS and associated classifier).

Berrendero, Cuevas and Torrecilla. On near perfect classification and functional Fisher rules via reproducing kernels. Manuscript. arXiv:1507.04398v2.

Some background

Definition: If $X = \{X_t, t \in [0, T]\}$ is a L^2-process with covariance function $K(s, t)$, define $(\mathcal{H}_0(K), \langle \cdot, \cdot \rangle)$ by

$$\mathcal{H}_0(K) := \{f : \ f(s) = \sum_i^n a_i K(s, t_i), \ a_i \in \mathbb{R}, \ t_i \in [0, T], \ n \in \mathbb{N}\}$$

$$\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_j, t_i),$$

where $f(x) = \sum_i \alpha_i K(x, t_i)$ and $g(x) = \sum_j \beta_j K(x, s_j)$.

The RKHS associated with K, $\mathcal{H}(K)$, is defined as the completion of $\mathcal{H}_0(K)$. More precisely, $\mathcal{H}(K)$ is the set of functions $f : [0, T] \to \mathbb{R}$ obtained as t pointwise limit of a Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0(K)$.

Some background (II)

Reproducing property: $f(t) = \langle f, K(\cdot, t)\rangle_K$, for all $f \in \mathcal{H}(K)$.

Natural congruence: If $\bar{\mathcal{L}}(X)$ is the L^2-completion of the linear span of X, $\Psi_X(\sum_i a_i X_{t_i}) = \sum_i a_i K(\cdot, t_i)$ defines a congruence between $\bar{\mathcal{L}}(X)$ and $\mathcal{H}(K)$.

The model

The model

$$\begin{cases} P_0: & X(t) = & m_0(t) + \xi(t), & t \in [0,1] \\ P_1: & X(t) = & m_1(t) + \xi(t), & t \in [0,1] \end{cases}$$

- $\xi(t)$ Gaussian with $\mathbb{E}(\xi(t)) = 0$.
- $K(s,t) = \mathbb{E}(\xi(s)\xi(t))$
- Prior probabilities: $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$.
- $m(t) = m_1(t) - m_0(t)$.

Parzen's result

Theorem 7A (Parzen, Ann. Math. Stat. 1961)

Under this model, if K is continuous

$$P_0 \sim P_1 \Leftrightarrow m \in \mathcal{H}(K),$$

and if $P_0 \sim P_1$

$$\frac{dP_1}{dP_0}(X) = \exp\left\{\langle m, X\rangle_K - \frac{1}{2}\langle m, m\rangle_K\right\}$$

- $\langle m, X\rangle_K \equiv \Psi_X(m)$.
- $\langle K(\cdot, t), X\rangle_K = X(t)$.

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Equivalent measures: the new optimal rule

Bayes Rule (Theorem 2.2)

Under the given model, if $m \in \mathcal{H}(K)$ then

$$g^*(X) = 1 \Leftrightarrow \eta^*(X) = \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

Equivalent measures: the new optimal rule

Bayes Rule (Theorem 2.2)

Under the given model, if $m \in \mathcal{H}(K)$ then

$$g^*(X) = 1 \Leftrightarrow \eta^*(X) = \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

Bayes error

(1) $\eta^*(X)|Y = 0 \sim N\left(-\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$

(2) $\eta^*(X)|Y = 1 \sim N\left(\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$

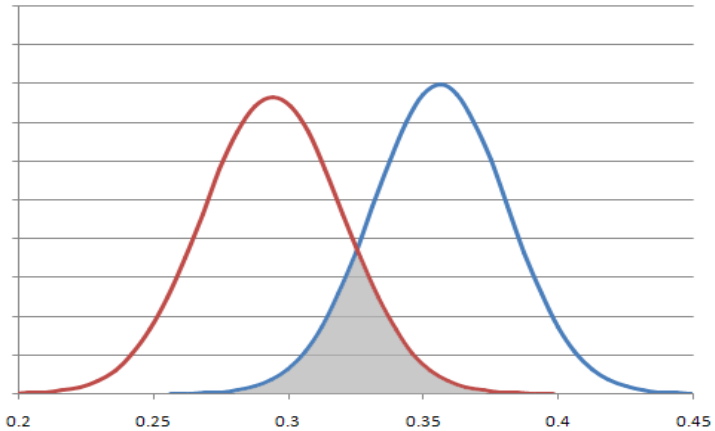
$$L^* = \mathbb{P}\{g^*(X) \neq Y\} = 1 - \Phi\left(\frac{\parallel m \parallel_K}{2}\right)$$

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

The singular case

"We argue that those [functional classification] problems have unusual, and fascinating, properties that set them apart from their finite dimensional counterparts. In particular we show that, in many quite standard settings, the performance of simple [linear] classifiers constructed from training samples becomes perfect as the sizes of those samples diverge [...]. That property never holds for finite dimensional data, except in pathological cases." Delaigle and Hall, J. R. Statist. Soc. B 2012



The model

$$\begin{cases} P_0: & X(t) = & \xi(t), & t \in [0,1] \\ P_1: & X(t) = & m(t) + \xi(t), & t \in [0,1] \end{cases}$$

$\xi(t)$ gaussian with $\mathbb{E}(\xi(t)) = 0$.

$$K(s,t) = \mathbb{E}(\xi(s)\xi(t)) = \sum_{j=1}^{\infty} \theta_j \phi_j(s)\phi_j(t).$$

Where $\theta_1 \geq \theta_2 \geq \ldots$ and K is strictly positive definite and uniformly bounded.

$$m(t) = \sum_{j=1}^{\infty} \mu_j \phi_j.$$

Prior probabilities: $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$.

Results

Theorem 1 (Delaigle and Hall, J. R. Statist. Soc. B 2012)

(a) When $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 < \infty$ the Bayes (minimal) error is $err_0 = 1 - \Phi\left(\frac{1}{2}(\sum_{j\geq 1}\theta_j^{-1}\mu_j^2)^{1/2}\right) > 0$ and the optimal classifier (that achieves this error) is the rule

$$T^0(X) = 1, \text{ if and only if } (\langle X,\psi\rangle_{L^2} - \langle m,\psi\rangle_{L^2})^2 - \langle X,\psi\rangle_{L^2}^2 < 0,$$

with $\psi(t) = \sum_{j=1}^{\infty}\theta_j^{-1}\mu_j\phi_j(t)$.

(b) If $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 = \infty$ then the minimal misclassification probability is $err_0 = 0$ and it is achieved, in the limit, by a sequence of classifiers constructed from T^0 by replacing the function ψ with $\psi^{(r)} = \sum_{j=1}^{r}\theta_j^{-1}\mu_j\phi_j(t)$, with $r = r_n \uparrow \infty$.

An unanswered question:

Why?

"The theoretical foundation for these findings is an intriguing dichotomy of properties and is as interesting as the findings themselves." Delaigle and Hall, 2012

An unanswered question:

Why?

"The theoretical foundation for these findings is an intriguing dichotomy of properties and is as interesting as the findings themselves." Delaigle and Hall, 2012

Because of the singularity

Our view of the "near perfect classification"

Theorem 2.4

(a) $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 < \infty$ if and only if $P_1 \sim P_0$. In that case, the Bayes rule g^* is

$$g^*(X) = 1 \text{ if and only if } \langle X, m\rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

This is a coordinate-free, equivalent expression of the optimal rule given by D. & H. The corresponding optimal (Bayes) classification error is $L^* = 1 - \Phi\left(\parallel m \parallel_{\mathcal{H}_K} /2\right)$.

(b) $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 = \infty$ if and only if $P_1 \perp P_0$. In this case the Bayes error is $L^* = 0$. Moreover, for any $\epsilon > 0$ we can construct a classification rule whose misclassification probability is smaller than ϵ (Theorem 2.5).

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

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.

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.

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

The Bayes rule under the sparsity assumption

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

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

The Bayes rule under the sparsity assumption

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

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

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 optimal rule coincides with the well-known Fisher linear rule based on the projections $x(t_1^*), \ldots, x(t_d^*)$.

RKHS-based variable selection

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

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

RKHS-based variable selection

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

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

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

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

RKHS-based variable selection

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

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

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

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

In practice we use a "greedy" algorithm to select the points.

An associated classifier

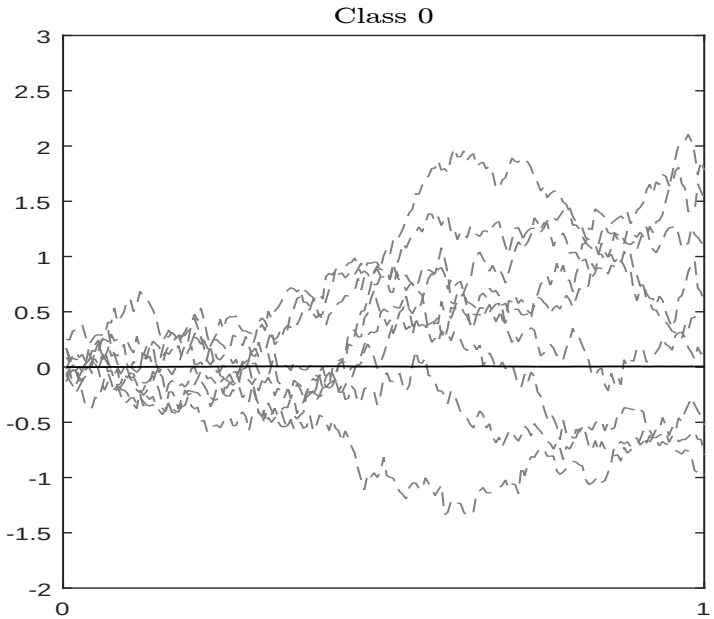
Theorem 2.6 (consistency)

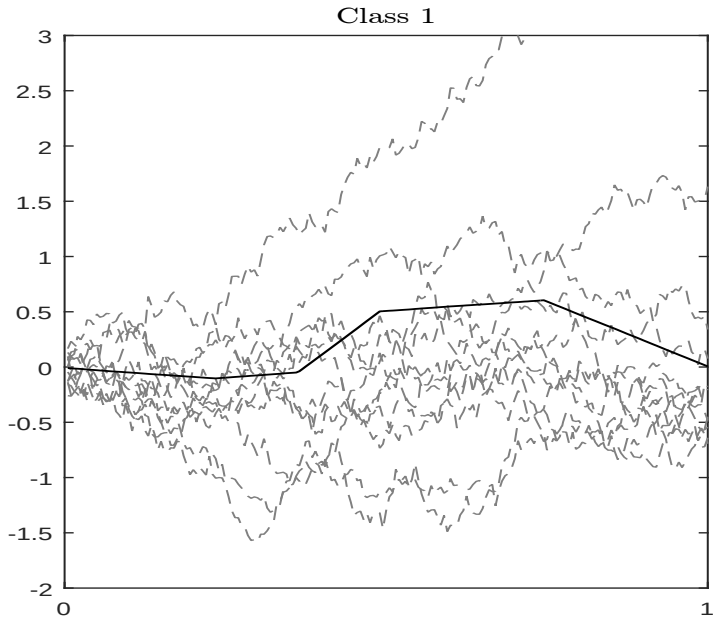
Let us consider the framework and conditions of our first theorem [i.e. the expression of the optimal rule for the abs. continuous case] and assume further that [SA] holds. Let $L^* = \mathbb{P}(g^*(X) \neq Y)$ the optimal misclassification probability corresponding to the Bayes rule. Denote by $L_n = \mathbb{P}(\hat{g}(X) \neq Y | X_1, \ldots, X_n)$ the misclassification probabilities of the estimated rules defined above (under the [SA] assumption of order n). Then, $L_n \to L^*$ a.s., as $n \to \infty$.

Some observations

- Two for one: variable selection method and classifier.
- Theoretically motivated (consistent).
- Greedy algorithm: no guarantee of convergence but affordable.
- It shows good performance in practice.
- Robust: the empirical results show also a remarkable robustness of the RK methodology against departures from the assumptions on which it is based.
- Flexible: additional information can be incorporated easily.

An example



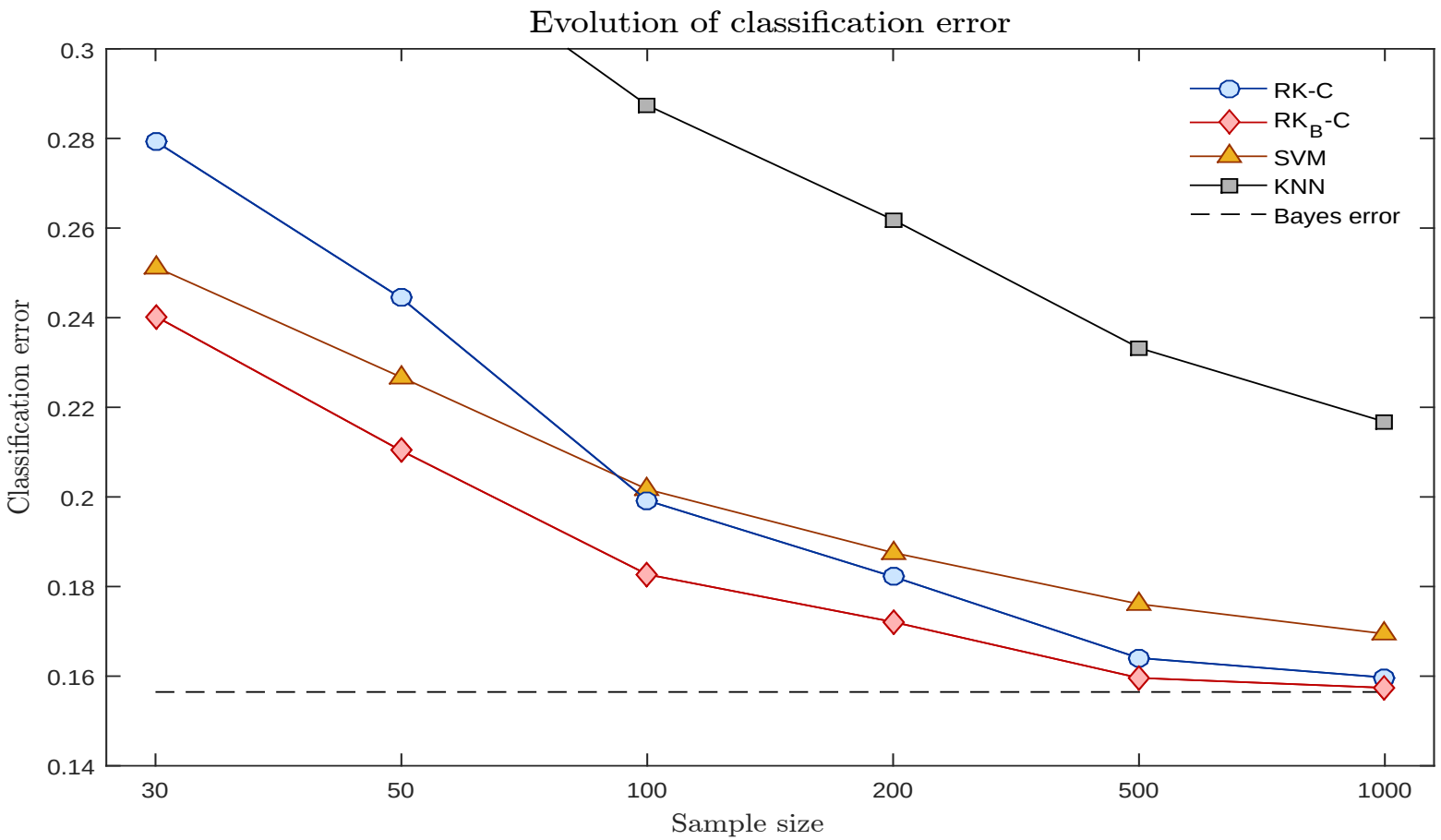


$K(s,t) = \min\{s,t\}$.

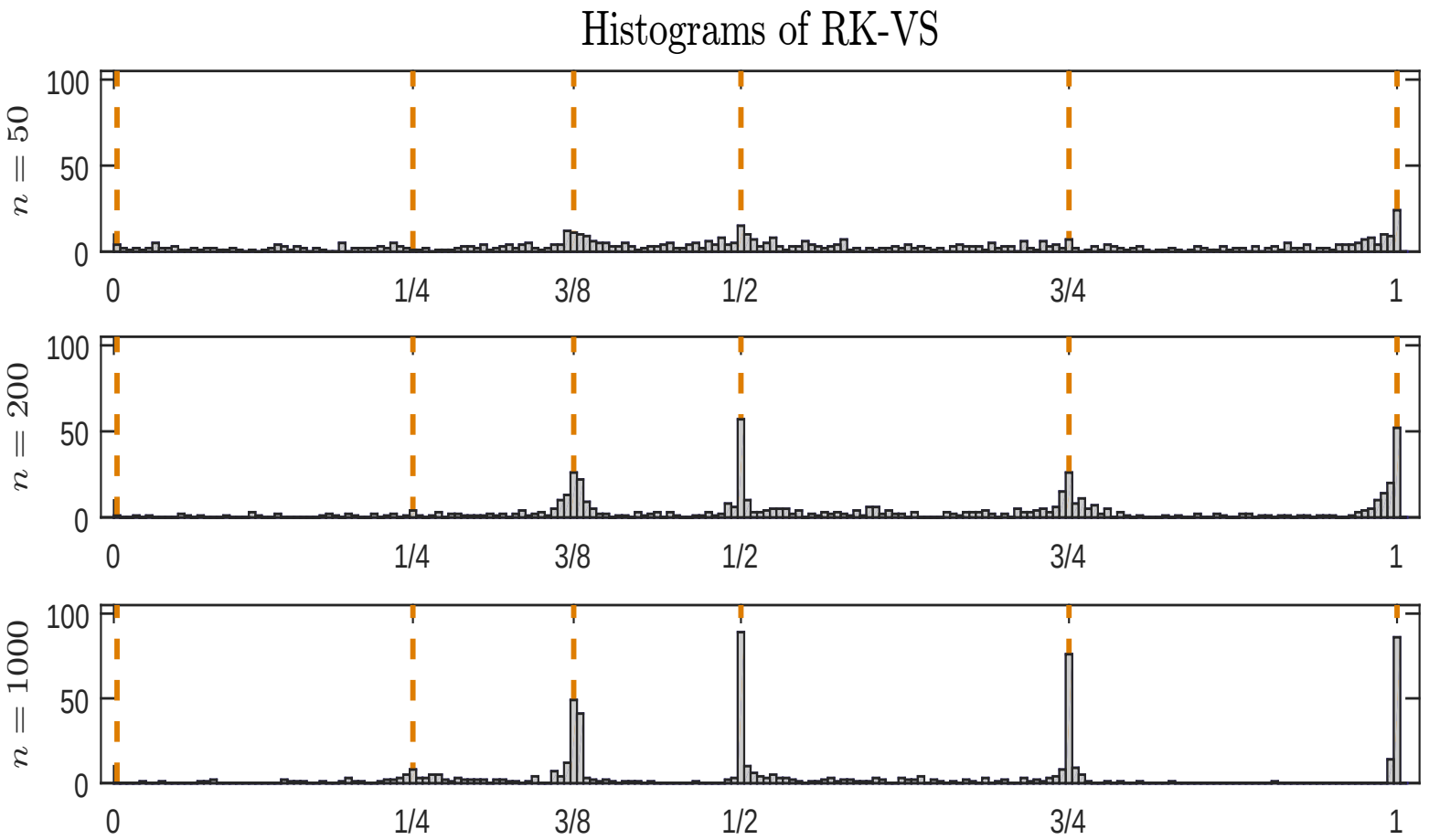
$t^* = \{0, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1\}$.

$L^* = 0.1587$.

An example (II)



An example (III)



Some results

Table: Misclassification percentages (and standard deviations) for the classification methods considered in Table 2 of Delaigle and Hall (2012) and the new RK-C method

Data	n	Classification rules				
		CENT_{PC1}	CENT_{PLS}	NP	CENT_{PCp}	**RK-C**
Wheat	30	0.89 (2.49)	0.46 (1.24)	0.49 (1.29)	15.0 (1.25)	**0.25 (1.58)**
	50	0.22 (1.09)	0.06 (0.63)	0.01 (0.14)	14.4 (5.52)	**0.02 (0.28)**
Phoneme	30	22.5 (3.59)	24.2 (5.37)	24.4 (5.31)	23.7 (2.37)	**22.5 (3.70)**
	50	20.8 (2.08)	21.5 (3.02)	21.9 (2.91)	23.4 (1.80)	**21.5 (2.36)**
	100	20.0 (1.09)	20.1 (1.12)	20.1 (1.37)	23.4 (1.36)	**20.1 (1.25)**

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Our second proposal

To use the minimun Redundancy Maximun Relevance method replacing de Mutual Information discrepancy with the Distance Correlation. mRMR is a contrasted filter method of variable selection proposed by Ding and Peng (2005), Peng et al. (2005) .

Berrendero, Cuevas and Torrecilla. The mRMR variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation* (to appear)

The mRMR algorithm

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

The mRMR 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 set 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 mRMR 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 set 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 objective is to choose the set S which maximizes (greedy)

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

Original mRMR relevance measure: MI

Mutual Information

General statistical independence measure between two random variables. It takes care of nonlinear dependences.
Let two continuous random variables X and Y, their marginal density functions $p(X)$ and $p(Y)$, and their joint density function $p(X, Y)$, mutual information is defined by

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

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

The new proposal

Distance correlation $\mathcal{R}$

- Distance correlation is a measure of dependence between random vectors proposed in Székely, Rizzo and Bakirov, *Ann Stat* (2007) and Székely, Rizzo, (2009, 2012, 2013).
- For all distributions with finite first moments, $\mathcal{R}$ generalizes the idea of correlation in two fundamental ways:
  - $\mathcal{R}(X, Y)$ is defined for X and Y in arbitrary dimensions.
  - $\mathcal{R}(X, Y) = 0$ characterizes independence of X and Y.
- It can be estimated without tuning parameters or smoothing.

Experiments

Measures under comparison

- Distance covariance (V)
- Distance correlation (R)
- Mutual information (MI)
- Fisher-Correlation criterion (FC)
- Standard correlation (C)

Classifiers

- k nearest neighbours (k-NN)
- Linear discriminant analysis (LDA)
- Naive Bayes (NB)
- Linear support vector machine (SVM)

`www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx`

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

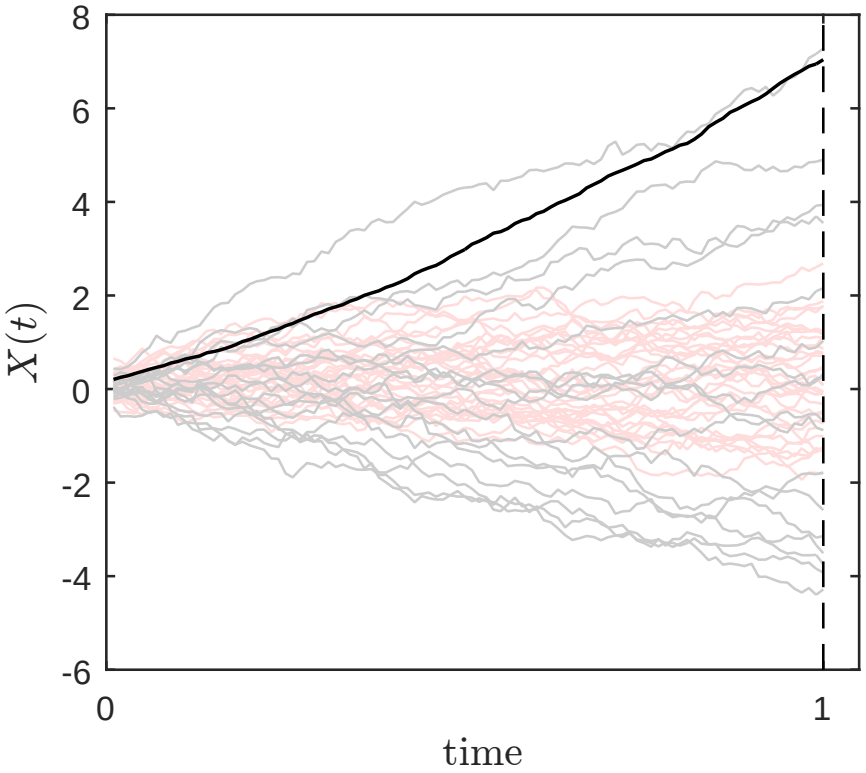
Our third proposal

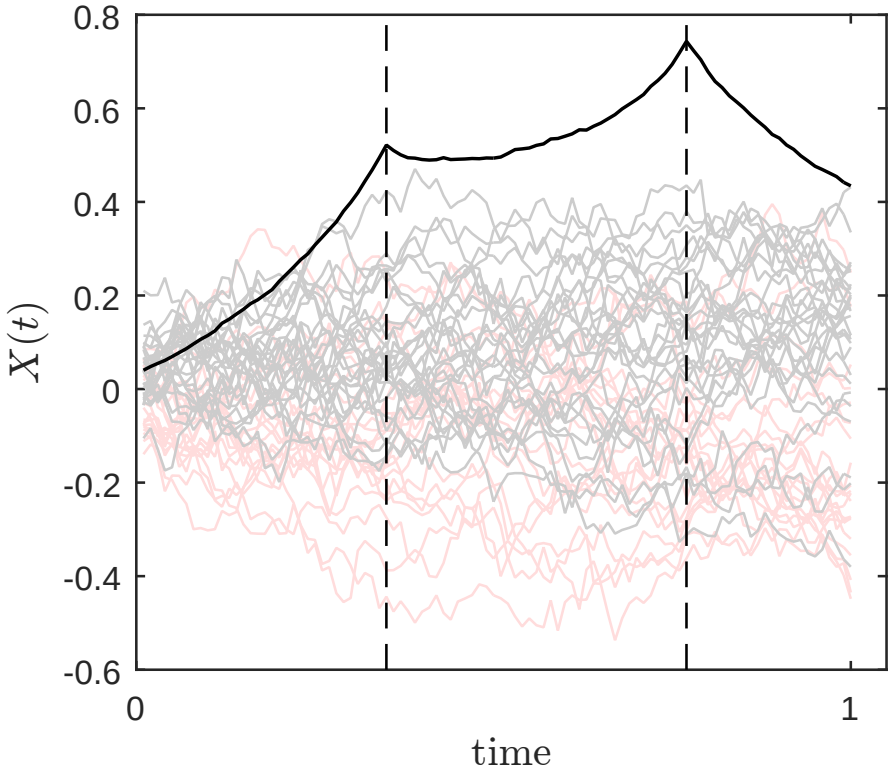
Maxima-Hunting criterion.

To select the points $t_1, \cdots, t_k$ in according to the local maxima of the Distance Correlation $\mathcal{R}^2(X(t), Y)$ (alternatively, the local maxima of the Distance Covariance $\mathcal{V}^2(X(t), Y)$).

Berrendero, Cuevas and Torrecilla. Variable selection in functional data classification: a maxima-hunting proposal. Statistica Sinica (to appear)

Some examples





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 in $\mathcal{V}_n^2(X(t), Y)$.
- The empirical results show a remarkable good performance of MH methods in comparison with other state-of-art alternatives.
  `www.uam.es/antonio.cuevas/exp/outputs.xlsx`
- It is also theoretically supported.

Theoretical results

- Some equivalent expressions for $\mathcal{V}^2(X(t), Y)$ in the binary case (Thm. 3.1).
- Several non-trivial examples where the relevant information is concentrated on the maxima of $\mathcal{V}^2(X(t), Y)$ (Props. 3.4, 3.5).

Theorem 3.2 (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 m local maxima at $t_1, \cdots, t_m$, then $\mathcal{V}_n^2(X_t, Y)$ has also eventually at least m maxima at $t_{1n}, \cdots, t_{mn}$ with $t_{jn} \to t_j$, as $n \to \infty$, a.s., for $j = 1, \ldots, m$.

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Empirical study

We have carried out exhaustive and reproducible experiments in order to assess the performance of our variable selection methods.

- 8 dimension reduction methods (with some variants) and three benchmark procedures.
- 4 different classifiers.
- Data
  - 100 simulation models with (4 different sample sizes).
  - 4 real data sets.
  - A real biomedical application.

    Barba et al. High fat diet and female sex induce metabolic changes and reduce oxidative stress in an additive maner in mice heart. Submitted
- Parameters are chosen by standard validation procedures.

`www.uam.es/antonio.cuevas/exp/outputs.xlsx`
`www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx`

And the winner is...

- There is no uniform winner. Different approaches, different targets.
- Good performance of the new proposals. PLS is the first competitor (not interpretable).
- On average MHR and RK-VS are better (encouraging).
- Stable results in different models and different classifiers.

Some recommendations

- Use mRMR-RD instead of othe mRMR formulations.
- Use RK-VS when the required assumptions are partially fulfilled.
- Use MHR when we are far from RK-VS hypotheses or very small sample sizes.

Outline

1. Introduction
   - FDA
   - Variable Selection
   - Functional classification
2. RKHS
   - The RKHS approach
   - The absolutely continuous case
   - The singular case
3. Variable selection
   - Variable selection and RKHS
   - mRMR-RD
   - Maxima hunting
4. Experiments
5. Conclusions and future work

Summary

a) General mathematical theory for the functional classification problem (RKHS associated with the covariance operator of the processes).
   a1) Explicit expressions for the Bayes (optimal) rule and error for the case of absolutely continuous Gaussian processes.
   a2) A complete mathematical treatement for the classification problem between to mutually singular processes (near perfect classification).

b) Functional variable selection.
   b1) A general theoretical motivation (expressed in terms of a sparsity assumption) for the problems of functional variable selection.
   b2) Three new variable selection methods: *RK-VS* (an RKHS-based selector), *MH* (a "maxima-hunting" method) and *mRMR-RD* (a modified version of the popular mRMR procedure).

c) Numerical experiments. We provide the largest simulation study on functional variable selection we are aware of. Some popular data examples are also analysed together with a further real example with metabolic data.

General conclusions

- Variable selection is extremely useful in terms of statistical efficiency in FDA. In our experience we have not found any reason against the use of variable selection in functional classification.
- Variable selection entails a gain in interpretability compared with other popular dimension reduction methods (e.g., PCA and PLS).
- Our new proposals are competitive. MH and RK-VS are theoretically motivated and easy to interpret. mRMR-RD also leads to an improvement in accuracy with respect to the original mRMR formulations.

General conclusions

A major aim in this thesis was to contribute to the mathematical foundation of FDA as a statistical counterpart for the stochastic processes theory.

- The use of RKHS theory provides a convincing model for variable selection and calculation of RN-derivatives.
- The RN derivatives can be successfully used to define new plug-in classifiers. The expressions of many RN derivatives are not so difficult to handle.
- RKHS appears as an appealing alternative to the classical L^2 setup for some problems. As a consequence, the near-perfect classification phenomenon can be explained in terms of the singularity of the measures.

Future work

- Extension of our results in functional classification to different settings: non Gaussian, non homoscedastic, multiclass...
- Explore the potential of application of the RKHS theory in FDA (regression, clustering, visualization...). d is still an open problem.
- Open problems in our variable selection methods: non-differentiable points in $\mathcal{R}^2(X(t), Y)$, "parametric" variable selection, mRMR theory, two-stage algorithms...
- Further applications of the distance correlation.
- Real applications.
- R package or MATLAB toolbox.

Thank you

joseluis.torrecilla@uam.es