



FACULTAD DE CIENCIAS  
DEPARTAMENTO DE MATEMÁTICAS

# On the Theory and Practice of Variable Selection for Functional Data

José Luis Torrecilla Noguerales

A dissertation submitted in partial fulfillment of the requirements for the Degree  
of Doctor of Mathematics

Under the supervision of  
José Ramón Berrendero Díaz and Antonio Cuevas González

Madrid, 2015



# Agradecimientos

Este trabajo no sería lo que es, y quizá ni siquiera sería, sin la participación de mucha gente que ha contribuido de alguna manera, incluso sin saberlo, a que esta tesis cobrase forma. A todas estas personas mi gratitud.

En primer lugar quiero expresar mi más sincero agradecimiento a mis directores José Ramón y Antonio, sin los cuales esta tesis no existiría. Gracias por darme la oportunidad de comenzar este proyecto y por haberme acompañado y apoyado en cada etapa. Gracias también por vuestra confianza, motivación y disponibilidad, sobre todo en los momentos complicados. Me siento muy afortunado de haberos tenido como tutores, de vosotros he aprendido mucho más que estadística.

Este agradecimiento es extensivo a Tati, Begoña, Raimundo, Mari Paz, Emilio, José Ramón y a todos los profesores que tanto me han enseñado en mi paso por el C.P. Badiel, el I.E.S. Liceo Caracense y la Universidad Autónoma de Madrid. Y también a Luisfer y a Samuel que, aunque nunca me evaluaron, me enseñaron valiosas lecciones y confiaron en mí para empezar a trabajar con *chavales*.

Mi agradecimiento a Ignasi Barba y David García-Dorado. Algunos de los resultados de esta tesis son fruto de la colaboración con su grupo de investigación en el Hospital Universitari Vall d'Hebron. También quiero agradecerle a Jane-Ling Wang la oportunidad de trabajar con ella y su grupo en la UC Davis, y a Christina el hacer de Davis un lugar tan acogedor.

Durante todos estos años en la UAM he tenido la suerte de contar con muchos buenos compañeros con los que discutir de matemáticas, compartir las dificultades del doctorado, o simplemente pasar un buen rato y recobrar el ánimo para enfrentarme a la tesis. Gracias a todos. En este punto me gustaría agradecerles expresamente a Alberto, Alessandro y Bego el buen ambiente que ha habido en el despacho y su ayuda, a Rocío todas las conversaciones en esas largas horas de tren y a Ángela, Dani, Gema, Jaime, José, Mariaje, Raúl, Sergio y Sofía, el apoyo y el compañerismo en distintos momentos. Y un agradecimiento especial para Carlos, que *me aguantó* durante toda la carrera con infinita paciencia, y para Carlos M.,

que le tomó el relevo en el máster y que, junto con Ángela, han sido importantes asesores sobre algunos aspectos de este trabajo.

No puedo olvidarme del Departamento de Matemáticas, el Instituto de Ingeniería del Conocimiento y al programa FPI del MICINN, que pusieron los medios para que pudiera llevar a cabo mi investigación.

También quiero dar las gracias a mis amigos de Guadalajara y a la gente del Centro Juvenil que han seguido la evolución de esta tesis, aun no entendiendo la mayor parte de lo que les contaba, y han soportado mis altibajos.

Gracias a Iris por apoyarme incondicionalmente y confiar tanto en mí, creyendo muchas veces más en mí que yo mismo. Gracias por conseguir sacarme una sonrisa incluso en los peores días y por escuchar una y mil veces todas esas *cosas incomprensibles*. Espero que la selección de variables o la clasificación funcional ya no sean ideas tan extrañas.

Quiero acabar dando las gracias a mi familia, a mis abuelos, a mi hermano Miguel Ángel (corrector de estilo ocasional) y especialmente a mis padres Natividad y Víctor, a quienes esta tesis está dedicada. Gracias por vuestro amor y sacrificios, sin vuestro apoyo nada de esto hubiera sido posible.

# Abstract

Functional Data Analysis (FDA) might be seen as a partial aspect of the modern mainstream paradigm generally known as Big Data Analysis. The study of functional data requires new methodologies that take into account their special features (e.g. infinite dimension and high level of redundancy). Hence, the use of variable selection methods appears as a particularly appealing choice in this context. Throughout this work, variable selection is considered in the setting of supervised binary classification with functional data  $\{X(t), t \in [0, 1]\}$ . By variable selection we mean any dimension-reduction method which leads to replace the whole trajectory  $\{X(t), t \in [0, 1]\}$ , with a low-dimensional vector  $(X(t_1), \dots, X(t_d))$  still keeping a similar classification error. In this thesis we have addressed the “functional variable selection” in classification problems from both theoretical and empirical perspectives.

We first restrict ourselves to the standard situation in which our functional data are generated from Gaussian processes, with distributions  $P_0$  and  $P_1$  in both populations under study. The classical Hajek-Feldman dichotomy establishes that  $P_0$  and  $P_1$  are either mutually absolutely continuous with respect to each other (so there is a Radon-Nikodym (RN) density for each measure with respect to the other one) or mutually singular. Unlike the case of finite dimensional Gaussian measures, there are non-trivial examples of mutually singular distributions when dealing with Gaussian stochastic processes. This work provides explicit expressions for the optimal (Bayes) rule in several relevant problems of supervised binary (functional) classification under the absolutely continuous case. Our approach relies on some classical results in the theory of stochastic processes where the so-called Reproducing Kernel Hilbert Spaces (RKHS) play a special role. This RKHS framework allows us also to give an interpretation, in terms of mutual singularity, for the “near perfect classification” phenomenon described by [Delaigle and Hall \(2012a\)](#). We show that the asymptotically optimal rule proposed by these authors can be identified with the sequence of optimal rules for an approximating sequence of classification problems in the absolutely continuous case.

The methodological contributions of this thesis are centred in three variable selection methods. The obvious general criterion for variable selection is to choose the “most representative” or “most relevant” variables. However, it is also clear that a purely relevance-oriented criterion could lead to select many redundant variables. First, we provide a new model-based method for variable selection in binary classification problems, which arises in a very natural way from the explicit knowledge of the RN-derivatives and the underlying RKHS structure. As a consequence, the optimal classifier in a wide class of functional classification problems can be expressed in terms of a classical, linear finite-dimensional Fisher rule.

Our second proposal for variable selection is based on the idea of selecting the local maxima  $(t_1, \dots, t_d)$  of the function  $\mathcal{V}_X^2(t) = \mathcal{V}^2(X(t), Y)$ , where  $\mathcal{V}$  denotes the *distance covariance*

association measure for random variables due to Székely et al. (2007). This method provides a simple natural way to deal with the relevance vs. redundancy trade-off which typically appears in variable selection. This proposal is backed by a result of consistent estimation for the maxima of  $\mathcal{V}_X^2$ . We also show different models for the underlying process  $X(t)$  under which the relevant information is concentrated on the maxima of  $\mathcal{V}_X^2$ .

Our third proposal for variable selection consists of a new version of the minimum Redundancy Maximum Relevance (mRMR) procedure proposed by Ding and Peng (2005) and Peng et al. (2005). It is an algorithm to systematically perform variable selection, achieving a reasonable trade-off between relevance and redundancy. In its original form, this procedure is based on the use of the so-called *mutual information criterion* to assess relevance and redundancy. Keeping the focus on functional data problems, we propose here a modified version of the mRMR method, obtained by replacing the mutual information by the new *distance correlation* measure in the general implementation of this method.

The performance of the new proposals is assessed through an extensive empirical study, including about 400 simulated models (100 functional models  $\times$  4 sample sizes) and real data examples, aimed at comparing our variable selection methods with other standard procedures for dimension reduction. The comparison involves different classifiers. A real problem with biomedical data is also analysed in collaboration with researchers of Hospital Vall d'Hebron (Barcelona). The overall conclusions of the empirical experiments are quite positive in favour of the proposed methodologies.

# Resumen

El Análisis de Datos Funcionales (FDA por sus siglas en inglés) puede ser visto como una de las facetas del paradigma general conocido como *Big Data Analysis*. El estudio de los datos funcionales requiere la utilización de nuevas metodologías que tengan en cuenta las características especiales de estos datos (por ejemplo, la dimensión infinita y la elevada redundancia). En este contexto, las técnicas de selección de variables parecen particularmente atractivas. A lo largo de este trabajo, estudiaremos la selección de variables dentro del marco de la clasificación supervisada binaria con datos funcionales  $\{X(t), t \in [0, 1]\}$ . Por selección de variables entendemos cualquier método de reducción de dimensión enfocado a reemplazar las trayectorias completas  $\{X(t), t \in [0, 1]\}$  por vectores de baja dimensión  $(X(t_1), \dots, X(t_d))$  conservando la información discriminante. En esta tesis hemos abordado la “selección de variables funcional” en problemas de clasificación tanto en su vertiente teórica como empírica.

Nos restringiremos esencialmente al caso general en que los datos funcionales están generados por procesos Gaussianos, con distribuciones  $P_0$  y  $P_1$  en las distintas poblaciones. La dicotomía de Hajek-Feldman establece que  $P_0$  y  $P_1$  sólo pueden ser mutuamente absolutamente continuas (existiendo entonces una densidad de Radon-Nikodym (RN) de cada medida con respecto a la otra) o mutuamente singulares. A diferencia del caso finito dimensional, cuando trabajamos con procesos Gaussianos aparecen ejemplos no triviales de distribuciones mutuamente singulares. En este trabajo se dan expresiones explícitas de la regla de clasificación óptima (Bayes) para algunos problemas funcionales binarios relevantes en el contexto absolutamente continuo. Nuestro enfoque se basa en algunos resultados clásicos de la teoría de procesos estocásticos, entre los que los Espacios de Hilbert de Núcleos Reproductores (RKHS) desempeñan un papel fundamental. Este marco RKHS nos permite también dar una interpretación del fenómeno de la “clasificación casi perfecta” descrito por [Delaique and Hall \(2012a\)](#), en términos de la singularidad mutua de las distribuciones.

Las contribuciones metodológicas de esta tesis se centran en tres métodos de selección de variables. El criterio más obvio para seleccionar las variables sería elegir aquellas “más representativas” o “más relevantes”. Sin embargo, un criterio basado únicamente en la relevancia probablemente conduciría a la selección de muchas variables redundantes. En primer lugar, proponemos un nuevo método de selección de variables basado en modelo, que surge de manera natural del conocimiento de las derivadas RN y de la estructura RKHS subyacente. Como consecuencia, el clasificador óptimo para una amplia clase de problemas de clasificación funcional puede expresarse en términos de la regla lineal de Fisher finito dimensional.

Nuestra segunda propuesta para selección de variables se basa en la idea de seleccionar los máximos locales  $(t_1, \dots, t_d)$  de la función  $\mathcal{V}_X^2(t) = \mathcal{V}^2(X(t), Y)$ , donde  $\mathcal{V}$  denota la *covarianza*

*de distancias*, medida de asociación entre variables aleatorias propuesta por Székely et al. (2007). Este procedimiento se ocupa de manera natural del equilibrio entre relevancia y redundancia típico de la selección de variables. Esta propuesta está respaldada por un resultado de consistencia en la estimación de los máximos de  $\mathcal{V}_X^2$ . Además, se muestran distintos modelos de procesos subyacentes  $X(t)$  para los que la información relevante se concentra en los máximos de  $\mathcal{V}_X^2$ .

La tercera propuesta para seleccionar variables es una nueva versión del método mRMR (mínima Redundancia Máxima Relevancia), propuesto en Ding and Peng (2005) y Peng et al. (2005). Este algoritmo realiza una selección de variables sistemática, consiguiendo un equilibrio relevancia-redundancia razonable. El procedimiento mRMR original se basa en la utilización de la *información mutua* para medir la relevancia y la redundancia. Manteniendo el problema funcional como referencia, se propone una nueva versión de mRMR en la que la información mutua es remplazada por la nueva *correlación de distancias*.

El rendimiento de las nuevas propuestas es evaluado mediante extensos estudios empíricos con el objetivo de comparar nuestros métodos de selección de variables con otros procedimientos de reducción de dimensión ya establecidos. Los experimentos incluyen 400 modelos de simulación (100 modelos funcionales  $\times$  4 tamaños muestrales) y ejemplos con datos reales. La comparativa incluye distintos clasificadores. Además se ha analizado un problema real con datos biomédicos en colaboración con investigadores del Hospital Vall d'Hebron (Barcelona). Los resultados del estudio son, en general, bastante positivos para los nuevos métodos.

# Contents

<b>Agradecimientos</b>	<b>I</b>
<b>Abstract</b>	<b>III</b>
<b>Resumen</b>	<b>V</b>
<b>Table of contents</b>	<b>IX</b>
<b>Lists</b>	<b>XI</b>
<b>Some notation</b>	<b>XV</b>
<b>1. Introduction</b>	<b>1</b>
1.1. Functional Data Analysis . . . . .	1
1.1.1. Some basic notions and difficulties in FDA . . . . .	3
1.2. Supervised Classification . . . . .	6
1.2.1. Classification with functional data . . . . .	7
1.2.2. Supervised classification and absolute continuity . . . . .	10
1.3. Functional data representation . . . . .	11
1.3.1. Smoothing and basis representation . . . . .	11
1.3.2. Other issues . . . . .	15
1.4. Variable selection . . . . .	16
1.4.1. Motivation . . . . .	16
1.4.2. Some general terminology and references on dimension reduction methods . . . . .	19
1.4.3. Functional variable selection methods . . . . .	25
1.5. Contributions and structure of the thesis . . . . .	28
1.5.1. Contributions . . . . .	29
1.5.2. Structure . . . . .	32

<b>2. RKHS-based functional classification</b>	<b>35</b>
2.1. Radon-Nikodym densities for Gaussian processes . . . . .	36
2.1.1. RKHS . . . . .	37
2.1.2. RKHS and Radon-Nikodym derivatives. Parzen's Theorem	38
2.2. Absolutely continuous Gaussian processes . . . . .	39
2.3. Singular Gaussian processes . . . . .	40
2.4. An RKHS-based proposal . . . . .	42
2.4.1. RKHS and variable selection . . . . .	42
2.4.2. An RKHS-based criterion for variable selection and its associated Fisher rule . . . . .	44
2.4.3. Practical issues . . . . .	46
2.5. Experiments . . . . .	50
2.5.1. Methodology . . . . .	51
2.5.2. Simulation outputs . . . . .	52
2.5.3. Real data . . . . .	55
2.6. Conclusions . . . . .	56
2.7. Proofs . . . . .	57
<b>3. Maxima-Hunting</b>	<b>63</b>
3.1. An auxiliary tool: the distance covariance . . . . .	64
3.2. Variable selection based on maxima hunting . . . . .	66
3.3. Theoretical motivation . . . . .	69
3.4. Empirical study . . . . .	74
3.4.1. The variable selection methods under study. Criteria for comparisons . . . . .	74
3.4.2. The simulation study . . . . .	76
3.4.3. Real data examples . . . . .	78
3.5. Overall conclusions . . . . .	82
3.6. Some additional results and proofs . . . . .	85
<b>4. mRMR</b>	<b>95</b>
4.1. The mRMR criterion . . . . .	96
4.1.1. Association measures . . . . .	96
4.1.2. Methodology . . . . .	100
4.2. The empirical study . . . . .	102
4.2.1. A few numerical outputs from the simulations . . . . .	103
4.2.2. Ranking the methods . . . . .	106
4.2.3. Real data examples . . . . .	107
4.3. A real application: NMR spectral fingerprints . . . . .	112
4.4. Final conclusions and comments . . . . .	116

<b>5. On the empirical studies</b>	<b>119</b>
5.1. Methods and implementation . . . . .	119
5.1.1. Dimension reduction methods . . . . .	120
5.1.2. Classifiers . . . . .	122
5.1.3. Computational details . . . . .	124
5.2. Simulations . . . . .	126
5.2.1. Models . . . . .	126
5.2.2. Methodology . . . . .	127
5.2.3. Additional results . . . . .	129
5.3. Real data . . . . .	133
5.3.1. Data sets . . . . .	133
5.3.2. Methodology . . . . .	135
5.3.3. Additional results . . . . .	137
<b>6. Conclusions</b>	<b>141</b>
6.1. Further work . . . . .	144
<b>6. Conclusiones</b>	<b>147</b>
6.1. Trabajo futuro . . . . .	151
<b>A. Simulation models</b>	<b>153</b>



# Lists

## Figures

1.1. Examples of stochastic processes . . . . .	4
1.2. B-spline representation: number of elements . . . . .	13
1.3. Basis representation example . . . . .	14
1.4. Example of derivatives . . . . .	17
1.5. Variable selection algorithms by evaluation criterion . . . . .	20
1.6. Univariate vs. multivariate selection . . . . .	24
2.1. Motivating example: trajectories . . . . .	47
2.2. Motivating example: evolution of RK-C error (sample size) . . . . .	48
2.3. Motivating example: evolution of RK-C error (number of variables) . . . . .	49
2.4. Motivating example: first selected variable by RK-VS . . . . .	49
2.5. Motivating example: first selected variable by $RK_B$ -VS . . . . .	50
3.1. Examples of distance covariance functions . . . . .	67
3.2. Criteria for sorting local maxima . . . . .	75
3.3. Ranking of methods with different classifiers . . . . .	84
4.1. Ranking of mRMR criteria with different classifiers . . . . .	109
4.2. NMR spectra: trajectories . . . . .	113
4.3. NMR spectra: 2D projection . . . . .	116
5.1. Methodology flowchart for simulations . . . . .	129
5.2. Real data trajectories . . . . .	133
5.3. Methodology flowchart for real data . . . . .	137

## Tables

2.1. RK-VS simulation outputs: classification accuracy . . . . .	52
--	----

2.2.	RK-VS simulation outputs: number of variables . . . . .	53
2.3.	RK-C: classification accuracy over all simulations . . . . .	54
2.4.	RK-C: classification accuracy over several special models . . . . .	54
2.5.	RK-C: misclassification percentages over two real data sets . . . . .	56
3.1.	MH: simulation outputs with $k$ -NN . . . . .	77
3.2.	MH: simulation outputs with LDA . . . . .	78
3.3.	MH: classification accuracy of several simulation models . . . . .	79
3.4.	MH: classification accuracy with real data . . . . .	80
3.5.	MH: number of variables with real data . . . . .	81
3.6.	MH: ranking of methods . . . . .	83
4.1.	mRMR: Simulation outputs with NB . . . . .	104
4.2.	mRMR: Simulation outputs with $k$ -NN . . . . .	105
4.3.	mRMR: Simulation outputs with LDA . . . . .	105
4.4.	mRMR: Simulation outputs with SVM . . . . .	106
4.5.	mRMR: ranking of methods with NB . . . . .	107
4.6.	mRMR: ranking of methods with $k$ -NN . . . . .	108
4.7.	mRMR: ranking of methods with LDA . . . . .	108
4.8.	mRMR: ranking of methods with SVM . . . . .	110
4.9.	mRMR: real data outputs . . . . .	111
4.10.	NMR: Classification matrices with a 3-NN classifier . . . . .	114
4.11.	NMR: Classification matrices with mRMR-RD + LDA . . . . .	115
4.12.	NMR: Classification matrices with PLS+LDA . . . . .	115
5.1.	Simulations summary: classification accuracy . . . . .	130
5.2.	Simulations summary: number of variables . . . . .	131
5.3.	Description of real data sets . . . . .	134
5.4.	Real data summary: classification accuracy . . . . .	138
5.5.	Real data summary: number of variables . . . . .	139

## Theorems and others

2.1.	Theorem (Parzen 1961, Thm. 7A) . . . . .	38
2.2.	Theorem (Bayes Rule for homoscedastic Gaussian problems) . . . . .	39
2.3.	Theorem (Delaique and Hall 2012a, Thm.1) . . . . .	41
2.4.	Theorem (Another view on near perfect classification) . . . . .	42
2.5.	Theorem (Singular case classifier) . . . . .	42
2.1.	Remark (Sparsity example) . . . . .	44
2.6.	Theorem (Consistency of the RKHS-based classifier) . . . . .	45

3.1. Definition (Distance covariance) . . . . .	64
3.2. Definition (Estimator of $\mathcal{V}^2$ ) . . . . .	65
3.1. Theorem (Expressions for $\mathcal{V}^2$ ) . . . . .	68
3.2. Theorem (Uniform convergence of $\tilde{\mathcal{V}}_n^2$ ) . . . . .	69
3.1. Lemma (Asymptotic equivalence of estimators) . . . . .	69
3.1. Proposition (Bayes rule stochastic trend) . . . . .	71
3.2. Proposition (Bayes rule linear trend) . . . . .	71
3.3. Proposition (Bayes rule “peak” trend) . . . . .	71
3.3. Theorem (Bayes rule under heteroscedasticity) . . . . .	72
3.1. Remark (Additional examples) . . . . .	72
3.4. Proposition (Maxima of $\mathcal{V}^2$ ) . . . . .	73
3.2. Remark (Other examples) . . . . .	73
3.5. Proposition (Global maximum of $\mathcal{V}^2$ ) . . . . .	73
3.2. Lemma (Uniform convergence of one-sample U-statistics) . . . . .	85
3.3. Lemma (Uniform convergence of two-sample U-statistics) . . . . .	87
3.4. Theorem (Shepp 1966, Thm. 1) . . . . .	91



## Some notation

Throughout this thesis, we will denote the whole stochastic process by  $X$ . Provided that no confusion is possible, its value at a generic point  $t$  will be denoted by  $X(t)$  or, when convenient, by  $X_t$ . The trajectories drawn from  $X$  are denoted by  $x$ .  $Y$  stands for the class label associated with  $X$ .

As usual, transposition and inversion are denoted by the superscripts  $\top$  and  $^{-1}$  respectively, and  $f'$  stands for the derivative of a real function  $f$ . Likewise,  $X'$  corresponds to an independent copy of the random variable  $X$ , and empirical estimators are denoted by either a “hat” or the subscript  $n$ . Non-standard operators are defined on their first use.

For the sake of readability, we include here a relation of the main abbreviations and symbols used in this dissertation.

## Abbreviations

$B$	Standard Brownian motion.
$BB$	Brownian bridge.
$BT$	Brownian motion with a trend.
$cf$	Characteristic function.
$dcor$	Distance correlation.
$dcov$	Distance covariance.
DCT	Dominated Convergence Theorem.
FC	Fisher-Correlation criterion.
FDA	Functional Data Analysis.
fMRI	Functional Magnetic Resonance Imaging.
FPCA	Functional Principal Components Analysis.

$k\text{CV}$	$k$ -fold cross-validation.
$k\text{-NN}$	$k$ Nearest Neighbours.
LDA	Linear Discriminant Analysis.
LOOCV	Leave-one-out cross-validation.
MH	Maxima Hunting.
MI	Mutual Information.
mRMR	minimum Redundancy Maximum Relevance.
NB	Naïve Bayes classifier.
NIR	Near infrared.
NMR	Nuclear Magnetic Resonance.
$OU$	Ornstein-Uhlenbeck process.
PCA	Principal Component Analysis.
PLS	Partial Least Squares.
RK-C	Reproducing Kernel Classifier.
RK-VS	Reproducing Kernel Variable Selection.
$RK_B\text{-C}$	RK-C under a Brownian assumption.
$RK_B\text{-VS}$	RK-VS under a Brownian assumption.
RKHS	Reproducing Kernel Hilbert Spaces.
RN	Radon-Nikodym.
SLLN	Strong Law of Large Numbers.
SVM	Support Vector Machine.

## Symbols

$P_0 \ll P_1$	The measure $P_0$ is absolutely continuous with respect to $P_1$ .
$P_0 \sim P_1$	$P_0$ and $P_1$ are equivalent ( $P_0 \ll P_1$ and $P_1 \ll P_0$ ).
$P_0 \perp P_1$	Both measures are mutually singular.
$\langle f, g \rangle_K$	Inner product in the space $\mathcal{H}(K)$ if $f, g \in \mathcal{H}(K)$ . Otherwise, the congruence defined in Remark (a) of Theorem 2.2.
$\  \cdot \ _K$	Norm in the space $\mathcal{H}(K)$ .
$\mathcal{C}[a, b]$	Space of real continuous functions on $[a, b]$ endowed with the supremum norm.
Cov	Standard covariance.
$d$	Number of selected variables.
$\mathcal{D}$	Dirichlet space defined in Subsection 2.1.2.

$\mathcal{D}_n$	Training sample.
$dP_1(x)/dP_0$	Radon-Nikodym derivative of $P_1$ with respect to $P_0$ .
$\mathbb{E}$	Mathematical expectation.
$\epsilon(t)$	Noise process.
$\eta(x)$	Regression function $\eta(x) = \mathbb{E}(Y X = x)$ .
$\mathcal{F}$	Generic functional space.
$g^*$	Bayes (optimal) rule.
$g_n$	Data-driven classifier.
$h$	Tuning parameter in the estimation of the local maxima.
$\mathcal{H}(K)$	Hilbert space associated with $K$ .
$\mathcal{I}$	Set of indices.
$\mathbb{I}$	Indicator function.
$I(\cdot, \cdot)$	Association measure between two random variables.
$K(\cdot, \cdot)$	Covariance operator.
$L^*$	Bayes error.
$L_n$	Empirical error.
$m(t)$	Mean function.
$n$	Sample size.
$N$	Dimension of the discretization grid.
$\Omega$	Sample space.
$p$	$\mathbb{P}(Y = 1)$ .
$\mathbb{P}$	Probability measure.
$P_0, P_1$	Distributions of the stochastic processes defining the “populations” denoted also $P_0, P_1$ .
$\Phi$	Cumulative distribution function of the standard normal.
$\Phi_{m,k}$	Peak-type functions defined in Section 3.3.
$\mathcal{R}$	Distance correlation.
$\mathcal{R}_X^2$	$\mathcal{R}^2(X(t), Y)$ .
$\mathcal{V}$	Distance covariance.
$\mathcal{V}_X^2$	$\mathcal{V}^2(X(t), Y)$ .
$\text{Var}$	Variance.



In our lust for measurement, we frequently measure that which we can rather than that which we wish to measure... and forget that there is a difference.

George Udny Yule

Not everything that can be counted counts, and not everything that counts can be counted.

William Bruce Cameron



# Chapter 1

## Introduction

### 1.1. Functional Data Analysis

Functional Data Analysis (FDA) is a small part of that huge topic in contemporary science and technology known as Big Data. More specifically, FDA deals (using mathematical and computational tools) with those problems involving the use of data which are "big" in the sense that they are recorded "in continuous time" so that they are in fact real functions. Functional data appear in many significant areas from medicine to economics, taking the form of electrocardiograms, functional magnetic resonance imaging, spectroscopy, biometric signals, paths in space, climate time series or economics indexes. Thus, it seems clear that the proper collection and treatment of these data in order to obtain the best information from them is a fundamental task. Nevertheless, the functional nature of the data makes many classical statistical approaches inappropriate or directly useless, so new approximations and methods are needed.

Since term FDA was probably first coined by [Ramsay \(1982\)](#), the boom of what we mean today by functional data analysis is relatively recent, with no more than two decades of history (since the available techniques did not allow the adequate registration and process of functional data before that time). But despite its novelty, the high research activity in this area has produced a big amount of advances and associated bibliography. A full review of all FDA development to this days exceeds the scope of this thesis by far. However, a comprehensive approximation to the topics which are more closely related to our work (data representation, supervised classification and variable selection) is given in the next sections in order to provide a framework for our research as clear and complete as possible. But first let us comment some basic and general references in the FDA literature that readers interested in this theme might find helpful, as well as

point out some of the main differences between the functional and the multivariate settings.

The former monograph by [Ramsay and Silverman \(2005, 1<sup>st</sup> ed. in 1997\)](#) provided the first collection of ideas and techniques for functional data analysis and has had a major influence in this field ever since. In this practical-oriented book the attention is centred in the  $L^2$  space, and both smoothing techniques and basis representation play a central role (a functional data is assumed to be a realization of a smooth process). Some of these ideas are applied to real-data problems of different areas in [Ramsay and Silverman \(2002\)](#) and the computational details (in both *R* and *MATLAB* languages) can be found in [Graves et al. \(2009\)](#). The original *R* code was distributed by the authors in the *fda* package contributing to the increasing popularity of the FDA. From another point of view, the reference book by [Ferraty and Vieu \(2006\)](#) presents a comprehensive (theoretical and practical) treatment of the nonparametric approach to functional classification, prediction and forecasting in the wider setting of complete normed (and semi-normed) spaces. In this nonparametric setting, the monograph by [Bosq and Blanke \(2008\)](#) provides new mathematical tools for prediction problems with functional data with a major emphasis on the theoretical aspects. On the other hand, the book by [Horváth and Kokoszka \(2012\)](#) centres on inferential methods and their applications with special attention to dependent functional data. Finally, the recent book by [Hsing and Eubank \(2015\)](#) is a first attempt of collecting the mathematical concepts which are relevant to the theoretical foundations of FDA. The selected topics include Reproducing Kernel Hilbert Spaces (RKHS), factor analysis, regression and discriminant analysis.

The increasing interest in FDA is also revealed by the number of special issues and overview papers devoted to these topics that have been published in different journals. Some recent statistical surveys cover essential themes such as regression, classification, clustering and dimension reduction. For example, [Cuevas \(2014\)](#) provides a comprehensive survey on FDA theory and methods, and [Wang et al. \(2015\)](#) pays special attention to the functional regression problem including inverse regression and nonlinear models. It is also worth mentioning the collective book by [Ferraty and Romain \(2011\)](#) in which each chapter is a survey of a different topic by different authors, and the last two chapters of [Goldfarb et al. \(2011\)](#), by [Delsol et al. \(2011\)](#) and [González-Manteiga and Vieu \(2011\)](#). Finally, the applications of FDA in specific fields are also covered in thematic overviews such as [Burfield et al. \(2015\)](#), in chemometrics, or [Ullah and Finch \(2013\)](#) in biomedicine, with a singular systematic style.

Finally, it is noteworthy that in this booming field of statistics with functional

data, the computational and numerical aspects, as well as the real data applications, have had (understandably) a major role so far. However, the underlying probabilistic theory, connecting the models (i.e., the stochastic processes) which generate the data is far less developed. As pointed out by [Biau et al. \(2015\)](#), “*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*”. The present thesis can be seen as a contribution to partially fill this gap regarding the relevant supervised (binary) classification setting and the associated dimension reduction problem via variable selection.

### 1.1.1. Some basic notions and difficulties in FDA

The references mentioned above and many other works place FDA as an area of interest with many potential applications. So it is time to define what a functional data is. We have said that functional data can be curves, images, surfaces or more complex structures, i.e., any observation living in a functional (infinite dimensional) space. In this way, [Ferraty and Vieu \(2006\)](#) defines a functional data  $x$  as an observation of a random variable  $X$  which takes values in a functional space  $\mathcal{F}$ . This thesis focuses on the most common case of real functions defined in a bounded interval, which arises in a wide variety of situations: spectrometry, genetics, medicine, economics, biometrics, etc. Therefore, we precise the definition of functional data in terms of stochastic processes (this approach if followed, e.g., in [Hsing and Eubank \(2015\)](#)). 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  $\Omega$ . Then a functional data is just a realization (often called “trajectory”) of a stochastic process for all  $t \in \mathcal{I}$ . Provided that no confusion is possible, we will denote the whole process by  $X$ . Its value at a generic point  $t$  will be denoted by  $X(t)$  or, when convenient, by  $X_t$ .

In the functional setting stochastic processes play the role of random variables in classical statistics. Continuing with this analogy, Gaussian processes occupy the place of the normal distribution in  $\mathbb{R}^n$ . A stochastic process is said to be Gaussian if and only if, for all  $t_1, \dots, t_k \in \mathcal{I}$  the  $k$ -dimensional random vector  $(X(t_1), \dots, X(t_k))$  has a normal distribution. Like their multivariate counterparts, the distribution of Gaussian processes are fully determined by the mean function and the covariance operator, although in the functional case the family of Gaussian processes is not a parametric model. These processes have many other well-studied and valuable properties that can be found in any standard reference (see for example [Doob \(1953\)](#)). We will focus on this “central” type of processes several times throughout this work with special attention to the Standard Brown-

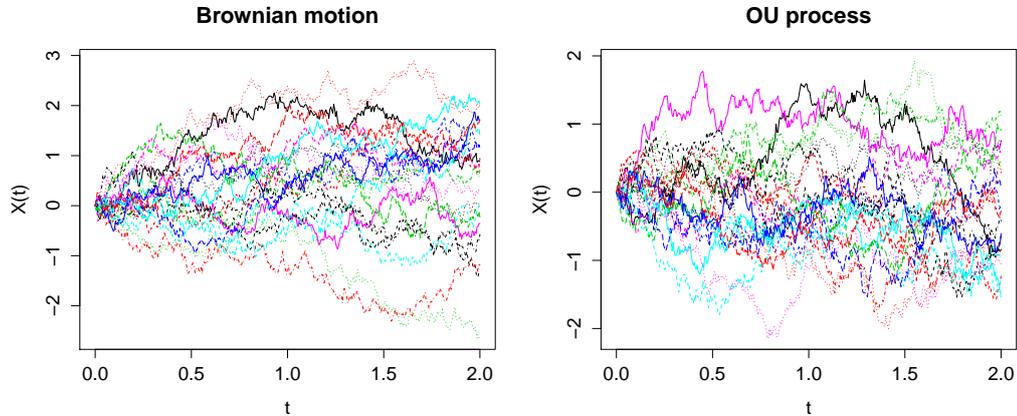


Figure 1.1: Some trajectories of a standard Brownian motion (left) and a Ornstein-Uhlenbeck process (right).

ian motion. The Wiener process or standard Brownian motion  $B(t)$  is a Gaussian process with stationary independent increments such that  $m(t) = \mathbb{E}(B(t)) = 0$  and  $K(s, t) = \text{Cov}(B(s), B(t)) = \min(s, t)$ . It is widely used in fields like finances, mathematics, physics or chemistry, since more complicated random processes can be ultimately described in terms of  $(t)$ . Some trajectories of the Brownian motion are plotted in Figure 1.1 (left panel) along with several realizations of the Ornstein-Uhlenbeck process (right panel) which is also used in our experiments. A more detailed description of the processes considered in the simulation experiments can be found in Subsection 5.2.1. The book by Mörters and Peres (2010) is a complete (nicely written) monograph about the Brownian motion.

The election of the function space  $\mathcal{F}$  where the trajectories live, is a strategic issue since it determines the collection of mathematical tools available. Probably  $L^2[a, b]$ , the space of real square-integrable functions on  $[a, b]$ , is the most popular choice. This space endowed with the usual norm induced by the inner product

$$\langle x, y \rangle^{1/2} = \left( \int_a^b x(t)y(t)dt \right)^{1/2},$$

is a separable Hilbert space (with all the advantages that would come from it). Another commonly used setting is to consider the space of real continuous functions on  $[a, b]$ , denoted by  $\mathcal{C}[a, b]$ , endowed with the supremum norm,

$$\|x\| = \sup_{t \in [a, b]} |x(t)|.$$

This is a Banach space so it is more difficult to work in this setting. Nevertheless one can still use many existing classical results (see e.g. Billingsley (2013)).

Other functional settings have been considered in the literature, often for very specific problems. One of the most interesting approaches is the use of subspaces endowed with a suitable semi-metric (Ferraty and Vieu, 2006). Maybe the semi-norm based on the derivatives  $\|x\| = \langle x', x' \rangle^{1/2}$  is the better known example. In this work we will mostly use either the spaces  $L^2$  and  $\mathcal{C}$ , or the Reproducing Kernel Hilbert Space (RKHS) associated with the kernel of the underlying process (introduced in Chapter 2). Without loss of generality we will usually consider these spaces defined on the interval  $[0, 1]$ .

The special features of these function spaces (and, in particular, their infinite-dimensional character) are the main source of problems and the reason of the particularities that appear in FDA. However this also gives raise to the study of new phenomena (as the so-called “near perfect classification”) and further theoretical and practical research. We just mention below some of the most representatives difficulties and differences (with respect to finite-dimensional statistics) which appear in FDA.

- The usual norms are no longer equivalent, so different norms could reveal (or hide) different information. The chosen metric must be then coherent with the data, which is not always easy to achieve.
- There are multiple possible representations for the same data set (depending, for example, from the basis we chose).
- Graphical tools have been mainly derived for the usual  $L^2[a, b]$  space. Then, data belonging to other spaces can be hard to display properly.
- Probability measures are difficult to handle. No natural translation-invariant measure plays the role of Lebesgue measure in  $\mathbb{R}^n$ , so there are not natural density functions.
- There is no obvious order structure, so notions like centrality or outliers are more difficult to formalize and distribution functions cannot be defined.
- The orthogonality notion is also lost if we do not work in a Hilbert space.
- Those variables which are “close together” in the family  $X(t), t \in [0, 1]$  are often highly correlated, leading to nearly singular covariance matrices for which many usual methods fail. Redundancy also has detrimental effects in many standard classifiers (even when they do not use covariance matrices).
- Functional data are difficult to record. In practice some kind of discretization or dimension reduction method must be used.

- Function spaces are “difficult to fill”. This means that usually one needs huge sample sizes to get many functional data in a small neighbourhood of a given function. As a consequence, non-parametric methods (which are typically of “local” nature) have often slow convergence rates so that large sample sizes are needed.

Some of these points, specially those concerning the infinite dimensionality, redundancy between variables and representation issues, will be further developed in the next sections and chapters. We will also see the implications of choice of the space where the functional data are supposed to “live”.

## 1.2. Supervised Classification

The discrimination procedures (also called *supervised classification methods* in modern terminology) are now a commonplace in the standard use of statistics. Their origins go back to the classical work by Fisher (1936), motivated by biological taxonomy problems. Today, biomedical sciences remain as a major field of application of discrimination techniques but other areas, like engineering, provide also several important problems (signal theory, image analysis, speech recognition...). The books by Devroye et al. (2013), Hastie et al. (2009) and Duda et al. (2012) offer insightful, complementary perspectives of this topic. In the rest of this thesis we will focus on the binary discrimination problem, even though many methods and results can be immediately extended to the multiclass case.

While the statement and basic ideas behind the discrimination problem are widely known, we need to briefly recall them for the sake of clarity. Suppose that an explanatory random variable  $X$  (say, e.g., the result of a medical analysis) taking values in a *feature space*  $\mathcal{F}$  can be observed in the individuals of two populations  $P_0$  and  $P_1$  (e.g.,  $P_0$  could be the population of healthy individuals and  $P_1$  that of people suffering from a certain illness). Let  $Y$  denote a binary random variable, with values in  $\{0, 1\}$ , indicating the membership to  $P_0$  or  $P_1$ . On the basis of a data set  $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$  of  $n$  independent observations drawn from  $(X, Y)$ , the discrimination problem aims at predicting the membership class  $Y$  of a new observation for which only the variable  $X$  is known. In the medical example, the goal would be to classify a patient as healthy or ill from the observation of  $X$ , in view of the experience provided by the data base  $\mathcal{D}_n$  of well-classified patients (this accounts for the expression *supervised classification*).

A *classifier* or *classification rule* is just a measurable function  $g : \mathcal{F} \rightarrow \{0, 1\}$ . It is natural to assess the performance of a classifier by the corresponding *classi-*

fication error  $L = \mathbb{P}(g(X) \neq Y)$ . It is well known that this classification error is minimized by the so-called *Bayes classifier*,

$$g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}, \quad (1.1)$$

where  $\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$ . The corresponding minimal “classification error” (i.e., the misclassification probability)  $L^* = \mathbb{P}(g^*(X) \neq Y)$  is called *Bayes error*. Since this optimal (Bayes) classifier is in general unknown, the purpose of (binary) supervised classification is just to construct data-driven classifiers  $g_n: \mathcal{F} \rightarrow \{0, 1\}$ , with  $g_n(x; \mathcal{D}_n) = g_n(x)$ , aimed at providing reasonable (in some sense) approximations of  $g^*$ . A common strategy is the so-called *plug-in approach*, consisting in replacing  $\eta(x)$  with a suitable data-driven estimator. The goodness of these classifiers is assessed in relation with the Bayes error, in this sense a sequence of classifiers  $\{g_n\}$  is *weakly consistent* if  $L_n \rightarrow L^*$  in probability as  $n \rightarrow \infty$ , and it is *strong consistent* if  $L_n \rightarrow L^*$  almost surely (a.s.) when  $n \rightarrow \infty$ .

Since the distribution of  $(X, Y)$  is also unknown in general, the error associated with a classifier  $g_n$  is unknown too. However it can be easily estimated by the *empirical risk*

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{g_n(X_i) \neq Y_i\}}.$$

This empirical risk (calculated over an independent test sample) is the usual criterion for comparison among different methods in our simulation experiments. This can also be used to construct new classifiers with the only goal of minimizing this error. Similarly to wrapper methods described in Section 1.4, the idea lies in the generation of a range of classification rules and the election of that which minimizes the empirical risk. This methodology is supposed to produce, in general, more accurate but less general classifiers (in the sense that the minimization of this error is completely data-dependent) than the plug-in approach. Empirical-risk classifiers are also supposed to converge faster to the Bayes error, but it is a controversial point (Audibert and Tsybakov, 2007).

### 1.2.1. Classification with functional data

The general setup outlined before remains valid in the functional setting, but here the feature space  $\mathcal{F}$  will be an infinite-dimensional functional space. Typical choices for  $\mathcal{F}$  are  $\mathcal{F} = L^2[0, 1]$  and  $\mathcal{F} = \mathcal{C}[0, 1]$ . Thus, our data are of type  $(X_1, Y_1), \dots, (X_n, Y_n)$ , where the  $X_i$  are iid trajectories in  $L^2[0, 1]$  or  $\mathcal{C}[0, 1]$ , drawn from a stochastic process  $X = X(t) = X(\omega, t)$ .

Although the formal statement of the supervised functional classification essentially coincides with that of the classical multivariate one, some important differences appear when dealing with functions instead of vectors. From the point of view of the classification rules, the similar setting allows us a more or less direct extension of many standard multivariate classifiers such as  $k$  Nearest Neighbours ( $k$ -NN) or kernel methods, but with some restrictions or inconveniences. Perhaps, the most noticeable case is that of the popular linear Fisher rule, or Linear Discriminant Analysis (LDA). The infinite dimension and high collinearity of functional data makes the covariance operator non-invertible and the associated (by discretization) covariance matrix nearly singular, so LDA is not feasible for FDA (the same is applicable to any method which requires the inversion of this operator). Many different strategies have been developed in order to overcome this problem: regularization methods adding different penalization to the covariance matrix (Friedman, 1989; Hastie et al., 1995), the use of a suitable basis representation (see next Section), or more specific methodologies such as the functional linear discrimination analysis by James and Hastie (2001), which deals with fragments of curves and sparse data. In Chapter 2 we propose a new adaptation of the Fisher rule suitable for functional data (which does not require any regularization or new representation of the data). Furthermore, even when the classifier extension is possible, it entails theoretical or/and computational costs. For example, it is well known that  $k$ -NN and kernel rules are universally consistent in  $\mathbb{R}$  while this consistency is no longer valid in the functional case without non-trivial assumptions (C erou and Guyader, 2006; Abraham et al., 2006).

Differences between the multivariate and the functional cases are beyond the construction of classification rules. A good example of this is the *near perfect classification* phenomenon described by Delaigle and Hall (2012a). The authors show that in the functional setting there are non-trivial problems where classifying without any error is possible, and also problems for which linear methods often achieve the best results if near perfect classification is not possible. Note that this can never happen in finite dimensional spaces except for degenerate problems. The phenomenon is also characterized in terms of the convergence of certain series, in words of the authors: “*The theoretical foundation for these findings is an intriguing dichotomy of properties and is as interesting as the findings themselves.*”. In Chapter 2 we will show that this dichotomy can be also explained in terms of the probability distributions of the involved processes. In particular, the near perfect classification phenomenon is related with the orthogonality of the probability measures involved in the classification problem. We will also construct a new interpretable linear classifier which estimates the optimal one under some conditions.

Dozens of classifiers could be used according to the data under study, the goals of the analysis, computational or time requirements, etc. Several functional classifiers have been considered in the literature; see, e.g., [Baíllo et al. \(2011\)](#) for a survey. For other recent proposals see, for example, [Cuesta-Albertos et al. \(2015\)](#) and [Martin-Barragan et al. \(2014\)](#). In addition, after a dimension reduction any multivariate classifier is a valid choice (at least in principle). In this thesis, we are primarily concerned with the comparison of several variable selection methods under the same conditions (on the more general possible way). Our goal is to get good variable selection methods (working properly in a wide range of situations) rather than to get the best classification rate for a specific problem. For these reasons, we have chosen a small number of popular and not too complex classifiers but with good performance in practice, which are commonly used as benchmarks in the literature. In fact, we will see that some of these simple classifiers achieve really good results for different problems.

Maybe the simplest one is the so-called  $k$  nearest neighbours rule, according to which an observation  $x$  is assigned to  $P_1$  if and only if the majority among the  $k$  sample observations  $X_i$  nearest to  $x$  fulfil  $Y_i = 1$ . In general,  $k$ -NN could be considered (from the limited experience so far available; see e.g., [Baíllo et al. \(2011\)](#); [Dudoit et al. \(2002\)](#)) a sort of benchmark, reference method for functional supervised classification. Simplicity, ease of motivation and general good performance (it typically does not lead to gross classification errors) are perhaps the most attractive features of this method.

Other, more recent, extremely popular classifier is the so-called Support Vector Machine (SVM); see [Cortes and Vapnik \(1995\)](#). These classifiers depend on an auxiliary function called “kernel”. The SVM classifier based on a linear kernel is particularly successful (see e.g. [Díaz-Uriarte and Alvarez de Andrés \(2006\)](#) or [Gönen \(2011\)](#)) and is probably the linear method of reference today. This reproducing kernel based methodology is often used for comparisons in reference where similar studies are carried out ([Ding and Peng, 2005](#); [Peng et al., 2005](#)).

Finally, we will consider the popular *Fisher’s linear classifier* (LDA) used often in classical discriminant analysis. This is a well-known rule which is commonly utilized as a reference because of its popularity, simplicity and good performance in many low-dimensional real problems (see, e.g. [Hand \(2006\)](#)). This is specially recommended when distributions are Gaussian, which will be frequent in our context. However, note that LDA can be used only on the “reduced data” resulting from a dimension reduction method while  $k$ -NN and SVM can deal with the entire data.

Other recurrent classifiers in similar studies were also considered but they are just occasionally commented (or not included) in this dissertation for the sake of clarity and concision. The results for these classifiers are nearly analogous to those presented in this work and will be briefly commented in Chapter 5.

### 1.2.2. Supervised classification and absolute continuity

As we will comment below, the relationship between the probability measures involved in the classification problem entails strong consequences regarding the optimal rule and the optimal classification error. In fact, the absolute continuity or mutual singularity of these measures determine whether one can achieve a perfect classification in some models or not, and Radon-Nikodym derivatives (in the absolute continuous case) are related with the calculation of explicit expressions for the Bayes rules.

The expression  $P_1 \ll P_0$  indicates that  $P_1$  is absolutely continuous with respect to  $P_0$  (i.e.  $P_0(A) = 0$  entails  $P_1(A) = 0$ ). Note that, from the Hájek-Feldman dichotomy for Gaussian measures (Feldman, 1958),  $P_1 \ll P_0$  implies also  $P_0 \ll P_1$ , so that both measures are in fact mutually absolutely continuous (or “equivalent”). This is often denoted by  $P_0 \sim P_1$ .

When  $P_0$  and  $P_1$  are completely known in advance and  $P_1 \ll P_0$ , the optimal classification rule (often called *Bayes rule*) is

$$g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}} = \mathbb{I}_{\left\{\frac{dP_1(x)}{dP_0} > \frac{1-p}{p}\right\}}, \quad (1.2)$$

where  $\mathbb{I}$  denotes the indicator function,  $\eta(x) = \mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x)$ ,  $p = \mathbb{P}(Y = 1)$  and  $dP_1(x)/dP_0$  is the Radon-Nikodym derivative of  $P_1$  with respect to  $P_0$ ; see (Baíllo et al., 2011, Thm. 1) for additional details.

If the Radon-Nikodym derivative  $dP_1(x)/dP_0$  is explicitly known, there is not much else to be said. However, in practice, this is not usually the case. Even if the general expression of  $dP_1(x)/dP_0$  is known, it typically depends on the covariance  $K(s, t) = \text{Cov}(X(s), X(t))$  and mean functions  $m_i(t) = \mathbb{E}(X(t)|Y = i)$  of  $P_i$ ,  $i = 1, 2$ .

As said above, the term “supervised” accounts for the fact that, in any case, a data set of “well-classified” independent observations  $\mathcal{D}_n$  is assumed to be available beforehand. Therefore, a common strategy is to use these data to estimate the optimal rule (1.2). This plug-in approach is often implemented in a non-parametric way (e.g., estimating  $\eta(x)$  by a nearest-neighbour estimator) which

does not require much information on the precise structure of  $\eta(x)$  or  $dP_1(x)/dP_0$ . However, in some other cases we have a quite precise information on the structure of  $dP_1(x)/dP_0$ , so that we can take advantage of this information to get better plug-in estimators of  $g^*(x)$ . This idea will be developed in Chapters 2 and 3.

## 1.3. Functional data representation

In this section we refer to the difficulties of FDA appearing during the data preprocessing due to three principal causes: the choice of the functional space and representation of data, the infinite dimension of the observations and the data collection procedure.

### 1.3.1. Smoothing and basis representation

From a practical point of view, it is impossible to observe a complete functional data. Nowadays, high-tech sensors allow for monitoring processes in finer and finer grids, but at the end some sort of discretization used. Therefore, what we have in practice is not the process  $\{X(t), t \in [0, T]\}$  but a high-dimensional vector  $(x(t_1), \dots, x(t_N))$  in the discrete grid  $t_1, \dots, t_N$ . And this vector of highly correlated covariates represents the functional data. Indeed, we will often have a collection of  $n$  of these observations that is usually represented as a matrix with  $n$  files and  $N$  columns; this is the *training sample* or *training set*. A question to discuss is whether these vectors are true functional data. In our opinion the answer is affirmative: on the one hand, in some cases the grid can be as fine as desired so the process is virtually observable at any instant  $t$ . On the other hand, we can choose a functional model to approximate the data. So, the idea is that behind these vectors there are real functions with all the corresponding implications, or in words of [Ramsay and Silverman \(2005\)](#): “*The term functional in reference to observed data refers to the intrinsic structure of the data rather than to their explicit form*”.

Furthermore, the presence of noise presents similar problems. In practice, most functional data are contaminated with random noise. This is also called measurement error, although sometimes the source of noise will not be an error in the measurement. Noise is usually modelled considering that we observe a data  $y(t) = X(t) + \epsilon(t)$ , where  $X(t)$  is a real function and  $\epsilon(t)$  is random noise independent from  $X(t)$  such that  $\mathbb{E}[\epsilon(t)] = 0$  and  $\text{Var}\epsilon(t) = \sigma^2$ . These errors are sometimes insignificant, for example when recording the height of children along time, but in other cases noise is a critical point, as head movements when taking fMRI’s (functional magnetic resonance images).

The goal in both cases is obtaining the original function from the (noisy) observed vector, or sometimes, getting a smoother version for further actions, for example to take derivatives. There are FDA tools that allow us to “recover” the original function (or a very close approximation) from the observation or just get a smooth enough approximation. These methods are basically grouped in two big families: basis representation and smoothing.

The basis representation is a recurrent tool in functional data analysis. It plays the double role of getting a continuous function and reducing the dimension by truncating the expansion series. A full review of these topics can be found in (Ramsay and Silverman, 2005, ch. 3). Assuming that the data  $x(t)$  live in a functional space  $\mathcal{F}$  and let  $\{\phi_i(t)\}_{\mathbb{N}}$  be a basis of that space, then  $x(t)$  can be represented by the series  $x = \sum_{i=1}^{\infty} c_i \phi_i(t)$ , where  $\{c_i\}$  are the coefficients corresponding to the basis. In practice, the infinite series is truncated at certain element  $k$  and we approximate  $x(t)$  by  $\tilde{x}(t) = \sum_{i=1}^k c_i \phi_i(t)$ . This number  $k$  is a sort of smoothing parameter that must be carefully chosen. Larger values of  $k$  produce good approximations (perhaps incurring in over-fitting) but keep the high dimensionality problem. Besides, smaller values of  $k$  make the data easy to handle but some important information could be lost. Unluckily, there is no general rule to set the best number of elements. This phenomenon (over or under-smoothing) is illustrated in Figure 1.2 where a (noisy) functional observation is approximated by B-splines. The grey curve corresponds to the real function, circles are the observation points and blue and red lines stand for two B-splines approximations with  $k = 5$  and  $k = 40$  respectively. The smaller value of  $k$  cannot approximate the underlying model (under-fitting) while the larger does not replicate the original function but the noise (over-fitting).

The functional space  $\mathcal{F}$  is typically supposed to be  $L^2[0, 1]$ . In this case there exist an orthonormal basis  $\{e_i\}$  and the coefficients can be easily calculated by means of the inner product,  $\tilde{x}(t) = \sum_{i=1}^k \langle x(t), e_i(t) \rangle e_i(t)$ . Of course, the properties of  $\tilde{x}(t)$  depend on the basis functions. Hence, an adequate choice of the basis, according to the nature of the data, is needed. Among the wide variety of functional basis (exponential, polynomial, etc.) maybe the most frequently used are the following.

*Fourier.* The Fourier basis in the interval  $[0, T]$  has the form  $\tilde{x}(t) = c_0 \phi_0 + \sum_r c_{2r-1} \phi_{2r-1}(t) + c_{2r} \phi_{2r}(t)$ , where  $\phi_0(t) = 1/\sqrt{T}$ ,  $\phi_{2r-1}(t) = \sin(r\omega t)/\sqrt{T/2}$ ,  $\phi_{2r}(t) = \cos(r\omega t)/\sqrt{T/2}$ . These are periodic functions of period  $2\pi/\omega$ . If the discretization grid is equispaced and the period is equal to  $T$ , then the basis is orthonormal. Fourier basis is very easy to derive and the coefficients can be efficiently calculated via the *Fast Fourier Transform* algorithm (FTT).

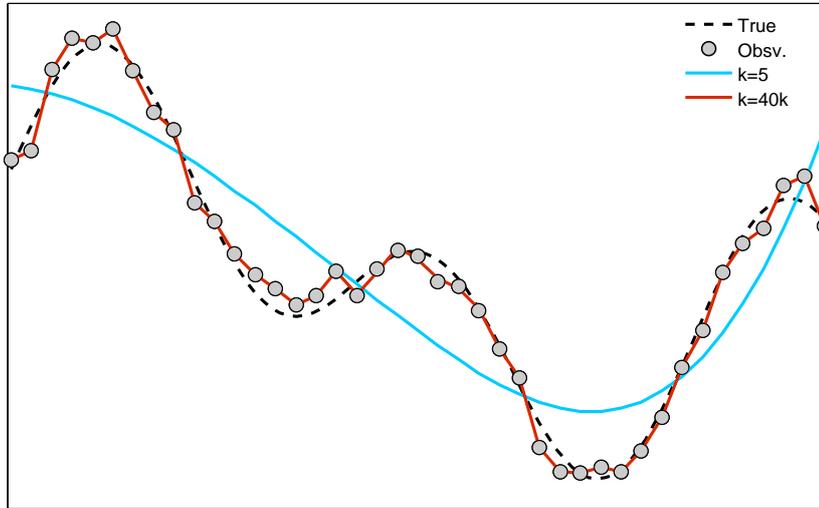


Figure 1.2: Approximation of a functional data via B-splines. Grey: true data. Blue: under-fitted approximation ( $k = 5$ ). Red: over-fitted approximation ( $k = 40$ ).

*B-Splines.* Splines basis might be the most popular basis nowadays (for non-periodical data). This basis combines the efficiency of polynomials (which are included in it) with a greater flexibility, so that it usually needs just a few elements. The technique basically lies on dividing the time interval and making a polynomial approximation in each subinterval while taking care of the breakpoints. Many variants can be used for particular cases, see [De Boor \(1978\)](#) for some examples. We will use expansions of this type when taking derivatives in practice.

*Wavelets.* The idea behind this representation is that any function in  $L^2$  can be properly approximated by suitable *mother wavelet* function  $\psi$  and its translations and dilations of the form  $\psi_{m,k}(t) = 2^{m/2}\psi(2^m t - k)$ ,  $m, k \in \mathbb{N}$ . The use of these basis is associated with treatment of signals since they can deal with discontinuous and nondifferentiable functions in a natural way, but now it is an extended practice also in FDA (see e.g., [Pigoli and Sangalli \(2012\)](#); [Antoniadis et al. \(2013\)](#)). In some examples of this thesis, the so-called Haar basis (formed by square pulses) will play a relevant role.

*Empirical.* Empirical bases are constructed from the data aiming at optimizing some particular target. The most popular one is that obtained using Functional Principal Components Analysis. This is an extension of the multivariate functional data analysis through the Karhunen-Loève expansion. This approach, which tries to capture the variance of the data in the new representation, have been successfully employed in many FDA techniques ([Ramsay and Silverman, 2005](#)). However, it does not consider possible relations with other variables (e.g. the class label), so other representations which take into account these relationships seem

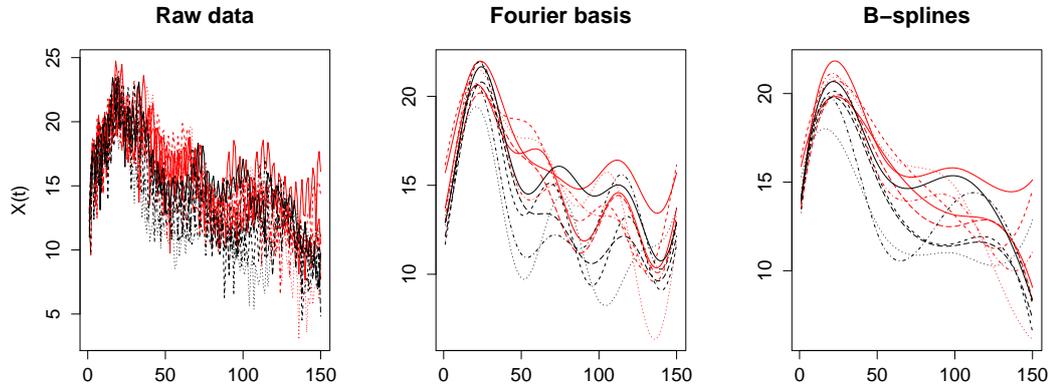


Figure 1.3: Some Phoneme trajectories with different representations: raw data (left), Fourier basis (middle) and B-Splines (right). Observations of different classes are in different colours.

to be more suitable if supervised classification is the final goal. This is the case of Partial Least Squares (PLS) which aims at maximizing the covariance between the data and the response variable (Delaigle and Hall, 2012b).

Figure 1.3 illustrates the effects of two different bases mentioned above. Left panel shows some raw trajectories of the well known *Phoneme* dataset (which will be discussed in Section 5.3). Middle and right panels present the same trajectories approximated with a Fourier basis and B-splines respectively, with 6 elements each.

Beyond these approximations, there exists a great variety of smoothing methods (with or without a basis representation) to remove noise or just making the discretized data continuous. The classical way to proceed is convolving the data with a smooth weighting function (maybe Nadaraya-Watson is the best known). This approach produces linear models which are determined by two elements of the weighting function: the kernel function (often Gaussian) and the smoothing parameter or bandwidth. The book by Ramsay and Silverman (2005, ch. 4-6) provides a good summary of specific smoothing methods including different penalties and constraints. More complex models without homoscedasticity, independence or with a more difficult data structure have been considered in the literature (see e.g. Yao et al. (2005) for a method to estimate the variance of the noise in sparse longitudinal data).

In the end, the way to treat functional samples is a widely discussed topic in FDA with no universal solution. Each data set is different and probably requires a specific treatment. However, although these are interesting and relevant topics,

this thesis is not primarily concerned with this preprocessing steps but with the comparison of different classifiers and dimension reduction techniques under the same conditions. Then, with this criterion in mind and provided that there is no standard way to proceed, we have followed the next general practical rules (which are common in many classification and variable selection works). First, we will assume that the noise has been removed in a previous step. Second, we have worked, when possible, with raw data. That is, we have worked preferably with discretized data without any smoothing or basis representation. Although the election of an adequate partition is not as trivial as it may appear at first glance (there are particular aspects of the data, such as the curvature, that have to be taken into account) the major risk choosing a fine enough equispaced discretization grid is the increment in computing time, while the proper election of a basis or the bandwidth is much more delicate and can entail a loss of discriminant information. Moreover, it is far from clear whether smoothing would be a good practice when classifying functional data. The work by [Carroll et al. \(2013\)](#) shows that the usual smoothing parameters (with good and even optimal performance in prediction and hypotheses testing) fail in this context, and undersmoothing is recommended as a practical guideline. Indeed, the best results with two of the three evaluated classifiers are achieved with the raw data without any smoothing. Our experience also points to that direction, that is, smoothing is a critical issue that can entail a loss in classification accuracy. Nevertheless, in examples with extremely rough trajectories, certain level of smoothing is frequently useful. Some examples along both lines will be given in this dissertation.

### 1.3.2. Other issues

Other common sources of problems when working with functional data are the registration and display of the observations. Many practical concerns arise during the registration of the data which can frustrate even the simplest analysis. The range of registration problems includes missing or incomplete data, shift registration, different scales, etc. The consideration of these issues is far beyond the scope of this work, but some reference can be found in the general bibliography on FDA (Section 1.1). At a practical level, we will consider these problems to be solved in advance for our data.

The way of displaying functional data is more relevant for this thesis. What do we mean by this? An appropriate display of functional data is not an easy task. Since we can obtain very different information looking at them in a way or another. Usually, the standard plot of  $x(t)$  as a function of the time is less informative than other possible choices. For example, important concepts in multivariate data analysis, such as centrality or proximity, are no longer so easy to determine

or visualize in FDA. That is a direct consequence of the functional nature commented in Section 1.1. Because of the non equivalence of the metrics in functional spaces, the space in which we place the data plays a fundamental role in the information we can deduce from the data (including visual representations). Therefore, as pointed out before, a good choice of the space and an appropriate metric is a key point in FDA, and what might be considered as a disadvantage actually opens a world of possibilities, which the multivariate setting lacks. Good examples of this phenomenon are the use of different semi-metrics (see [Ferraty and Vieu \(2006\)](#)) or the research about functional depth measures (see, e.g., [Cuevas et al. \(2007\)](#); [López-Pintado and Romo \(2009\)](#)). Probably the best known examples are those concerning derivatives. It is well known that, at times, derivatives are much more informative than the sample curves, providing new insights (many uses of the derivatives are well documented in [Ramsay and Silverman \(2005\)](#); [Ferraty and Vieu \(2006\)](#)). This also happens in the supervised classification problems when the information provided by the derivatives turns out to be essential in some cases. A typical example which illustrates this phenomenon is given by the near infrared spectroscopy (NIR) problems. Usually, NIR data consist of very smooth and homogeneous curves with small differences between the classes, so the classification is hard. However, taking the derivatives, often reveals big differences and the classification problem becomes much easier. Figure 1.4 shows this fact through two real NIR datasets: the classical *Tecator* data and a set of *Wheat* samples (both data sets are explained in Section 5.3). Trajectories of different classes are plotted in different colours. Faded lines stand for the sample trajectories while thick lines represent the mean function of each class. First row corresponds to *Tecator* and the second one to *Wheat* samples. Left panel shows several trajectories of the original data sets and right panels present the derivatives. The gain is quite obvious and it is empirically confirmed: while raw data performance is quite modest, the derivatives achieve near perfect results (see Sections 2.5 and 3.4 for details). Finally, let us recall that in practice, taking the derivatives commonly requires a previous smoothing step to make the data differentiable. Throughout this thesis, the derivatives have been estimated (when needed) using spline smoothing.

## 1.4. Variable selection

### 1.4.1. Motivation

The use of high-dimensional or functional data entails some important practical issues. In addition to the inconveniences associated with increasing computation time and storage costs, high-dimensionality introduces noise and redundancy. Thus, there is a strong case for using different techniques of dimensionality reduc-

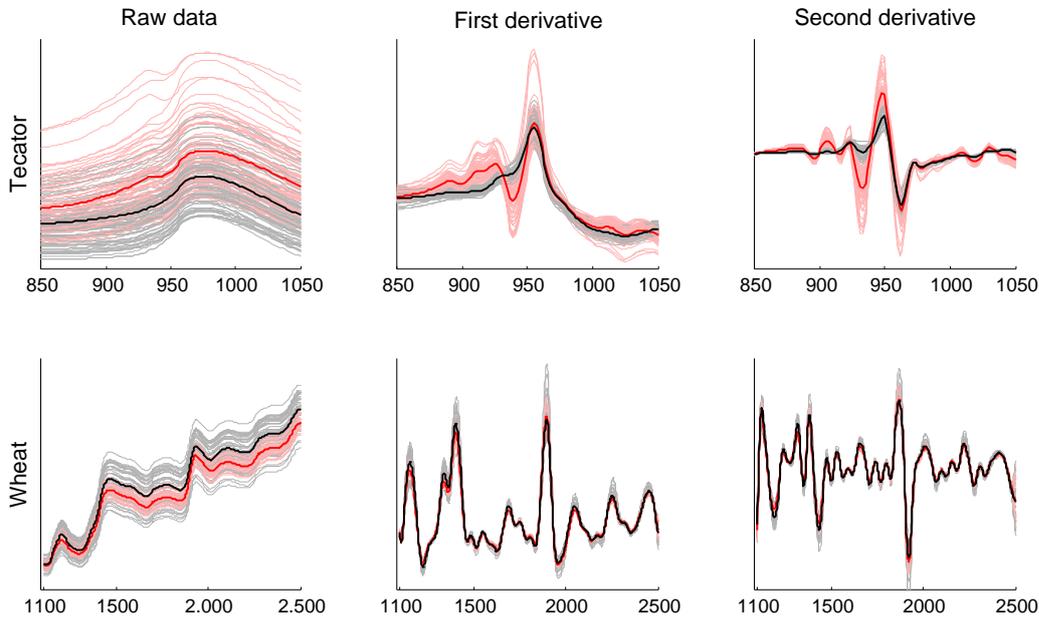


Figure 1.4: Raw trajectories (left) and derivatives (right) of Tecator (first row) and Wheat (second row) real data sets. Faded lines stand for sample trajectories are thick lines represent the mean function of each class.

tion.

We will consider here dimensionality reduction via variable selection techniques. The general aim of these techniques is to replace the original high-dimensional (perhaps functional) data with lower dimensional projections obtained by just selecting a small subset of the original variables in each observation. In the case of functional data, this amounts to replace each observation  $\{x(t), t \in [0, 1]\}$  with a low-dimensional vector  $(x(t_1), \dots, x(t_k))$ . Then, the chosen statistical methodology (supervised classification, clustering, regression,...) is performed with the ‘reduced’ low-dimensional data. The variables must be selected according to some desirable criterion of representativeness in order to achieve the original task with the same or better performance.

A first advantage of such a “radical” dimension reduction is interpretability. When compared with other dimension reduction methods based on general projections, such as Principal Component Analysis (PCA) or Partial Least Squares (PLS), the output of any variable selection method is always directly interpretable in terms of the original variables, provided that the required number  $k$  of selected variables is not too large. This general advantage is even bigger in the functional

setting where the interpretation of the synthetic components is still harder. In a way, variable selection appears as the most natural dimension reduction procedure in order to keep in touch, as much as possible, with the original data. This is a very appreciated property in certain research areas, specially in biomedicine where the identification of relevant metabolites, genes, etc. for further research is a fundamental task. For example, [Díaz-Uriarte and Alvarez de Andrés \(2006\)](#) point out that in microarray problems “probably a more challenging and relevant issue [than improving prediction rates] is to identify sets of genes with biological relevance”. They argue that given a real data set, different classifiers often obtain analogous results.

A second advantage is that variable selection makes sense in real problems. In most real situations, experts do not consider all the available variables since the relevant information is usually concentrated in some points of interest. See for instance [Golub et al. \(1999\)](#); [Lindquist and McKeague \(2009\)](#) among many other examples in experimental sciences or engineering: in [Golub et al. \(1999\)](#) the authors note that 50 genes (among almost 7000) are enough for cancer subtype classification. Likewise, [Lindquist and McKeague \(2009\)](#) point out that in some functional data regression (or classification) problems, as functional magnetic resonance imaging or gene expression, “the influence is concentrated at sensitive time points”.

Third, variable selection entails classification benefits. Classifying high dimensional and functional data involves some difficulties (curse of dimensionality, redundancy, noise, etc.) which result in a loss of classification accuracy. A suitable variable selection method can overcome (at least partially) these problems and lead to equivalent or even better discrimination performances with the reduced data sets. This has been empirically shown many times in the multivariate case, see for instance [Guyon et al. \(2002\)](#); [Ding and Peng \(2005\)](#); [Díaz-Uriarte and Alvarez de Andrés \(2006\)](#); [Karabulut et al. \(2012\)](#). This is also one of the main conclusions of the extensive experiments with functional data carried out in this thesis (see Chapters 2-5). In summary, variable selection has been empirically proved to be a reliable dimension reduction methodology when the objective is discrimination of high dimensional and functional data.

Finally, variable selection with functional data can be also theoretically motivated. Along this dissertation we will see that in several non-trivial functional models, the optimal classification rule depends only on finitely many variables, so the best we can do in these cases is finding them by means of a suitable selection method.

### 1.4.2. Some general terminology and references on dimension reduction methods

There are two main objectives that the variable selection methods should strive for in a classification setting,

- To identify relevant variables for a posterior investigation. As a matter of fact, variable selection is sometimes the main target itself in many cases where the focus is on model simplification. Here, the most important thing is to detect all variables with significant information (in the sense of discrimination), no matter how big the resulting set is or how redundant the selected variables are.
- To select small sets of variables that could be used for class prediction. In this case, we look for sets of non redundant variables that can achieve good predictive performances and are as small as possible.

In the present work we will focus on the second point, developing a theoretical framework and providing new selection methods for that aim. Next, we will briefly comment the principal approaches and methodologies of variable selection to this day in order to give a context of our new contributions. Let us recall that, although only a small proportion of the existent variable selection procedures have been designed for functional data, the general ideas that apply in the multivariate context remain usually valid in the FDA setup.

There is a vast literature on variable selection published by researchers in machine learning and mathematical statisticians. The approaches and the terminology used in these two communities are not always alike. Thus, in machine learning language, variables are commonly called *features* or *attributes* and variable selection is often referred to as *feature selection*, though this term is sometimes used in a more general way to include the generation of new features. It is very common as well (especially in the setting of regression models) to use the terms “sparse” or “sparsity” to describe situations in which variable selection is the first natural aim; see e.g., [Gertheiss and Tutz \(2010\)](#) and [Rosasco et al. \(2013\)](#). It has been also argued in [Kneip and Sarda \(2011\)](#) that the standard sparsity models are sometimes too restrictive, so that it is advisable to combine them with other dimension reduction techniques. The “relevant” variables in a functional model are sometimes called “impact points” ([McKeague and Sen, 2010](#)) or “most predictive design points” ([Ferraty et al., 2010](#)). Also, the term “choice of components” has been used by [Delaigle et al. \(2012\)](#) as a synonym of variable selection.

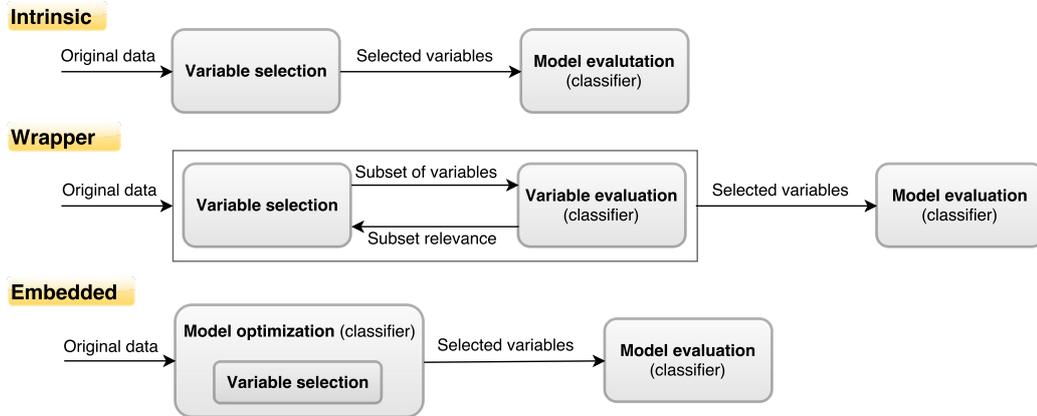


Figure 1.5: Flowcharts of different types of variable selection algorithms by evaluation criterion.

The monograph by [Guyon et al. \(2006\)](#) contains a complete survey on feature extraction (including selection) from the point of view of machine learning. It is organized around the results of a benchmark where several research groups competed on five large feature selection problems from different application domains. The second part is devoted to several specific methodologies used by the participants achieving the best results. The more recent book by [Liu and Motoda \(2012\)](#) provides the general background for variable selection and an overview of standard methods since the 70's for supervised and unsupervised classification, also in the machine learning framework. In [Saeys et al. \(2007\)](#), the authors make a complete review of supervised variable selection methods applied to bioinformatics. The overview paper by [Fan and Lv \(2010\)](#) has a more statistical orientation and [Arauzo-Azofra et al. \(2011\)](#) presents an interesting empirical comparison of several variable selection methods in the classification setting.

Without any exhaustiveness purpose in mind, we give here a short review of some of the principal aspects of variable selection methods in order to place our contributions in context. A variable selection algorithm is characterized by several essential features which allow us to establish different categories. These elements are mainly the search strategy, the measure of relevance and the evaluation criterion. Commonly, variable selection methods are classified according to the latter criterion, i.e., depending on the relation between the selection method and the predictor. Hence, variable selection algorithms are usually divided into three categories: intrinsic, wrapper and embedded ([Guyon et al., 2006](#); [Beniwal and Arora, 2012](#)). Figure 1.5 shows a schematic explanation of these approaches.

**Intrinsic** The methods we have called “intrinsic” are often denoted as “filter methods” in machine learning literature. In intuitive terms, the intrinsic methods aim at extracting (via variable selection) the information present in the data, independently of the use given to such data. Typically, variables are sorted by any relevance measure and those with lower score are removed. Thus, this approach is usually computationally simple, efficient and fast, so intrinsic methods readily scale high-dimensional problems. Since they are independent of the predictor they are more generalizable, that is, different classifiers can be evaluated with the same selected subset of variables. Also, the overfitting risk is smaller than in the other paradigms. However, this independence could be the main drawback of intrinsic techniques since they ignore any possible interaction with the classification rule. Besides the general references mentioned above, some intrinsic methods are reviewed in [Lazar et al. \(2012\)](#) for gene selection problems. A relevant example of intrinsic algorithms is the so called mRMR (minimum Redundancy Maximum Relevance) method, proposed by [Ding and Peng \(2005\)](#). See Chapter 4 for further details on this method along with a modified version and new applications.

**Wrapper** These are popular techniques since the publication of the paper by [Kohavi and John \(1997\)](#). Wrapper methods receive this name because the selection procedure “wraps” the predictor. The classification rule is used as a black box to assign scores to the different subsets of variables according to their discriminant power. In other words, the quality of a set of variables is directly measured by the performance of a predictor that only uses these variables. The algorithm carries out a double search, the first one considers all possible subsets of variables and then the classifier is estimated from each set. Hence, wrapper methods are computationally expensive and slow. In addition, they are not generalizable: the resulting selection is valid only for the considered classifier so, if different predictors must be studied, the whole process must be repeated for each predictor. The principal advantages of wrapper methods are that they take into account relations among variables in a natural way, and the connection with the predictor, which leads to better classification accuracy. However, this connection also entails a higher risk of overfitting than that of intrinsic methodologies. Many wrapper methods have been developed because of their good performance. SVM is one of the classifiers more utilized for these techniques methodology ([Maldonado and Weber, 2009](#)).

**Embedded** This third category is sometimes omitted or included as a special case of wrapper methods, but it represents a different approach. In this setup, variable selection and model estimation are performed simultaneously through the optimization of a target function. Then, embedded methods have closer connections between the selected variables and the predictor, so in this sense they

present same advantages and disadvantages as wrapper methods, but accentuated. Most interactions are taken into account, but embedded methods are totally dependent on the chosen model. However, mixing the selection process and the model estimation is usually less computationally expensive than the wrapper approach and can lead to a better use of the available data since it does not require to split the training sample for validation (Saeys et al., 2007). Maybe the most popular embedded methods are LASSO (Tibshirani, 1996) and variable selection via Random-Forest (Díaz-Uriarte and Alvarez de Andrés, 2006); see Scornet et al. (2015) for a general reference on random-forest with mathematical theory. Many other methodologies have been adapted to embedded variable selection, for example, SVM (Guyon et al., 2002).

**Hybrid** Finally, in the latest years there have appeared hybrid or two-steps methods which try to take advantage of the strengths of intrinsic and wrapper approaches avoiding their weaknesses. The general idea is to combine intrinsic and wrapper methods in two-steps algorithms. First, an intrinsic procedure is used in order to filter the informative variables by removing useless (and sometimes redundant) variables, and then a wrapper algorithm is applied to this reduced set; see for instance, Hua et al. (2009); Hsu et al. (2011).

We are especially interested in the “intrinsic” approaches to variable selection, in the sense that the final output should depend only on the data, not on any assumption on the underlying model (although the result should be interpretable in terms of the model).

Another fundamental issue in variable selection is how to decide whether a variable (or a set of variables) is relevant or not. Remember that our aim is to remove redundant or irrelevant variables in order to get the best classification performance with the smallest possible number of variables. There is not an universal definition for the relevance or the importance of a variable since this depends on each particular problem. However, some attempts to formalize these concepts have been done, see e.g. Yu and Liu (2004); Guyon et al. (2006). Roughly speaking, a group of variables is relevant in the classification setting if they have a high discriminant power, and irrelevant otherwise. Algorithms use different association measures  $I(\cdot, \cdot)$  to estimate the amount of discriminant information of a variable  $X$  in terms of its relation with the class  $Y$ . Thus, the usual relevance indicator is  $I(X, Y)$ . Standard correlation based measures were the first choice for  $I$  and they are still commonly used as relevance indices (Hall, 1999). Other options like the *gain ratio* (Quinlan, 1996), the *Gini index* (used by Breiman (2001) in random forest) or *Relief-F* (an extension of Relief to deal with sets of continu-

ous variables by [Robnik-Šikonja and Kononenko \(2003\)](#) ) are also popular among practitioners. A good summary of standard criteria of relevance can be found in [Guyon et al. \(2006, Chapter 3\)](#). Nowadays, the preferred relevance indices might be those based on distance between distributions. Two examples of this approach are considered in this thesis. The popular *mutual information* measure, which has led to a wide range of procedures [see [Vergara and Estévez \(2014\)](#) for a comprehensive survey], is commented in Chapter 4. Besides, we propose the use of the recent *distance correlation* measure ([Székely et al., 2007](#)) in two different ways in Chapters 3 and 4.

Depending on the use of the relevance measure, variable selection methods can be *univariate* or *multivariate* (in machine learning terminology). The former are usually called *ranking methods* since the variables are simply ranked by the relevance score  $I(X, Y)$ . In this setting, variables are considered separately, regardless of the classifier and ignoring any kind of dependence among them. These methods have been extensively used due to their simplicity, speed and good performance in a variety of problems ([Saeys et al., 2007](#); [Fan and Lv, 2010](#)). However, ranking methods do not take into account relationships among variables. In particular, they do not remove the redundancy among them, which is a critical point when dealing with functional data. On the other hand, the term “multivariate” stands for those methods which consider, in some sense, variable dependencies to overcome the redundancy problem. Also, some multivariate algorithms can take advantage of the positive interactions between variables (it is well known that some variables can be irrelevant individually but very informative when they work together). Some examples of this multivariate strategy are the methods based on mRMR ([Ding and Peng, 2005](#); [Ponsa and López, 2007](#)), which select iteratively the variables that maximize the relevance and minimize the redundancy at the same time, or the Correlation-based Feature Selection (CFS) by [Hall \(1999\)](#), that measures the correlation between pairs of variables. Another popular representatives of this approach are the Relief method and its versions for multiclass and regression problems, ReliefF and RRelief respectively ([Robnik-Šikonja and Kononenko, 2003](#); [Guyon et al., 2006](#)). The underlying idea in these methods is to take advantage of the nearest neighbours methodology to choose the best subset of variables.

The following example shows that the multivariate approach may be substantially better than the univariate one in the functional context. It refers to the well-known *Tecator data set* (a benchmark example very popular in the literature on functional data; see Section 5.3 for details). To be more specific, we use the first derivative of the curves in the Tecator data set, which is divided into two classes. We first use a simple ‘ranking procedure’ based on the mutual information, where

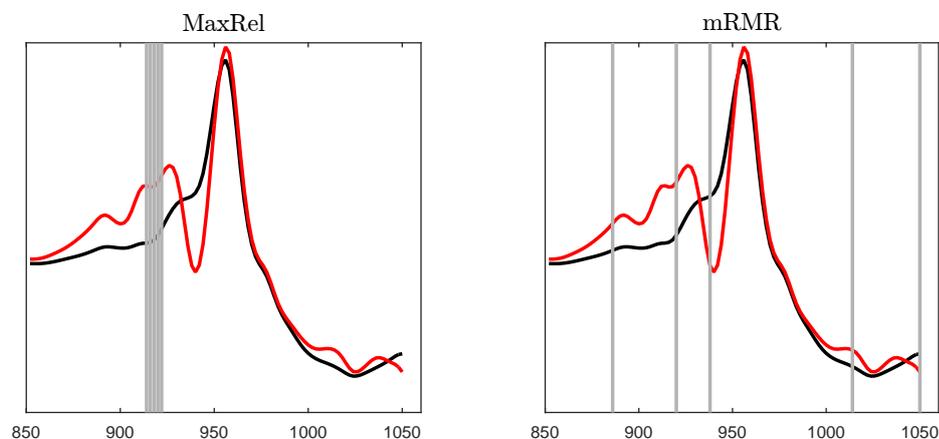


Figure 1.6: Mean functions for both classes considered in the Tecator data set (first derivative). Left panel shows the five variables selected by Maximum Relevance. Right panel corresponds to the variables selected by mRMR.

the variables are sequentially selected according to their relevance (thus ignoring any notion of redundancy). The result is shown in the left panel of Figure 1.6 (the selected variables are marked with grey vertical lines). In this case, the five selected variables provide essentially the same information. On the right panel we see the variables selected from the multivariate mRMR (with the same mutual information measure) procedure, which are clearly better placed to provide useful information. This visual impression is confirmed by comparing the error percentages obtained from a supervised classification method using only the variables selected by both methods. While the classification error obtained (using a 3-NN rule) with the variables selected by mRMR is 1.86%, the corresponding error obtained with those of the ranking method is 4.09%.

The last essential element in a variable selection technique is the search strategy. The search strategy defines how to explore the space of all possible combinations of variables until some stopping criterion is satisfied. Since the number of combinations is usually too large to carry out an exhaustive search (in fact, it is an NP-hard problem [Amaldi and Kann \(1998\)](#)), other strategies must be adopted. These methodologies are frequently heuristic and suboptimal, but allow us to deal with the problem and avoid the possible over-fitting related to the exhaustive search ([Reunanen, 2003](#)). Most search strategies correspond to the *forward selection* and *backward elimination* approaches, which sequentially add the most relevant or remove the least informative variable (according to some criterion), respectively. These approximations usually lead to the so-called *greedy algorithms*

which approximates the global optimum by a local one at each step and generate nested subsets of variables. The sequential forward search (SFS) and the analogous backward procedure (SBS) are two examples of this approach (Kittler, 1978). Some interesting alternatives are random searches where the space of variables is explored in an “arbitrary” way. Some relevant examples of this technique are genetic algorithms (Leardi et al., 1992) and simulated annealing (Brusco, 2014). See, e.g. Guyon et al. (2006); Liu and Motoda (2012) for further details and strategies. Besides, mixture approaches combining different strategies have been also considered, for example the “plus-1-take-away-r” methodology which combines forward and backward searches; see Vergara and Estévez (2014) for a survey focused on intrinsic methods.

The question of when to stop the search has not an unique answer so far. This is a sensitive issue, since too many or too less variables can affect heavily the results. There are numerous alternatives in the literature. Many times (specially in applied sciences) the number of variables is fixed arbitrarily based on cost, time or other reasons. For example, Golub et al. (1999) and Nguyen and Rocke (2002) uses the best 50 genes for their analysis. Other criteria can be a maximum number of loops (specially in random algorithms), a score threshold, to reach certain complexity level, etc.; see, e.g., Liu and Motoda (2012, Chapter 3) for further details and references. In this thesis we will use a standard validation step (via either cross-validation or using a validation sample) to set the number of variables (Guyon et al., 2006). Note however that we are not primarily concerned with the choice of the best number of variables but with establishing standard conditions in order to fairly compare the methods under study.

Many other aspects of variable selection methods could be considered, but they are beyond the scope of this work. We refer to the books and reviews cited in this section for further details and references.

### 1.4.3. Functional variable selection methods

In the functional setting, several relevant dimension reduction techniques are based upon the use of general finite dimensional projections. This is the case of functional principal component analysis (FPCA), see Li et al. (2013), although the so-called partial least squares (PLS) methodology is in general preferable when a response variable is involved; see Delaigle and Hall (2012b) for a recent reference on functional PLS. Functional PCA is adapted to sparse longitudinal data in Yao et al. (2005). Other common dimension reduction methods in the functional setting include sliced inverse regression (Hsing and Ren, 2009; Jiang and Liu, 2014) and additive models (Zhang et al., 2013). Also, the methods based on random pro-

jections could offer an interesting alternative. See, e.g., [Cuevas \(2014\)](#) for a short overview of dimension-reduction techniques together with additional references.

Nevertheless, we are concerned here with a different, more radical, approach to dimension reduction given by variable selection methods. As mentioned before, the aim of variable selection when applied to functional data is to replace every infinite dimensional observation  $x(t)$ ,  $t \in [0, 1]$  with a finite dimensional vector  $(x(t_1), \dots, x(t_k))$ . However, the reduction procedure must take into account the special characteristics of functional data, specially the high redundancy between close variables. Hence, the selection of the “variables”  $t_1, \dots, t_k$  should be a consequence of a trade-off between two mutually conflicting goals: representativeness and parsimony. In other words, we want to retain as much information as possible (thus selecting relevant variables) employing a small number of variables (thus avoiding redundancy).

Despite the huge amount of literature about variable selection for multivariate data, much less references are available when dealing with functional data, and most of them are centred in the linear regression framework. Today the most popular variable selection method among statisticians is perhaps the so-called *LASSO* procedure, proposed by [Tibshirani \(1996\)](#), as well as the *Dantzig selector*, a modification of *LASSO*, proposed by [Candes and Tao \(2007\)](#). These methods have a nice heuristic motivation and good theoretical properties; moreover, as shown by [Bickel et al. \(2009\)](#), they are asymptotically equivalent. Their application to the functional data setting has been analyzed by [Kneip and Sarda \(2011\)](#). Also, [Zhou et al. \(2013\)](#) adapt *SCAD* ideas for zero-coefficients to the functional setting. Other examples which also use  $L^1$  regularizations in this context are [Lee and Park \(2012\)](#); [Gertheiss et al. \(2013\)](#); [Zhao et al. \(2014\)](#). The Partial Linear Regression (PLR) model is extended to functional covariates by [Aneiros-Pérez and Vieu \(2006\)](#), and by [Ferraty and Romain \(2011, Chapter 3\)](#), where variable selection and estimation of coefficients are carried out simultaneously. Besides, let us also mention, with no attempt to be exhaustive, that the recent literature in functional variable selection includes a study of consistency in this setup ([Comminges and Dalalyan, 2012](#)). Also, [James et al. \(2009\)](#) gives an “interpretable” variable selection method which uses the derivatives of the function of coefficients jointly with a good and concise review of variable selection methods for high-dimensional linear regression. A wrapper procedure is proposed in the same linear regression framework by [Ferraty et al. \(2010\)](#). The “most predictive design points” are chosen to minimize the cross-validation error of a local linear regression method. The recent paper by [Aneiros and Vieu \(2014\)](#) introduces a two-stages procedure which uses the continuity of the functional predictors in order to get a better performance.

The application of these methods has been mostly developed for models. In fact, their formal implementation relies essentially on the assumed model. In the present work our approach to variable selection is slightly different, in the sense that we look for “intrinsic” methods based on the data and not relying on any assumption on a particular model (e.g. linear regression). Moreover, throughout this thesis we will consider variable selection in the setting of functional supervised classification (the extension to more general regression problems is also possible with some obvious changes). Surprisingly, variable selection for functional data classification is rarely addressed in both machine learning and statistical literature. Most of the existing references are focused on the classification of functional magnetic resonance images (Grosenick et al., 2008; Ryali et al., 2010) and near infrared spectra (Xiaobo et al., 2010). But in most cases curves are just treated as multivariate data for which the usual methods apply, and sometimes they are only used to extract some new (synthetic) variables which are the real objects in the variable selection phase. This happens, for example, in Gómez-Verdejo et al. (2009) which initially analyse the same problem as we do: variable selection for functional data classification. However, these authors propose to transform the functions into vectors of different artificial components extracted from the curves. Then, a new multivariate feature selection method is applied to these new high-dimensional observations. From an entirely functional point of view, Delaigle et al. (2012) provide a variable selection method for classification and clustering using the same approach of minimizing the cross-validation error as Ferraty et al. (2010). In both cases several heuristic adjustments are proposed in order to lighten the computational load typical of wrapper approaches. A new type of logistic regression model for binary classification of functional data is proposed by Lindquist and McKeague (2009), who consider Brownian-like predictors (e.g. fMRI and gene expression). A similar approach is followed in McKeague and Sen (2010) for fractional Brownian trajectories. Matsui (2014) introduces a  $L^1$ -penalized logistic model for multiclass classification.

Finally, some similar approaches should be mentioned even though they are not purely variable selection methods. For example, one can think on selecting intervals instead of points. In this way, the paper by Li and Yu (2008) provides a classification method for functional data based on “short curve segments”. Tian and James (2013) propose the selection of some basic “interpretable” functions to represent the curves (a sort of wavelet basis) before classification. These elements are chosen by minimizing the misclassification error through a stochastic search. An interpretable SVM-based classifier which allows us to consider “different levels of interpretability”, is provided by Carrizosa et al. (2011). Finally, a different approach is followed by Fraiman et al. (2015), where the selection is carried out after a “satisfactory” analysis of the data (regression, classification or principal

components). The goal is being able to “replicate” the result of the previous satisfactory analysis with a reduced dataset. This is obtained via several functions which capture relevant features of the original data (pointwise evaluation, local averages, moments, etc.). The selection is performed on a number of these special functions with a mixed search strategy (exhaustive and stochastic).

Our purpose in this thesis is to contribute to the study of the variable selection problem in a functional framework when classification is the final goal. On the one hand, a full theoretical motivation is given for these techniques, and on the other hand some new methodologies feasible for functional data are proposed. In particular, we present three intrinsic procedures for variable selection, i.e. not relying on any particular assumption on the dependence model. These methods have a sound functional motivation, and all of them adopt procedures to deal with the strong redundancy of the functional data sets. The use of the forward search strategy contributes to an easy comprehension and implementation, which is usually fast. All these methods are intrinsic, that is, they are (in principle) suitable for general problems and different classifiers. Despite these common features, the new proposed methods correspond to three different ideas. The first method, see [Berrendero et al. \(2015a\)](#), is called Reproducing Kernel Variable Selection (RKVS) and is based on the maximization of the Mahalanobis distance between the multivariate means corresponding to the selected variables of the two classes. It can be justified from an RKHS point of view and is explained in Chapter 2. Second, the Maxima-Hunting procedure (MH), see [Berrendero et al. \(2015c\)](#), relies only on the direct use of the distance correlation measure by [Székely et al. \(2007\)](#). This is fully described in Chapter 3. The last method, presented in Chapter 4, represents a modification of the mRMR algorithm, based on an idea we believe specially suitable for functional data. So we adapt this idea here, in combination with the distance correlation, for its use in the functional case ([Berrendero et al., 2015b](#)).

## 1.5. Contributions and structure of the thesis

This is concerned with the problem of supervised binary classification with functional data. We consider the functional data as trajectories drawn from a stochastic process. As a consequence, we will try 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.

In short, the original contributions in this work are the following:

- a) *General mathematical theory for the functional classification problem.* It is closely related with the reproducing kernel Hilbert space (RKHS) associated with the covariance operator of the processes.
  - a1) We provide some explicit expressions for the Bayes (optimal) rule and its corresponding error for the problem of classifying between absolutely continuous Gaussian processes.
  - a2) A complete mathematical treatment is given for the case of the classification between mutually singular processes, which corresponds to the so-called near perfect classification phenomenon.
- b) *Functional variable selection.*
  - b1) A general theoretical motivation (expressed in terms of a sparsity assumption) is given for the problems of functional variable selection.
  - b2) We propose 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.

The papers [Berrendero et al. \(2015b\)](#) and [Berrendero et al. \(2015c\)](#) correspond respectively to the mRMR-RD and MH proposals mentioned above. The contributions of point *a*) and the RK-VS method are essentially included in the manuscript [Berrendero et al. \(2015a\)](#). A more detailed explanation of this outline is given in the next subsections.

### 1.5.1. Contributions

The contributions of this thesis are threefold. Firstly, the supervised classification of functional data is tackled in a relatively unexplored point of view. Problems are characterized in terms of the absolute continuity or mutual singularity of the underlying probability measures, which entails some intriguing consequences in the functional classification setting. In the absolutely continuous case  $P_1 \ll P_0$ , some classical results concerning calculation of Radon-Nikodym derivatives for probability measures in function spaces are used in order to obtain and interpret explicit expressions for the optimal classifier in some binary functional discrimination problems of practical interest. This approach leads to a new class of plug-in classifiers. In some relevant cases the optimal rules turn out to depend on a finite

number of variables, so that the use of variable selection methods is especially appropriate. These results provide a theoretical basis for the techniques of variable selection in functional classification models. Usually these methods are considered in the literature from an exclusively algorithmic or computational point of view. Therefore, it is of some interest to motivate them in “population terms”, by identifying some specific models where these techniques have a clear meaning. The present dissertation can be seen as a partial contribution to this kind of motivations.

We also consider the mutually singular case  $P_0 \perp P_1$ , i.e., when there exist a Borel set  $A$  such that  $P_0(A) = 1$  and  $P_1(A) = 0$ . Note that this mutually singular (or “orthogonal”) case is rarely found in the finite-dimensional classification setting, except in a few trivial or artificial cases. However, in the functional framework the singular case is an important general situation. We show that this mutual singularity notion is behind the near perfect classification phenomenon described in [Delaigle and Hall \(2012a\)](#). The point is to look at this phenomenon from the slightly different (coordinate free) RKHS perspective. We also show that an approximately optimal (“near perfect”) classification rule to discriminate between  $P_0$  and  $P_1$  when  $P_1 \perp P_0$ , can be obtained in terms of the optimal rules corresponding to a sequence of problems  $(P_0^n, P_1^n)$  with  $P_1^n \ll P_0^n$ .

Second, we offer three new intrinsic methods for variable selection preceding functional discrimination. In the first place we propose a flexible RKHS-based variable selection mechanism which arises in a natural way from the theoretical framework. Unlike other popular variable selection methods in classification, this new proposal allows the user to incorporate, in a flexible way, different amounts of information (or assumptions) on the underlying model. We also provide a consistent closely related linear classifier. An empirical study shows that both the variable selection method and the associated classifier, perform very well and are clearly competitive with respect to existing competing alternatives. We also argue, as an important additional advantage, the simplicity and ease of interpretation of the RKHS-based procedures.

The second proposed method is based on a direct use of the distance covariance measure  $\mathcal{V}^2$ , or alternatively the distance correlation measure  $\mathcal{R}^2$ , proposed by [Székely et al. \(2007\)](#): we just propose to select the variables indices  $t_1, \dots, t_k$  in the functional data  $X(t)$ ,  $t \in [0, 1]$ , which correspond to local maxima of these distance covariance/correlation functions between  $X(t)$  and the response variable  $Y$ . So one always know the exact meaning of the selected variables: they are just those locally maximizing the dependence with the response variable. We will motivate this “maxima hunting” (MH) idea via some theoretical models for which

the optimal (Bayes) rule turns out to be explicitly known. The practical implementation of the method, for a given data set, arises as a result of the estimation of  $\mathcal{V}^2(X_t, Y)$  or  $\mathcal{R}^2(X_t, Y)$ . This is backed by a consistency result and we also provide some new alternative versions of the distance covariance measure when  $Y$  is a binary variable. An exhaustive empirical study shows the good performance of this new approach when compared with other competitors.

The third proposed selection method is a modified version of the mRMR algorithm by [Ding and Peng \(2005\)](#), in which the association between the variables is calculated by means of different measures. We consider several versions of the mRMR and compare them by an extensive empirical study. Two of these versions are new: they are based on the distance covariance and distance correlation association measures commented above. Our results suggest that the new version based on the distance correlation measure represents a clear improvement of the mRMR methodology in the functional framework. This method has been also successfully applied in a real problem for discriminating mice with different sex and diet from their NMR spectral fingerprints ([Barba et al., 2015](#)).

Additionally to these new proposals and results, the third major goal of this work is to provide an extensive and replicable empirical study aimed at assessing the performance (in the setting of binary functional classification) of several intrinsic variable selection methods. In this empirical study (which includes a large number of simulations and a few real data examples) the variable selection methods are viewed as particular instances of the dimension reduction methodology. Thus we have included in the comparisons different selection techniques and the PLS method as a reference, since it is, by now, the most usual procedure for dimension reduction in functional data analysis before discrimination. In all cases the classifiers under study are chosen as a sort of all-purposes benchmark methods. Of course other functional classifiers could be considered but since the study is centred on intrinsic variable selection methods, the interpretability of the results would be greatly complicated if new “variables” (in this case, different classifiers or tuning parameters) were introduced. To our knowledge, this is the largest empirical study on variable selection so far.

Finally, as a consequence of all the revision work we have carried out, this thesis presents a general overview of variable selection in a functional classification framework. Also some empirical result and details about some well known real data sets have been summarized.

## 1.5.2. Structure

This thesis is organized in six chapters and one appendix. After this introductory chapter, which states the problem and reviews the general literature about the related topics, the next three chapters contain the main contributions of this work. Roughly speaking, each chapter corresponds to a different approach to the variable selection problem when classifying functional data.

In Chapter 2 we study the supervised classification of functional data from the novel point of view based on the RKHS theory. Some background on this theory is given in Section 2.1. As a consequence of this approach, we provide theoretical motivation for variable selection when classifying Gaussian processes, as well as shed some light on some phenomenons around functional classification. First, the explicit expressions of the optimal rules for the case of equivalent processes (Thm. 2.2) can be found in Section 2.2. Second, the mutual singular setting is considered in the next section: this orthogonality is shown to be behind the near perfect classification phenomenon (Thm. 2.4) and an approximately optimal classification rule for this case is derived in Theorem 2.5. In the third place, the RKHS-based variable selection method and the related classifier are proposed in Section 2.4. A consistency result for RKHS-based classifier is given in Thm. 2.6. The performance of these new techniques is assessed in Section 2.5. Section 2.7 contains all the proofs of this chapter.

The maxima hunting method is presented in Chapter 3. In Section 3.1 we provide a survey of the main ideas and results concerning the distance covariance and distance correlation measures. Some useful simplified versions for  $\mathcal{V}^2$  are obtained in Theorem 3.1, for the particular case where  $Y$  is a binary variable. The maxima hunting method for variable selection is described in Section 3.2. Results of consistent estimation (Thm. 3.2, Lemma 3.1) for the maxima of  $\mathcal{V}^2$  is also proved in that section. In Section 3.3 we give several models (identified in terms of the conditional distributions  $X(t)|Y = j$ ) for which the optimal classification rule depends only on a finite number of variables. We also show that in some of these models the variables to be selected coincide with the maxima of  $\mathcal{V}^2$ . Finally, some empirical results with both simulation and real data examples are given in Section 3.4 together with a brief discussion. All the proofs of this chapter are in Section 3.6.

Chapter 4 corresponds to the new version of the minimum Redundancy Maximum Relevance (mRMR) algorithm. Section 4.1 contains a summary and some remarks about the mRMR methodology. The different association measures under study (which are used to define the different versions of the mRMR method)

are explained in Subsection 4.1.1. The empirical study, consisting of simulation experiments and some representative real data sets, is explained in Section 4.2. Finally, the real application is described in section 4.3 and some conclusions are given.

Extensive simulation studies, comparing our variable selection methods with other dimension reduction procedures (as well as with the “baseline option” of doing no variable selection at all) have been tackled for all the new methods. Some particular results are given in the corresponding chapters but the general methodology, the considered models and other empirical issues are detailed in Chapter 5 aiming at not duplicating information and making the study replicable by interested researchers. The methods under study and their implementation details are commented in Section 5.1. The simulation study is fully explained in Section 5.2, including the description of the models and the followed methodology. Four real data sets are discussed in Section 5.3 as well as the methodological differences. This Chapter also includes some additional outputs of our new proposals.

Finally, Chapter 6 contains some general conclusions that can be extracted from this work together with some potential lines of future research.

The full list of simulation models is in Appendix A.



## Chapter 2

# RKHS-based functional classification

Functional data classification has been always studied in the usual spaces of functions (essentially  $L^2[a, b]$ ,  $C[a, b]$  and some semi-normed spaces). In this chapter we propose to tackle the problem from the Reproducing Kernel Hilbert Spaces (RKHS) associated with the covariance function of the underlying process which generates the data. RKHS theory was first applied by Parzen (1961) in signal detection problems, where no samples were involved and the approach is quite different to that we will follow here. However, this and other works (Parzen, 1962; Kailath, 1971; Segall and Kailath, 1975) developed a collection of tools which are also very useful in our functional discrimination problem. But above all, the RKHS view opens the door to a new manner of dealing with stochastic processes in different spaces that are intrinsically connected with the processes and the classification problem. In words of Parzen (1961): “*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 series*”. Note that Parzen uses here the expression “time series” as a synonym of stochastic processes, which is rather unusual in the modern statistical terminology.

This chapter begins with a brief review of some relevant background about the RKHS theory related with the classification problem. Some important results for this chapter about characterizations of the probability measures (equivalence or mutually singularity) and expressions of the Radon-Nikodym derivatives (in the absolutely continuous case) for Gaussian processes are given in Subsection 2.1.2. The absolutely continuous case is considered in Section 2.2, where the previous results are used to derive explicit expressions of the optimal classification rule (Thm. 2.2) through the Equation (1.2). Section 2.3 is devoted to mutually sin-

gular measures. Here, the near perfect classification result by [Delaigle and Hall \(2012a\)](#) is explained in terms of the orthogonality of the distributions (Thm. 2.4) and Theorem 2.5 provides a method to approximate the Bayes rule in the singular framework by means of a sequence of approximating (absolutely continuous) problems. In Section 2.4 the RKHS approach is used in order to define a theoretical framework for variable selection under a reasonable sparsity assumption. This fact is exploited by constructing a flexible RKHS-based variable selection method and an associated classification rule in Subsection 2.4.2. These methods are backed by a consistency result (Thm. 2.6) and the good results of the experiments of Section 2.5. The possibility of adding extra information along with other issues are also explored in Subsection 2.4.3. Finally, some conclusions are drawn in Section 2.6 and all the proofs can be found in Section 2.7.

## 2.1. Radon-Nikodym densities for Gaussian processes: some background

It can be seen from the introductory section that the supervised classification problem can be formally stated, with almost no formal difference, either in the ordinary finite-dimensional situation (where  $X$  takes values on the Euclidean space  $\mathbb{R}^n$ ) or in the functional case (where  $X$  is a stochastic process). We have also seen that in spite of these formal analogies, the passage to an infinite-dimensional (functional) sample space  $\mathcal{F}$  entails some important challenges (see Section 1.2). For example, the classical Fisher linear rule, which is still very popular in the finite-dimensional setting, cannot be easily adapted to the functional case (see, Section 1.2 for more details). However, we are more concerned here with another crucial difference, namely the lack of a natural “dominant” measure in functional spaces, playing a similar role to that of Lebesgue measure in  $\mathbb{R}^n$ . If we are working with Gaussian measures in  $\mathbb{R}^n$ , the optimal rule (1.1) can be established in terms of the ordinary (Lebesgue) densities of  $P_0$  and  $P_1$ . Nevertheless, in the functional case we are forced to work with the “mutual” Radon-Nikodym derivatives  $dP_1/dP_0$ , provided that  $P_1 \ll P_0$  ([Baíllo et al., 2011](#)). Usually these derivatives are not easy to calculate or to work with. However, the good news is that in some relevant cases they are explicitly known and reasonably easy to handle. See [Baíllo et al. \(2011\)](#) and [Cadre \(2013\)](#) for some recent statistical applications of the Radon-Nikodym densities. In the following paragraphs we review, for posterior use, some results regarding the explicit calculation of Radon-Nikodym derivatives of Gaussian processes.

### 2.1.1. RKHS

We first need to recall some very basic facts on the theory of Reproducing Kernel Hilbert Spaces (RKHS); see [Berlinet and Thomas-Agnan \(2004\)](#), [Janson \(1997, Appendix F\)](#) for extra background.

Given a symmetric positive-semidefinite function  $K(s, t)$ , defined on  $[0, T] \times [0, T]$  (in our case  $K$  will be typically the covariance function of a process), let us define the space  $\mathcal{H}_0(K)$  of all real functions which can be expressed as finite linear combinations of type  $\sum_i a_i K(\cdot, t_i)$  (i.e., the linear span of all function  $K(\cdot, t)$ ). In  $\mathcal{H}_0(K)$  we consider the inner product  $\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_j, t_i)$ , where  $f(x) = \sum_i \alpha_i K(x, t_i)$  and  $g(x) = \sum_j \beta_j K(x, s_j)$ .

Then, 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  $f : [0, T] \rightarrow \mathbb{R}$  which can be obtained as  $t$  pointwise limit of a Cauchy sequence  $\{f_n\}$  of functions in  $\mathcal{H}_0(K)$ . The theoretical motivation for this definition is the well-known Moore-Aronszajn Theorem (see [Berlinet and Thomas-Agnan \(2004\)](#), p. 19). The functions in  $\mathcal{H}(K)$  have the “reproducing property”  $f(t) = \langle f, K(\cdot, t) \rangle_K$ .

If  $\{X(t), t \in [0, T]\}$  is an  $L^2$ -process (i.e.  $\mathbb{E}(X_t^2) < \infty$ , for all  $t$ ) with covariance function  $K(s, t)$ , the natural Hilbert space associated with this process,  $\bar{\mathcal{L}}(X)$  is the closure (in  $L^2$ ) of the linear span  $\mathcal{L}(X) = \mathcal{L}(X_t, t \in [0, T])$ . The so-called *Loève Representation Theorem* ([Berlinet and Thomas-Agnan, 2004](#), p. 65) establishes that the spaces  $\mathcal{L}(X)$  and  $\mathcal{H}(K)$  are *congruent*. More precisely, the natural transformation  $\Psi(\sum_i a_i X_{t_i}) = \sum_i a_i K(\cdot, t_i)$  defines in fact, when extended by continuity, a congruence (that is an isomorphism which preserves the inner product) between  $\bar{\mathcal{L}}(X)$  and  $\mathcal{H}(K)$ .

Two interesting consequences of Loève’s result are: first, if a linear map  $\phi$ , from  $\bar{\mathcal{L}}(X)$  to  $\mathcal{H}(K)$ , fulfils  $\mathbb{E}(\phi^{-1}(h)X_t) = h(t)$ , for all  $h \in \mathcal{H}(K)$ , then  $\phi$  coincides with the congruence  $\Psi$  which maps  $X_t$  to  $K(t, \cdot)$ . Second,  $\mathcal{H}(K)$  coincides with the space of functions which can be defined in the form  $h(t) = \mathbb{E}(X_t U)$ , for some  $U \in \bar{\mathcal{L}}(X)$ .

Thus, in a very precise way,  $\mathcal{H}(K)$  can be seen as the “natural Hilbert space” associated with a process  $\{X(t), t \in [0, T]\}$ . In fact, as we will next see, the space  $\mathcal{H}(K)$  is deeply involved in some relevant probabilistic and statistical notions.

### 2.1.2. RKHS and Radon-Nikodym derivatives. Parzen's Theorem

The following result is a slightly simplified version of Theorem 7A in [Parzen \(1961\)](#); see also [Parzen \(1962\)](#). It will be particularly useful in the rest of this chapter.

**Theorem 2.1** (Parzen 1961, Thm. 7A). *Let us denote by  $P_1$  the distribution of a Gaussian process  $\{X(t), t \in [0, T]\}$ , with continuous trajectories, mean function denoted by  $m = m(t) = \mathbb{E}(X(t))$  and continuous covariance function denoted by  $K(s, t) = \text{Cov}(X(s), X(t))$ . Let  $P_0$  be the distribution of another Gaussian process with the same covariance function and with mean function identically 0. Then,  $P_1 \ll P_0$  if and only if the mean function  $m$  belongs to the space  $\mathcal{H}(K)$ . In this case,*

$$\frac{dP_1}{dP_0}(X) = \exp\left(\langle X, m \rangle_K - \frac{1}{2}\langle m, m \rangle_K\right). \quad (2.1)$$

In the case  $m \notin \mathcal{H}(K)$ , we have  $P_1 \perp P_0$ .

*Some important remarks on this result.*

- (a) Note that, except for trivial cases, the trajectories  $x$  of the process  $X(t)$  are not included, with probability one, in  $\mathcal{H}(K)$ ; see, e.g., ([Berlinet and Thomas-Agnan, 2004](#), p. 66) and [Lukić and Beder \(2001\)](#) for details. Thus, the expression  $\langle X, m \rangle_K$  is somewhat of an abuse of notation. It is formally defined (a.s.) as the random variable  $\Psi^{-1}(m)$ , where  $\Psi^{-1}$  is the inverse of the above defined congruence  $\Psi : \bar{\mathcal{L}}(X) \rightarrow \mathcal{H}(K)$  which maps  $X_t$  to  $K(t, \cdot)$ . The following expressions (see [Parzen \(1961\)](#), p. 974)) will be particularly useful in our calculations. Let  $m \in \mathcal{H}(K)$  be the real mean of  $X$ , for every  $t \in [0, T]$ , and  $h, g \in \mathcal{H}(K)$ ,

$$\begin{aligned} \langle X, K(\cdot, t) \rangle_K &= X(t) \\ \mathbb{E}(\langle X, h \rangle_K) &= \langle m, h \rangle_K \\ \text{Cov}(\langle X, h \rangle_K, \langle X, g \rangle_K) &= \langle h, g \rangle_K \end{aligned}$$

- (b) In the case where  $X(t) = B(t)$  is the standard Brownian Motion,  $K(s, t) = \min(s, t)$ . Then, it can be seen that  $\mathcal{H}(K)$  coincides with the so-called Dirichlet space  $\mathcal{D}[0, T]$  of those real functions  $g$  on  $[0, T]$  such that there exists  $g'$  almost everywhere in  $[0, T]$  with

$$g' \in L^2[0, T], \text{ and } g(t) = \int_0^t g'(s) ds. \quad (2.2)$$

The norm in  $\mathcal{D}[0, T]$  is defined by  $\|g\|_K = \left( \int_0^T g^2(t) dt \right)^{1/2}$ . Likewise, the inverse congruence  $\langle X, m \rangle_K$  can also be expressed as the stochastic integral  $\int_0^T m'(s) dB(s)$ .

Thus, Theorem 2.1 can be seen as an extension of the classical Cameron-Martin Theorem (Mörters and Peres, 2010, p. 24), for  $X(t) = B(t)$ . It also coincides with Shepp (1966, Thm. 1), when applied to the homoscedastic case in which  $P_0$  and  $P_1$  are the distributions of  $X(t)$  and  $m(t) + X(t)$ , respectively.

- (c) Some additional references on the topic of Radon-Nikodym derivatives in function spaces are Varberg (1961, 1964), Kailath (1971) and Segall and Kailath (1975), among others.

## 2.2. Absolutely continuous Gaussian processes

In this section we consider the supervised classification problem, as stated in Section 1.2, under the following general model

$$\begin{cases} P_0 : m_0(t) + \epsilon_0(t) \\ P_1 : m_1(t) + \epsilon_1(t) \end{cases}, \quad (2.3)$$

where, for  $i = 0, 1$ ,  $\{\epsilon_i(t), t \in T\}$  are “noise processes” with mean 0 and continuous trajectories, and  $m_i(t)$  are some continuous functions, defining the respective “trends” of  $P_0$  and  $P_1$ . The following result provides the Bayes (optimal) rule and the corresponding minimal error probability for this case, under the usual assumption of homoscedasticity

**Theorem 2.2** (Bayes Rule for homoscedastic Gaussian problems). *In the classification problem under the model (2.3) assume*

- (a) *the noise processes  $\epsilon_i$  are both Gaussian with common continuous covariance function  $K(s, t)$ .*
- (b)  *$m_1 - m_0 \in \mathcal{H}(K)$ , where  $\mathcal{H}(K)$  denotes the reproducing kernel Hilbert space associated with  $K$ ; we denote  $m_1 - m_0 \equiv m$ .*

*Then, the optimal Bayes rule is given by  $g^*(X) = \mathbb{I}_{\{\eta^*(X) > 0\}}$ , where*

$$\eta^*(x) = \langle x - m_0, m \rangle_K - \frac{1}{2} \|m\|_K^2 - \log \left( \frac{1-p}{p} \right), \quad (2.4)$$

*$\|\cdot\|_K$  denotes the norm in the space  $\mathcal{H}(K)$ ,  $p = \mathbb{P}(Y = 1)$  and  $\langle x - m_0, m \rangle_K$  stands for the congruence defined in the Remark (a) of Theorem 2.1.*

Also, the corresponding optimal classification error  $L^* = \mathbb{P}(g^*(X) \neq Y)$  is

$$L^* = (1-p)\Phi\left(-\frac{\|m\|_K}{2} - \frac{1}{\|m\|_K} \log\left(\frac{1-p}{p}\right)\right) + p\Phi\left(-\frac{\|m\|_K}{2} + \frac{1}{\|m\|_K} \log\left(\frac{1-p}{p}\right)\right),$$

where  $\Phi$  is the cumulative distribution function of a standard normal random variable. When  $p = 1/2$ , we have  $L^* = 1 - \Phi\left(\frac{\|m\|_K}{2}\right)$ .

While this theorem has interest on its own, we will mainly use it as an important auxiliary tool in the rest of the thesis. In particular, it will be used in the calculation of an approximate optimal rule for the singular case (see Section 2.3 below) and will be also the basis for the variable selection method we propose in Section 2.4.

### 2.3. Classification of singular Gaussian processes: another look at the “near perfect classification” phenomenon

The starting point in this section is again the classification problem between the Gaussian processes  $P_0$  and  $P_1$  defined in (2.3), with  $\epsilon_0$  and  $\epsilon_1$  identically distributed according to the Gaussian process  $\epsilon(t)$  with continuous covariance function  $K$ , and the mean functions are  $m_0(t) = 0$  and  $m_1(t) = \sum_{j=1}^{\infty} \mu_j \phi_j(t)$ , where the  $\phi_j$  are the eigenfunctions of the Karhunen-Loève expansion of  $K$ , that is,

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

Let us assume for simplicity that the prior probability is  $\mathbb{P}(Y = 1) = 1/2$ . This model has been considered by [Delaigle and Hall \(2012a\)](#). In fact, these authors solve completely the classification problem since they provide the explicit expression of the optimal rule. In addition, they find that, under some conditions on the coefficients  $\theta_j$  and  $\mu_j$ , the classification is “near perfect” in the sense that one may construct a rule with an arbitrarily small probability of classification error. To be more specific, the classification rule they proposed is the so-called “centroid classifier”,  $T$ , defined by

$$T(X) = 1 \text{ if and only if } D^2(X, \bar{X}_1) - D^2(X, \bar{X}_0) < 0, \quad (2.5)$$

where  $\bar{X}_0, \bar{X}_1$  denote the sample means of the training sample observations from  $P_0$  and  $P_1$ ,  $D(X, Z) = |\langle X, \psi \rangle_{L^2} - \langle Z, \psi \rangle_{L^2}|$ , with  $\langle X, \psi \rangle_{L^2} = \int_0^1 X(t)\psi(t)dt$  and

$$\psi(t) = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \phi_j(t), \quad (2.6)$$

provided that this series is convergent. The asymptotic version of the classifier (2.5) under the assumed model with  $m_0 = 0$  is

$$T^0(X) = 1, \text{ if and only if } (\langle X, \psi \rangle_{L^2} - \langle m_1, \psi \rangle_{L^2})^2 - \langle X, \psi \rangle_{L^2}^2 < 0, \quad (2.7)$$

assuming again the convergence in (2.6).

Now, a more precise summary of the above discussion is as follows.

**Theorem 2.3** (Delaigle and Hall 2012a, Thm.1). *Let us consider the binary classification problem (2.3) under the Gaussian homoscedastic model with  $m_0(t) = 0$  and continuous covariance function  $K$ , as described at the beginning of this section.*

- (a) *If  $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 < \infty$ , the minimal (Bayes) misclassification probability is given by  $err_0 = 1 - \Phi\left(\frac{1}{2}(\sum_{j \geq 1} \theta_j^{-1} \mu_j^2)^{1/2}\right)$  and the optimal classifier (that achieves this error) is the rule  $T^0$  defined in (2.7).*
- (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  $\psi$  with  $\psi^{(r)} = \sum_{j=1}^r \theta_j^{-1} \mu_j \phi_j(t)$ , with  $r = r_n \uparrow \infty$ .*

As pointed out in Delaigle and Hall (2012a), “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.”

Our purpose here is to show that the setup of Theorem 2.3 can be analysed from the point of view of RKHS theory, in such a way that the situation considered in part (a) corresponds to the absolutely continuous case  $P_1 \ll P_0$  (that is,  $P_1 \sim P_0$  in the Gaussian case) considered in Theorem 2.1, and part (b) corresponds to

the singular case  $P_0 \perp P_1$ . In other words, in the absolutely continuous case, we can calculate the explicit expression of the optimal rule. It can be expressed in terms on the Radon-Nikodym derivative  $dP_1/dP_0$  but we will also show that the corresponding expression coincides with that given by [Delaigle and Hall \(2012a\)](#) in terms of eigenvalues and eigenfunctions. Also, condition in part (b) of [Theorem 2.3](#) is equivalent to  $P_1 \perp P_0$ . This sheds some light, in probabilistic terms, on the “near perfect classification” phenomenon. These ideas are made concrete in the following result

**Theorem 2.4** (Another view on near perfect classification). *In the framework of the classification problem considered in [Theorem 2.3](#), we have*

(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

$$g^*(X) = 1 \text{ if and only if } \langle X, m \rangle_K - \frac{1}{2} \|m\|_K^2 > 0, \quad (2.8)$$

with the notation of [Equation \(2.4\)](#). This rule is a coordinate-free, equivalent expression of the optimal rule given in [Theorem 2.3 \(a\)](#). The corresponding optimal (Bayes) classification error is  $L^* = 1 - \Phi(\|m\|_K/2)$ .

(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  $L^* = 0$ .

We next make explicit the meaning of the near perfect classification phenomenon.

**Theorem 2.5** (Singular case classifier). *Again, in the singular case considered in [Theorem 2.4](#), the following statement holds: given any  $\epsilon > 0$ , we can construct a classification rule whose misclassification probability is smaller than  $\epsilon$ .*

## 2.4. A model-based proposal for variable selection and classification

### 2.4.1. RKHS and variable selection

We have seen in [Section 2.3](#) how the RKHS framework gives insight into the near perfect classification phenomenon. In this section we argue that it also offers a natural setting to formalize variable selection problems. Variable selection methods are quite appealing when classifying functional data since they help reduce noise and remove irrelevant information. Classification performance often

improves if instead of employing the whole data trajectories we only use their values at carefully selected points. 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. By contrast, the usual  $L^2[0, T]$  space lacks functions playing a similar role. Here, we take advantage of this fact to establish a simple condition under which only a few points of the trajectory we observe are relevant for classifying it. Then, we propose a method of variable selection to identify the relevant points. As we will see, the method is motivated by the expressions of Radon-Nikodym derivatives and optimal rules we have derived in the previous sections. In fact, we will see that our method for identifying the relevant points also yields a natural procedure for estimating the optimal rule.

Recall the general model (2.3) and observe that by Theorem 2.2, if  $m = m_1 - m_0 \in \mathcal{H}(K)$ , then the optimal rule to classify a trajectory  $x$  between  $P_0$  and  $P_1$  is  $g^*(x) = \mathbb{I}_{\{\eta^*(x) > 0\}}$ , where  $\eta^*(x)$  is given in Equation (2.4). The following condition will be important for the remainder of this section:

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

Note that this assumption is not very restrictive since the finite combinations of type  $\sum_{i=1}^d \alpha_i^* K(\cdot, t_i^*)$  are dense in the RKHS.

It turns out that, under this assumption, the Bayes rule depends on the trajectory  $x(t)$  only through the values  $x(t_1^*), \dots, x(t_d^*)$ . Indeed, the discriminant score  $\eta^*(x)$  of a trajectory  $x$  is given by:

$$\begin{aligned} \eta^*(x) &= \langle x - m_0, \sum_{i=1}^d \alpha_i^* K(\cdot, t_i^*) \rangle_K - \frac{1}{2} \left\| \sum_{i=1}^d \alpha_i^* K(\cdot, t_i^*) \right\|_K^2 - \log \left( \frac{1-p}{p} \right) \\ &= \sum_{i=1}^d \alpha_i^* \langle x - m_0, K(\cdot, t_i^*) \rangle_K - \frac{1}{2} \left\langle \sum_{i=1}^d \alpha_i^* K(\cdot, t_i^*), \sum_{j=1}^d \alpha_j^* K(\cdot, t_j^*) \right\rangle_K - \log \left( \frac{1-p}{p} \right) \\ &= \sum_{i=1}^d \alpha_i^* (x(t_i^*) - m_0(t_i^*)) - \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d \alpha_i^* \alpha_j^* K(t_i^*, t_j^*) - \log \left( \frac{1-p}{p} \right), \end{aligned}$$

where we have used the reproducing property to obtain the last equality.

A more familiar expression for the optimal rule is obtained taking into account that [SA] implies the following relationship between  $\alpha_1^*, \dots, \alpha_d^*$  and  $t_1^*, \dots, t_d^*$ :

$$m_{t_1^*, \dots, t_d^*} = K_{t_1^*, \dots, t_d^*} \cdot (\alpha_1^*, \dots, \alpha_d^*)^\top, \quad (2.9)$$

where  $u^\top$  denote the transpose of  $u$ ,  $K_{t_1, \dots, t_d}$  is the  $d \times d$  matrix whose  $(i, j)$  entry is  $K(t_i, t_j)$ , and  $m_{t_1, \dots, t_d} = (m(t_1), \dots, m(t_d))^\top$ . Using (2.9) we can write

$$\eta^*(x) = \sum_{i=1}^d \alpha_i^* \left( x(t_i^*) - \frac{m_0(t_i^*) + m_1(t_i^*)}{2} \right) - \log \left( \frac{1-p}{p} \right), \quad (2.10)$$

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

This shows that under [SA], the Bayes rule coincides with the well-known Fisher linear rule based on the projections  $x(t_1^*), \dots, x(t_d^*)$ . This conclusion could be expected since we are dealing with finite dimensional homoscedastic Gaussian distributions. The above discussion just provides an independent derivation within the RKHS setup.

*Remark 2.1* (Sparsity example). A simple example for which the sparsity assumption holds is the following: consider model (2.3) where  $\epsilon_0$  and  $\epsilon_1$  are Brownian motions,  $m_0 \equiv 0$  and  $m_1$  is a continuous, piecewise linear function such that  $m_1(0) = 0$ . According to the computations above, the discriminant score of a trajectory  $x(t)$  only depends on the values of  $x$  at the points where  $m_1$  is not differentiable (and, possibly, also on  $x(0)$  and  $x(T)$ ). This can be more easily derived from the representation of the discriminant scores in terms of stochastic integrals (see Subsection 2.1.2, remark (b)).

## 2.4.2. An RKHS-based criterion for variable selection and its associated Fisher rule

Assume hereafter that [SA] holds and that we observe independent random samples  $X_{j,1}, \dots, X_{j,n_j}$  of trajectories from the model  $P_j$ , for  $j = 0, 1$ . Our first goal is to use the training samples for identifying a set of  $d$  points close to  $t_1^*, \dots, t_d^*$ , the only relevant points for the classification problem. In view of the observations of the previous paragraphs, once  $d$  relevant points  $\hat{t}_1, \dots, \hat{t}_d$  have been identified, the Fisher rule based on them is the natural estimator of the optimal classification rule. Hence, our RKHS approach, combined with the sparsity assumption [SA], leads us to both a natural variable selection method and a classification procedure based on the selected variables. We next develop this idea.

From the expression for the Bayes error  $L^*$  we gave in Theorem 2.2, it is easy to see that  $L^*$  is a monotone decreasing function of  $\|m\|_K$ . Moreover, under [SA]

and using (2.9),

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

Then, if we knew  $m$ , we would choose the points maximizing  $\psi(t_1, \dots, t_d) := m_{t_1, \dots, t_d}^\top K_{t_1, \dots, t_d}^{-1} m_{t_1, \dots, t_d}$ . Since  $m$  is unknown, we propose to replace it by its obvious estimator  $\hat{m}(t) = \hat{m}_1(t) - \hat{m}_0(t)$ , where  $\hat{m}_j(t) := n_j^{-1} \sum_{i=1}^{n_j} X_{1,j}(t) = \bar{X}_j(t)$ , for  $j = 0, 1$ . The criterion we suggest for variable selection is to choose points  $\hat{t}_1, \dots, \hat{t}_d$  such that  $\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) \geq \hat{\psi}(t_1, \dots, t_d)$  for all  $t_1, \dots, t_d$ , where

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

Notice that  $\hat{\psi}(t_1, \dots, t_d)$  is the Mahalanobis distance between the mean vectors  $(\bar{X}_0(t_1), \dots, \bar{X}_0(t_d))$  and  $(\bar{X}_1(t_1), \dots, \bar{X}_1(t_d))$  relative to the covariance matrix of the finite dimensional distribution corresponding to  $t_1, \dots, t_d$ .

Given the points  $\hat{t}_1, \dots, \hat{t}_d$ , a natural estimate of the optimal classification rule is provided by the Fisher rule based on the corresponding projections, that is,  $\hat{g}(x) = \mathbb{I}_{\{\hat{\eta}(x) > 0\}}$ , where

$$\hat{\eta}(x) = \sum_{i=1}^d \hat{\alpha}_i \left( x(\hat{t}_i) - \frac{\hat{m}_0(\hat{t}_i) + \hat{m}_1(\hat{t}_i)}{2} \right) - \log \left( \frac{1-p}{p} \right), \quad (2.11)$$

with  $(\hat{\alpha}_1, \dots, \hat{\alpha}_d)^\top = K_{\hat{t}_1, \dots, \hat{t}_d}^{-1} \hat{m}_{\hat{t}_1, \dots, \hat{t}_d}$ . Note that we assume that the covariance function is known. This amounts to assume that we are dealing with a given model: for example, we assume that we want to discriminate between a standard Brownian motion and a Brownian with drift. Of course, the case in which the covariance structure is estimated can be considered as well (see Subsection 2.4.3). In the following result we establish the consistency (i.e. the asymptotic optimality) of this procedure.

**Theorem 2.6** (Consistency of the RKHS-based classifier). *Let us consider the framework and conditions in Theorem 2.2 and assume further that [SA] holds. Let  $L^* = \mathbb{P}(g^*(X) \neq Y)$  the optimal misclassification probability corresponding to the Bayes rule defined in (2.10). Denote by  $L_n = \mathbb{P}(\hat{g}(X) \neq Y | X_1, \dots, X_n)$  the misclassification probabilities of the rules defined in (2.11). Then,  $L_n \rightarrow L^*$  a.s., as  $n \rightarrow \infty$ .*

### 2.4.3. Practical issues

There are several difficulties concerning the approach introduced in the previous paragraph. First, the number  $d$  of points to be selected is assumed to be known. Second,  $\hat{\psi}(t_1, \dots, t_d)$  is a non-convex function with potentially many local maxima. Third, matrix  $K_{t_1, \dots, t_d}$  and prior probability  $p$  may not be known either. In order to deal with the last difficulty,  $K_{t_1, \dots, t_d}$  and  $p$  must be replaced by suitable consistent estimators  $\hat{K}_{t_1, \dots, t_d}$  and  $\hat{p}$ . The appropriate estimator  $\hat{K}_{t_1, \dots, t_d}$  depends on the assumptions we are willing to make about the processes involved in the classification problem. For instance, if all we want to assume is that they are Gaussian, we could use the pooled sample covariance matrix. However, under a parametric model, only a few parameters should be estimated in order to get  $\hat{K}_{t_1, \dots, t_d}$ .

In practice, we can use the following procedure to deal with the other two difficulties:

1. Initial step: consider a large enough grid of points in  $[0, T]$  and find  $\hat{t}_1$  such that  $\hat{\psi}(\hat{t}_1) \geq \hat{\psi}(t)$  when  $t$  ranges over the grid. Observe that this initial step amounts to find the point maximizing the signal-to-noise ratio since

$$\hat{\psi}(t) = \frac{\hat{m}(t)^2}{\hat{\sigma}_t^2} = \frac{(\bar{X}_1(t) - \bar{X}_0(t))^2}{\hat{\sigma}_t^2},$$

for a suitable estimator  $\hat{\sigma}_t^2$  of the variance at  $t$ .

2. Repeat until convergence: once we have computed  $\hat{t}_1, \dots, \hat{t}_{d-1}$ , find  $\hat{t}_d$  such that  $\hat{\psi}(\hat{t}_1, \dots, \hat{t}_{d-1}, \hat{t}_d) \geq \hat{\psi}(\hat{t}_1, \dots, \hat{t}_{d-1}, t)$  for all  $t$  in rest of the grid.

Whereas we have no guarantee that the greedy algorithm above converges to the global maximum of  $\hat{\psi}(t_1, \dots, t_d)$ , it is computationally affordable and shows good performance in practice. The resulting variable selection method is denoted RK-VS. The result of applying linear Fisher rule to the variables selected by RK-VS yields the classifier denoted RK-C.

*A motivating example. The gains associated with model information*

The new RK methods can incorporate information on the assumed underlying model. For example if (as it often happens in parametric inference) we are willing to assume that the data trajectories come from a Brownian Motion with different (unknown) mean functions, we would like to use this information in our variable selection + classification task. Thus, we will denote by  $\text{RK}_B$  (plus -VS or -C) our

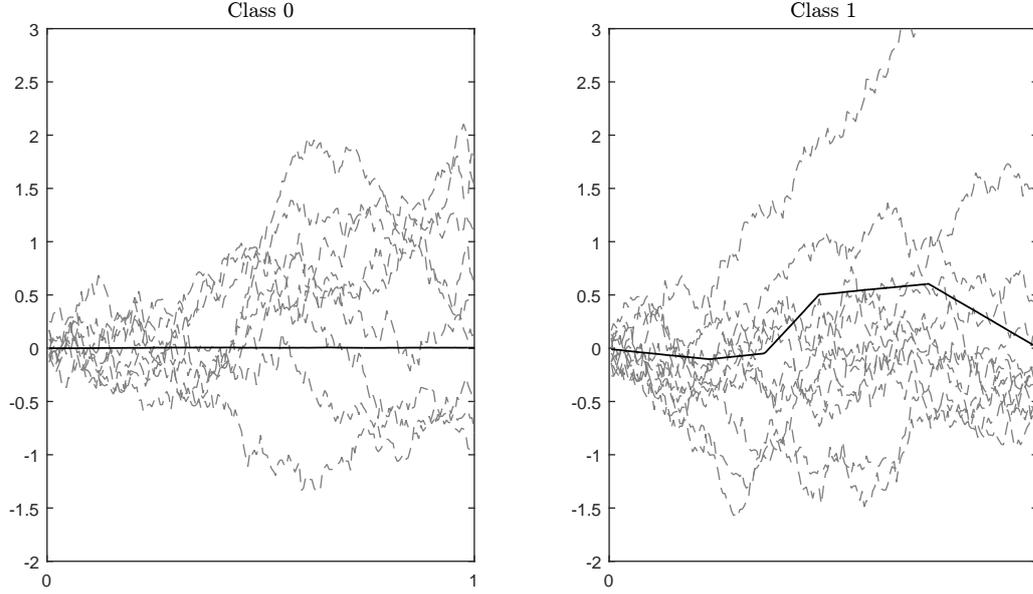


Figure 2.1: Mean functions and some trajectories (dashed lines) of population 0 (left panel) and population 1 (right panel).

RKHS based methods in which we incorporate this information by assuming that the common covariance function of  $P_0$  and  $P_1$  is  $K(s, t) = \min\{s, t\}$ .

To gain some insight on our RK methods it is interesting to compare  $RK_B$  with the standard RK versions in which  $K(s, t)$  is estimated from the sample. To this end, consider a simulated example under the general model (2.3) in which  $P_0$  and  $P_1$  are Brownian motions whose mean functions fulfil  $m(t) = m_1(t) - m_0(t) = \sum_{i=1}^r a_i \Phi_{m,k}(t)$ , where  $t \in [0, 1]$ , the  $a_i$  are constants and the  $\{\Phi_{m,k}\}$  are continuous piecewise linear functions as those considered in Mörters and Peres (2010, p. 28); in fact, it is proved there that the  $\{\Phi_{m,k}\}$  form an orthonormal basis of the Dirichlet space  $\mathcal{D}[0, 1]$  which, as commented above, is the RKHS space corresponding to this model. As a consequence, the equivalence condition in Theorem 2.2 is automatically fulfilled. In addition, given the simple structure of the “peak” functions  $\Phi_{m,k}$ , it is easy to see that the sparsity assumption [SA] also holds in this case. To be more specific, in our simulation experiments we have taken  $m_0(t) = 0$ ,  $m_1(t) = \Phi_{1,1}(t) - \Phi_{2,1}(t) + \Phi_{2,2}(t) - \Phi_{3,2}(t)$ , and  $p = \mathbb{P}(Y = 1) = 1/2$ , so that the Bayes rule given by Theorem 2.2 depends only on the values  $x(t)$  at  $t = 0, 1/4, 3/8, 1/2, 3/4$  and 1 and the Bayes error is 0.1587. Some trajectories (dashed lines) and the population mean functions are displayed in Figure 2.1.

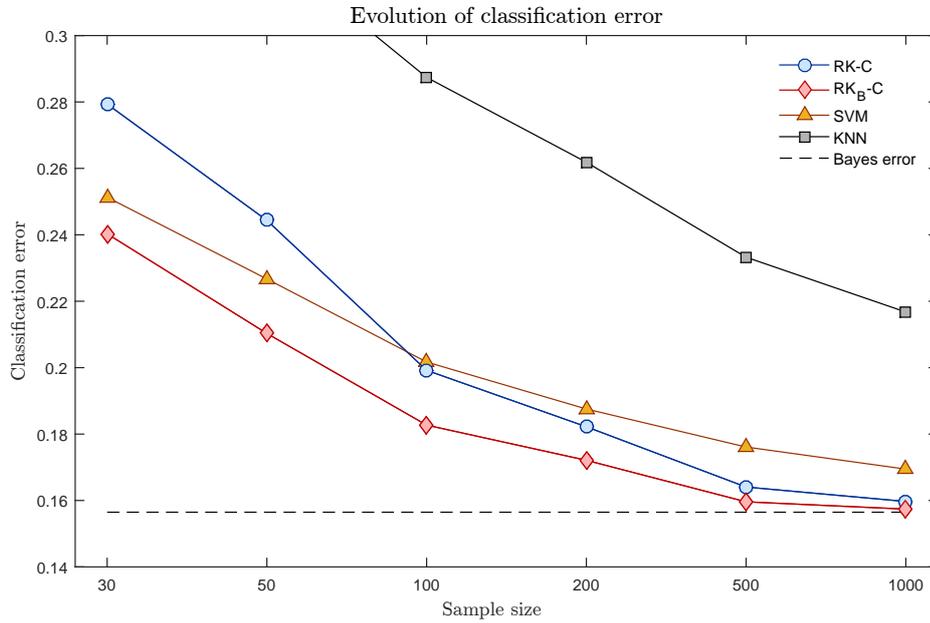


Figure 2.2: Evolution of the classification error of RK-C and RK<sub>B</sub>-C in terms of the sample size.

Now, we analyse the performance of RK and RK<sub>B</sub> in this example. Figure 2.2 shows the evolution of the classification error as the sample size increases for RK-C (blue line with circles), RK<sub>B</sub>-C (red line with diamonds),  $k$ -nearest neighbor rule with the Euclidean distance ( $k$ -NN, gray line with squares) and the support vector machine classifier with a linear kernel (SVM, orange line with triangles). The last two rules are applied to the complete trajectories, without any variable selection. The dashed black line indicates the Bayes error. Each output is obtained by averaging 100 independent runs with test samples of size 200; for each sample size, the number of selected variables (RK-C and RK<sub>B</sub>-C), the number  $k$  of neighbours ( $k$ -NN) and the cost parameter (SVM) are set through a validation sample. Likewise, Figure 2.3 shows the averaged classification error (over 100 runs) in terms of the number of selected variables for RK-C and RK<sub>B</sub>-C for  $n = 100$  (left panel) and  $n = 500$  (right panel). Finally, Figures 2.4 and 2.5 show the frequency of selection of each variable among the first six (by construction, we know there are just six relevant points) corresponding to 100 independent runs of RK-VS and RK<sub>B</sub>-VS, for three different sample sizes. The theoretical relevant points are marked by vertical dashed lines. So, to sum up, whereas Figures 2.2 and 2.3 summarize the results in terms of classification performance, Figures 2.4 and 2.5 are more concerned with the capacity of identifying the true relevant variables.

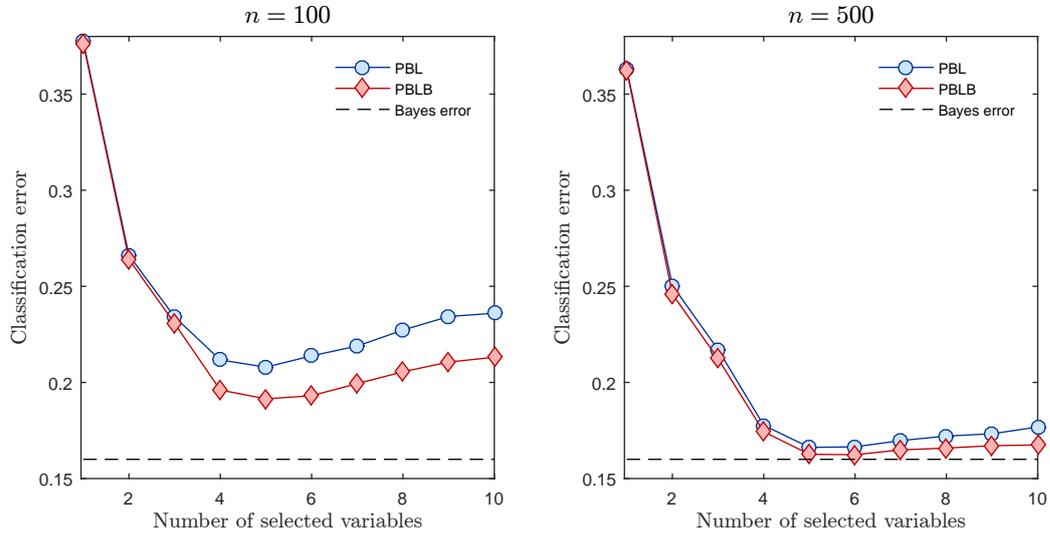


Figure 2.3: Evolution of the classification error of RK-C and  $RK_B$ -C in terms of the number of selected variables for  $n = 100$  (left) and  $n = 500$  (right).

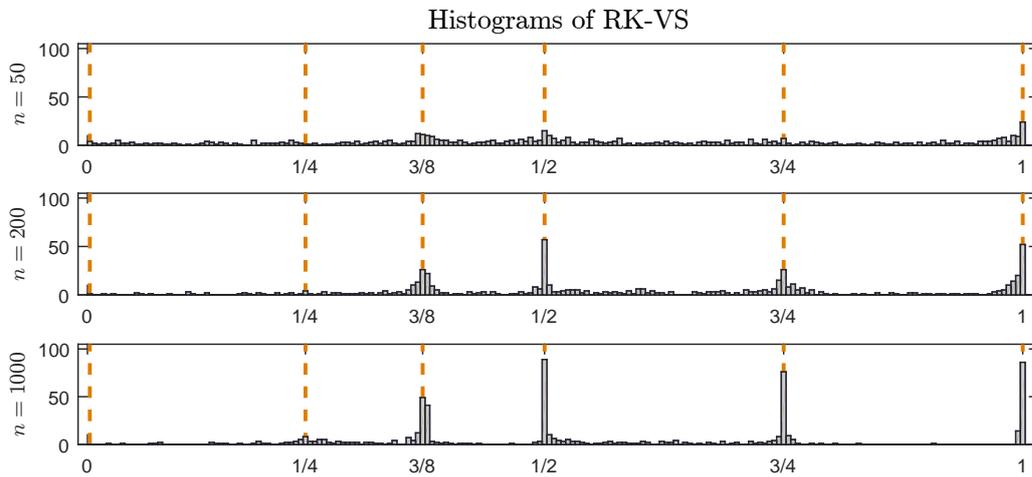


Figure 2.4: Histograms of the six first selected variables by RK-VS over 100 runs for sample sizes 50 (top panel), 200 (middle panel) and 1000 (bottom panel).

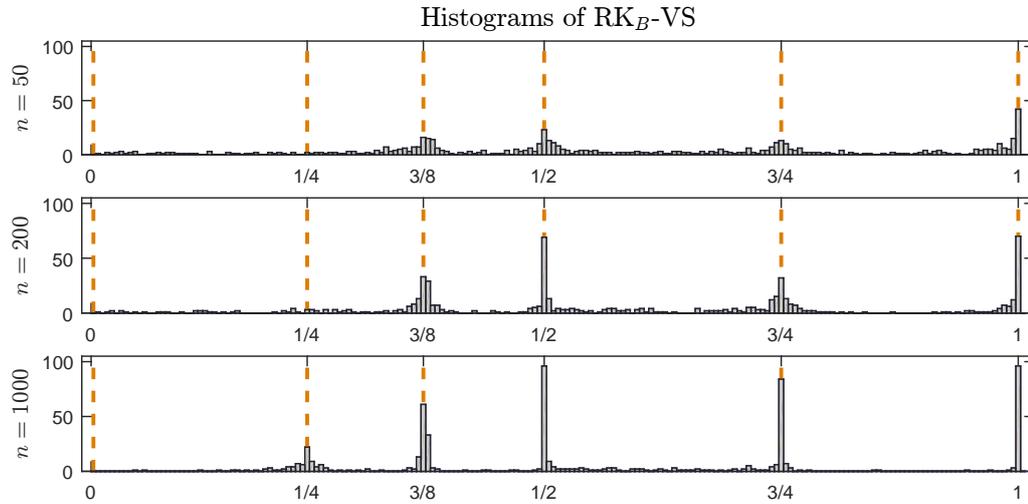


Figure 2.5: Histograms of the six first selected variables by  $RK_B$ -VS over 100 runs for sample sizes 50 (top panel), 200 (middle panel) and 1000 (bottom panel).

These results are quite positive;  $RK$ -C seems to be a good estimator of the optimal classifier as the error rate converges swiftly to the Bayes error even when the number of variables is unknown and fixed by validation. Observe that the convergence seems to be slower for other standard classifiers such as  $k$ -NN and SVM (Figure 2.2). Figure 2.3 shows that for the true number of variables (six) and enough observations, the algorithm achieves the best performance. By contrast, a wrong choice of the number of variables can entail an important increase of the misclassification rate, so this is a sensitive issue. In addition, the selected variables (represented in Figures 2.4 and 2.5) are mostly in coincidence with the theoretical ones. Even for small sample sizes,  $RK_B$ -VS and  $RK$ -VS variables are grouped around the relevant variables. Only the variable  $X(0)$  is omitted since it is in fact nearly irrelevant (see Figure 2.3). This good performance in detecting the important variables is in principle better than one might expect for a greedy algorithm (that, therefore might not provide the true global optimum). Finally, let us note that the inclusion of some additional information seems specially beneficial for smaller sample sizes.

## 2.5. Experiments

Our purpose in Section 2.4 was twofold: we proposed both a variable selection method and an associated classifier. We check here the performance of the proposal from both points of view. Let us recall that common models, methods, data sets and methodological details are fully explained in Chapter 5 and the full

list of simulation models is in Appendix A. However, all of these elements involved in these experiments are briefly described aiming at the completeness and consistency of this chapter.

### 2.5.1. Methodology

We compare RK (and  $RK_B$ ) methods with other variable selection procedures and classification rules by means of a simulation experiment based on the 94 functional models considered in Appendix A for which the mean functions  $m_0$  and  $m_1$  are different (otherwise any linear method is blind to discriminate between  $P_0$  and  $P_1$ ). Just a few of these models satisfy all hypotheses used in previous sections; others differ in several aspects so that we can check the behaviour of our proposal when some departures from the assumptions are present. Training samples of sizes  $n = 30, 50, 100, 200$  are considered for each model. Sample trajectories are discretized in 100 equispaced points in the interval  $[0,1]$ . The criterion of comparison is the classification accuracy for an independent test sample of size 200. The number of selected variables as well as the classification parameters (if needed) are fixed in a validation step, using, for each test sample, another independent validation sample of size 200. The final output is the average classification accuracy over 200 runs of each experiment.

Apart from **RK-VS** and  **$RK_B$ -VS**, the following **variable selection methods** (chosen among the winners in Berrendero et al. (2015b,c)) are considered in the study:

- **mRMR-RD**: this modification of the minimum redundancy maximum relevance algorithm (mRMR) is fully described in Chapter 4. We consider here the version which uses the difference criterion and the distance correlation measure.
- **MHR**: the maxima hunting method for variable selection measure is defined in Chapter 3. We also consider the distance correlation based alternative.
- **PLS**: partial least squares, a well-known dimension reduction technique; see e.g. Delaigle and Hall (2012b) and references therein.

Regarding the **classifiers**, we compare our **RK-C** and  **$RK_B$ -C** methods (Fisher linear rule **LDA** applied to the selected variables) with the standard nearest neighbours rule, denoted  $k$ -**NN** and the support vector machine classifier, denoted **SVM**, based on a linear kernel.

Table 2.1: Percentage of correct classification with the three considered classifiers.

Classifier	$n$	Dimension reduction methods				
		mRMR-RD	PLS	MHR	RK-VS	RK <sub>B</sub> -VS
LDA	30	81.04	82.87	82.44	81.50	80.89
	50	82.37	83.78	83.68	83.44	82.54
	100	83.79	84.70	84.97	85.30	84.46
	200	84.88	85.46	85.90	86.51	85.90
$k$ -NN	30	81.88	82.45	82.46	82.28	81.92
	50	82.95	83.49	83.43	83.75	83.25
	100	84.31	84.77	84.73	85.59	84.95
	200	85.38	85.79	85.91	87.16	86.50
SVM	30	83.22	84.12	84.62	84.28	84.12
	50	84.21	85.04	85.44	85.60	85.20
	100	85.27	86.03	86.29	86.96	86.48
	200	86.10	86.79	86.86	87.90	87.50

## 2.5.2. Simulation outputs

We first focus on the performance of the proposed methods when considered as variable selection methodologies (RK-VS and RK<sub>B</sub>-VS), to be used later in combination with different standard classifiers. All considered dimension reduction methods are data-driven, i.e., independent of the classifier, so we can use the more convenient one to our goals. For illustrative purposes we show the results with LDA,  $k$ -NN and SVM.

Some results are shown in Tables 2.1 and 2.2. Variable selection methods and PLS are in columns and each row corresponds to a sample size and a classifier. Each output Table 2.1 is the average classification accuracy of the 94 models over 200 runs. Table 2.2 contains the corresponding average number of variables (or PLS components) selected by each method and classifier. Boxed outputs denote the best result for each sample size and classifier. The full results of the 1128 experiments (94 models  $\times$  4 samples sizes  $\times$  3 classifiers) are available from the author.

The results are quite similar for all considered classifiers: RK-VS methodology outperforms the other competitors on average with a better performance for bigger sample sizes. Although RK-VS could have more difficulties to estimate the covariance matrix for small sample sizes, it is very close to MHR, which seems to

Table 2.2: Average number of selected variable (or components) with the considered classifiers.

Classifier	$n$	Dimension reduction methods				
		mRMR-RD	PLS	MHR	RK-VS	RK <sub>B</sub> -VS
LDA	30	4.9	2.6	5.4	2.7	3.7
	50	5.9	2.8	6.1	2.8	4.1
	100	7.2	3.3	7.0	3.2	4.8
	200	8.1	4.0	7.5	3.9	5.6
$k$ -NN	30	7.8	4.3	6.2	7.6	8.1
	50	8.0	4.8	6.2	7.3	7.9
	100	8.4	5.5	6.2	6.7	7.6
	200	8.6	6.2	5.9	6.3	7.2
SVM	30	9.3	3.3	8.0	9.3	10.0
	50	9.4	3.8	7.9	8.7	9.6
	100	9.7	4.6	7.9	8.0	9.2
	200	9.8	5.6	7.5	7.6	8.9

be the winner in this setting. Besides, the number of variables selected by RK-VS is comparable to those of mRMR-RD and MHR for both  $k$ -NN and SVM but it is about half of the number selected by mRMR-RD and MHR for LDA (the number of PLS components is often smaller but they lack interpretability). Note that, according with the available experimental evidence (Berrendero et al., 2015b,c), the competing selected methods (mRMR-RD, MHR and PLS) have themselves a good general performance. So, the outputs in Table 2.1 are remarkable and encouraging especially taking into account that only 7 out of 94 models under study fulfil all the regularity conditions required for RK-VS. Note that the ‘‘Brownian assumption’’ implicit in the RK<sub>B</sub>-VS method does not entail a big loss of accuracy with respect to the ‘‘non-parametric’’ RK-VS version.

Finally, it is perhaps worthwhile to assess the performance of RK/RK<sub>B</sub> algorithms when strictly considered as classification rules, rather than as variable selection methodologies.

Table 2.3 provides again average percentages of correct classification over 200 runs of the previously considered 94 functional models. The results are grouped by sample size (in rows). Classification methods are in columns. The full detailed outputs are available from the authors. The difference with Table 2.1 is that, in this case, the classifiers  $k$ -NN and SVM are used with no previous variable selection. So, the original whole functional data are used. This is why we have replaced

Table 2.3: Average classification accuracy (%) over all considered models

$n$	$k$ -NN	SVM	RK-C	RK <sub>B</sub> -C	LDA-Oracle
30	79.61	83.86	81.50	80.89	84.97
50	80.96	85.01	83.44	82.54	86.23
100	82.60	86.20	85.30	84.46	87.18
200	83.99	87.07	86.51	85.90	87.69

Table 2.4: Average accuracy (%) over the models satisfying the assumptions of Thm. 2.6

$n$	$k$ -NN	SVM	RK-C	RK <sub>B</sub> -C	LDA-Oracle
30	83.20	87.29	88.30	89.95	90.91
50	84.90	88.81	89.81	90.69	91.41
100	86.61	89.88	90.81	91.18	91.64
200	87.94	90.48	91.13	91.30	91.71

the standard linear classifier LDA (which cannot be used in high-dimensional or functional settings) with the **LDA-Oracle** method which is just the Fisher linear classifier based on the “true” relevant variables (which are known beforehand since we consider models for which the Bayes rule depends only on a finite set of variables). Of course this classifier is not feasible in practice; it is included here only for comparison purposes.

As before, RK-C results are better for higher sample sizes and the distances between SVM or LDA-Oracle and RK-C are swiftly shortened with  $n$ ; and again, RK<sub>B</sub>-C is less accurate than RK-C but not too much. While the global winner is SVM, the slight loss of accuracy associated with the use of RK-C and RK<sub>B</sub>-C can be seen as a reasonable price for the simplicity and ease of interpretability of these methods. Note also that the associated procedure of variable selection can be seen as a plus of RK-C. In fact, the combination of RK-VS with SVM outperforms SVM based on the whole functional data (see Table 2.1).

Table 2.4 shows average percentages of correct classification over 200 runs of the subset of 7 models that satisfy the assumptions in Theorem 2.6, which establishes the asymptotic optimality of the procedure proposed in Section 2.4. It is not surprising that for these models RK-C and RK<sub>B</sub>-C have a better performance than  $k$ -NN and SVM, even for small sample sizes. In fact the percentages of correct classification are very close to those of LDA-Oracle meaning there is not much room for improvement under these assumptions.

### 2.5.3. Real data

Finally, we study the RK-C performance in two real data examples. We have chosen the “easiest” and the “hardest” data sets (from the point of view of supervised classification) of those considered in [Delaigle and Hall \(2012a\)](#). Given the close connections between our theoretical framework and that of these authors, the use of the same benchmark data sets seems pertinent.

Thus, we follow the same methodology as in the cited paper, that is, we divide the data set randomly in a training sample of size  $n$  ( $n = 30, 50, 100$ ) and a test sample with the remainder observations. Then, the RK-C classifier is constructed from the training set and it is used to classify the test data. The misclassification error rate is estimated through 200 runs of the whole process. The number of variables selected by RK-C is fixed by a standard leave-one-out cross-validation procedure over the training data.

We consider the *Wheat* and the *Phoneme* data sets. *Wheat* data correspond to 100 near infrared spectra of wheat samples measured from 1100nm to 2500nm in 2nm intervals. Following [Delaigle and Hall \(2012a\)](#) we divide the data in two populations according to the protein content (more or less than 15) and use the derivative curves obtained with splines. For this wheat data the near perfect classification is achieved. *Phoneme* is a popular data set in functional data analysis. It consists of log-periodograms obtained from the pronunciation of five different phonemes recorded in 256 equispaced points. We consider the usual binary version of the problem which is not easy to solve. As in the reference paper we make the trajectories continuous with a local linear smoother and remove the noisiest part keeping the first 50 variables. More details and references on this data can be found in Chapter 5.

Table 2.5 shows exactly the same results of Table 2 in [Delaigle and Hall \(2012a\)](#) plus an extra column (in boldface) for our RK-C method. Since we have followed the same methodology, the results are completely comparable despite the minimum differences due to the randomness.  $\text{CENT}_{PC1}$  and  $\text{CENT}_{PLS}$  stand for the centroid classifier (2.7), where the function  $\psi$  is estimated via principal components or PLS components, respectively. NP refers to the classifier based in the non-parametric functional regression method proposed by [Ferraty and Vieu \(2006\)](#) and  $\text{CENT}_{PCp}$  denotes the usual centroid classifier applied to the multivariate principal component projections. The outputs correspond to the average (over 200 runs) percentages of misclassification obtained for each method, sample size and data set. The values in parentheses correspond to the standard deviation of these errors.

Table 2.5: 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 <sub>PC1</sub>	CENT <sub>PLS</sub>	NP	CENT <sub>PCp</sub>	<b>RK-C</b>
Wheat	30	0.89 (2.49)	0.46 (1.24)	0.49 (1.29)	15.0 (1.25)	<b>0.25 (1.58)</b>
	50	0.22 (1.09)	0.06 (0.63)	0.01 (0.14)	14.4 (5.52)	<b>0.02 (0.28)</b>
Phoneme	30	22.5 (3.59)	24.2 (5.37)	24.4 (5.31)	23.7 (2.37)	<b>22.5 (3.70)</b>
	50	20.8 (2.08)	21.5 (3.02)	21.9 (2.91)	23.4 (1.80)	<b>21.5 (2.36)</b>
	100	20.0 (1.09)	20.1 (1.12)	20.1 (1.37)	23.4 (1.36)	<b>20.1 (1.25)</b>

The results show that the RK-C classifier is clearly competitive against the remaining methods. In addition, there is perhaps some interpretability advantage in the use of RK-C, as this method is based in dimension reduction via variable selection so that the "reduced data" are directly interpretable in terms of the original variables. Let us finally point out that the variable selection process is quite efficient: in the wheat example, near perfect classification is achieved using just one variable; in the much harder phoneme example, the average number of selected variables is three.

## 2.6. Conclusions

We have proposed an RKHS-based method for both variable selection and binary classification. It is fully theoretically motivated in terms of the RKHS space associated with the underlying model. The method can be adapted, in a very natural way, to incorporate information on the covariance structure of the model. In our empirical study we have explored the Brownian case via  $RK_B$ : the method defined in Subsection 2.4.2 when we assume that  $K(s, t) = \min(s, t)$ .

We next summarize our study of the RK methods in the following conclusions.

- a) The identification of the RKHS associated with a supervised classification problem represents several important theoretical and practical advantages. Apart from providing explicit expressions of the optimal Bayes rule (via the corresponding Radon-Nikodym derivatives), the RKHS approach provides a theoretical explanation for the near perfect classification phenomenon in terms of the mutual singularity of the involved measures.
- b) Perhaps more importantly, the RKHS approach provides a theoretical scenario to motivate the use of variable selection. The point is that, under the

RKHS framework, the family of models fulfilling the sparsity assumption [SA] is dense in the whole class of considered models.

- c) The RKHS-based variable selection and classification procedures are quite accurate and computationally inexpensive with important advantages in terms of simplicity and interpretability. The simulation outputs show that RK-VS procedure is especially successful as a variable selection method. As a classifier RK-C is still competitive and especially good when the underlying assumptions are fulfilled.
- d) The empirical results show also a remarkable robustness of the RK methodology against departures from the assumptions on which it is based.

## 2.7. Proofs

*Proof of Theorem 2.2.* Equation (2.4) follows straightforwardly from the combination of (1.2) and (2.1). To prove the expression for the Bayes error notice that  $\langle X - m_0, m \rangle_K$  lies in  $\bar{\mathcal{L}}(X - m_0)$  and therefore the random variable  $\eta^*(X)$  is Gaussian both under  $Y = 1$  and  $Y = 0$ . Furthermore, Equations (6.19) and (6.20) in Parzen (1961) yield

$$\begin{aligned}\mathbb{E}(\eta^*(X)|Y = 0) &= -\|m\|_K^2/2 - \log\left(\frac{1-p}{p}\right), \\ \mathbb{E}(\eta^*(X)|Y = 1) &= \|m\|_K^2/2 - \log\left(\frac{1-p}{p}\right), \\ \text{Var}(\eta^*(X)|Y = 0) &= \text{Var}(\eta^*(X)|Y = 1) = \|m\|_K^2.\end{aligned}$$

The result follows using these values to standardize the variable  $\eta^*(X)$  in  $L^* = (1-p)\mathbb{P}(\eta^*(X) > 0|Y = 0) + p\mathbb{P}(\eta^*(X) < 0|Y = 1)$ .  $\square$

*Proof of Theorem 2.4.* Observe that, if  $\theta_j > 0$  for all  $j \geq 1$ ,

$$m_1 = \sum_{j=1}^{\infty} \mu_j \phi_j = \sum_{j=1}^{\infty} \frac{\mu_j}{\sqrt{\theta_j}} \sqrt{\theta_j} \phi_j,$$

where  $\{\sqrt{\theta_j} \phi_j : \theta_j > 0\}$  is an orthonormal basis of  $\mathcal{H}(K)$  [see, e.g., Theorem 4.12, p. 61 in Cucker and Zhou (2007)]. Then, by Parseval's formula,  $m_1 \in \mathcal{H}(K)$

if and only if  $\|m_1\|_K^2 = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j^2 < \infty$ . As a consequence, we have the desired equivalence:

$$P_1 \sim P_0 \Leftrightarrow m_1 \in \mathcal{H}(K) \Leftrightarrow \|m_1\|_K < \infty \Leftrightarrow \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j^2 < \infty.$$

Moreover,

$$\text{err}_0 = 1 - \Phi \left( \frac{1}{2} \left( \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j^2 \right)^{1/2} \right) = 1 - \Phi \left( \frac{1}{2} \|m_1\|_K \right),$$

what gives the coordinate-free expression of the Bayes error.

In order to obtain a coordinate-free expression of the Bayes rule, notice that (2.7) holds if and only if

$$\langle m_1, \psi \rangle_{L^2}^2 - 2 \langle m_1, \psi \rangle_{L^2} \langle X, \psi \rangle_{L^2} < 0. \quad (2.12)$$

Since  $m_1 = \sum_{j=1}^{\infty} \mu_j \phi_j$ , with  $m_1 \neq 0$ , and  $\psi = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \phi_j$ , it holds  $\langle m_1, \psi \rangle_{L^2} = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j^2 = \|m_1\|_K^2 \neq 0$ . Therefore, (2.12) holds if and only if

$$\langle X, \psi \rangle_{L^2} - \frac{\|m_1\|_K^2}{2} > 0.$$

To end the proof it is enough to show  $\langle X, m_1 \rangle_K = \langle X, \psi \rangle_{L^2}$ . The linearity of  $\langle X, \cdot \rangle_K$  and the fact that  $\theta_j$  and  $\phi_j$  are respectively eigenvalues and eigenfunctions of the integral operator with kernel  $K$  imply

$$\langle X, m_1 \rangle_K = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \langle X, \theta_j \phi_j \rangle_K = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \int_0^T \langle X, K(\cdot, u) \rangle_K \phi_j(u) du.$$

Now, from Equation (6.18) in Parzen (1961),

$$\int_0^T \langle X, K(\cdot, u) \rangle_K \phi_j(u) du = \int_0^T X(u) \phi_j(u) du = \langle X, \phi_j \rangle_{L^2}.$$

Finally, combining the two last displayed equations,

$$\langle X, m_1 \rangle_K = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \langle X, \phi_j \rangle_{L^2} = \langle X, \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \phi_j \rangle_{L^2} = \langle X, \psi \rangle_{L^2}.$$

□

*Proof of Theorem 2.5.* Let  $X = \sum_j Z_j \phi_j$ , the Karhunen-Loève expansion of  $X$ , with the  $Z_j$  uncorrelated. For a given trajectory  $x = \sum_j z_j \phi_j$ . Define  $x^n = \sum_j^n z_j \phi_j$ , This is a trajectory drawn from the process  $X^n = \sum_j^n Z_j \phi_j$ , whose distribution under  $P_i$  is denoted by  $P_{in}$  (for  $i = 0, 1$ , the covariance function is  $K_n(s, t) = \sum_{i=1}^n \mathbb{E}(Z_i^2) \phi_i(s) \phi_i(t)$  and the mean function is

$$m_n(t) = \sum_j^n \mathbb{E}(Z_j) \phi_j(t),$$

Note that, under  $P_0$ ,  $\mathbb{E}(Z_j) = 0$ , so that the mean function is 0. From Karhunen-Loève Theorem (see [Ash and Gardner \(2014\)](#), p. 38)  $m_n(t) \rightarrow m(t)$  for all  $t$  (in fact this results holds uniformly in  $t$ ).

Note also that  $m_n \in \mathcal{H}(K)$ . Again this follows from the fact that  $\{\sqrt{\theta_j} \phi_j : \theta_j > 0\}$  is an orthonormal basis of  $\mathcal{H}(K)$  [see, e.g., [Theorem 4.12](#), p. 61 in [Cucker and Zhou \(2007\)](#)].

We now prove that we must necessarily have  $\lim_n \|m_n\|_K = \infty$ . Indeed, if we had  $\lim_n \|m_n\|_K < \infty$  for some subsequence of  $\{m_n\}$  (denoted again  $\{m_n\}$ ) we would have that such  $\{m_n\}$  would be a Cauchy sequence in  $\mathcal{H}(K)$ , since for  $q > p$ ,  $\|m_p - m_q\|_K \leq \left| \|m_q\|_K - \|m_p\|_K \right|$ . This, together with the pointwise convergence  $m_n(t) \rightarrow m(t)$  leads, from Moore-Aronszajn Theorem (see [Berlinet and Thomas-Agnan \(2004\)](#), p. 19) to  $m \in \mathcal{H}(K)$ . But, from [Parzen's Theorem 2.1](#), this would entail  $P_1 \ll P_0$ , in contradiction with  $P_1 \perp P_0$ . We thus conclude  $\|m_n\|_K \rightarrow \infty$ .

Then, given  $\epsilon > 0$ , choose  $n$  such that

$$\begin{aligned} & (1-p)\Phi\left(-\frac{\|m_n\|_K}{2} - \frac{1}{\|m_n\|_K} \log\left(\frac{1-p}{p}\right)\right) \\ & + p\Phi\left(-\frac{\|m_n\|_K}{2} + \frac{1}{\|m_n\|_K} \log\left(\frac{1-p}{p}\right)\right) < \epsilon, \end{aligned} \quad (2.13)$$

Now, consider the problem  $X^n \sim P_{1n}$  vs  $X^n \sim P_{0n}$ . Note that  $X^n \sim P_{in}$  if and only if  $X \sim P_i$ , for  $i = 0, 1$ . Since  $m_n \in \mathcal{H}(K_n)$ , we have  $P_{0n} \sim P_{1n}$  (using again [Parzen's Theorem 2.1](#)).

Hence, according to the theorem we have proved (on the expression of the optimal rules in the absolutely continuous case under homoscedasticity), the optimal rule is  $g_n(X) = \mathbb{I}_{\{\eta_n(X) > 0\}}$ , where

$$\eta_n(x) = \langle x, m_n \rangle_K - \frac{1}{2} \|m_n\|_K^2 - \log\left(\frac{1-p}{p}\right), \quad (2.14)$$

whose probability of error, is exactly the expression on the left-hand side of (2.13). So this probability is  $\leq \epsilon$ .  $\square$

*Proof of Theorem 2.6.* Let us consider, without loss of generality, that  $p = \mathbb{P}(Y = 1) = 1/2$ . We have shown in Subsection 2.4.1 that the Bayes rule for our problem is the linear discriminant rule for the selected variables  $(X(t_1^*), \dots, X(t_n^*))$ . So the corresponding Bayer error is  $L^* = 1 - \Phi(\psi(t_1^*, \dots, t_d^*)^{1/2}/2)$ , where  $\psi(t_1, \dots, t_d) := m_{t_1, \dots, t_d}^\top K_{t_1, \dots, t_d}^{-1} m_{t_1, \dots, t_d}$  and  $\Phi$  is the cumulative distribution function of the standard Gaussian distribution. Recall that  $\psi(t_1, \dots, t_d)$  is the (square) Mahalanobis distance between the vectors of mean functions  $(m_j(t_1^*), \dots, m_j(t_d^*))$  for  $j = 1, 2$ .

However, as pointed out in Subsection 2.4.2 our classification rule is an empirical approximation of this optimal classifier which is defined by replacing  $\psi(t_1^*, \dots, t_d^*)$  by the natural estimator  $\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) = \hat{m}_{\hat{t}_1, \dots, \hat{t}_d}^\top K_{\hat{t}_1, \dots, \hat{t}_d}^{-1} \hat{m}_{\hat{t}_1, \dots, \hat{t}_d}$ . A direct calculation shows that the conditional error  $L_n$  of this rule is then

$$L_n = 1 - \Phi\left(\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d)^{1/2}/2\right).$$

As  $\Phi$  is continuous, the desired conclusion  $L_n \rightarrow L^*$ , a.s. will readily follow if we prove  $\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) \rightarrow \psi(t_1^*, \dots, t_d^*)$  a.s., as  $n \rightarrow \infty$ .

Observe that  $\psi(t_1, \dots, t_d)$  is continuous and, therefore, uniformly continuous on the compact set  $[0, T]^d$ . Notice also that  $\hat{m} \rightarrow m$  uniformly a.s., as  $n \rightarrow \infty$ . This follows as a direct consequence of Mourier's Strong Law of Large Numbers for random elements taking values in Banach spaces; see, e.g., [Laha and Rohatgi \(1979, p. 452\)](#). Then, with probability 1, given  $\epsilon > 0$  there exists  $N$  such that for  $n \geq N$  and  $(t_1, \dots, t_d)^\top \in [0, T]^d$ ,

$$\hat{m}_{t_1, \dots, t_d}^\top K_{t_1, \dots, t_d}^{-1} \hat{m}_{t_1, \dots, t_d} - \epsilon \leq m_{t_1, \dots, t_d}^\top K_{t_1, \dots, t_d}^{-1} m_{t_1, \dots, t_d} \leq \hat{m}_{t_1, \dots, t_d}^\top K_{t_1, \dots, t_d}^{-1} \hat{m}_{t_1, \dots, t_d} + \epsilon.$$

Taking the maximum of the terms in these inequalities we get

$$\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) - \epsilon \leq \psi(t_1^*, \dots, t_d^*) \leq \hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) + \epsilon, \quad \text{a.s.}$$

That is, we have

$$\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) \rightarrow \psi(t_1^*, \dots, t_d^*), \quad \text{a.s., as } n \rightarrow \infty. \quad (2.15)$$

However, what we need is  $\hat{\psi}(\hat{t}_1, \dots, \hat{t}_d) \rightarrow \psi(t_1^*, \dots, t_d^*)$ . This would readily follow from (2.15) if we had  $\hat{\psi} \rightarrow \psi$ , uniformly on  $[0, T]^d$ , a.s. Denote, by simplicity,  $t = (t_1, \dots, t_d)$  and, given  $\epsilon_0 > 0$ ,

$$E(\epsilon_0) = \{v \in \mathbb{R}^d : \min\{\|v - x\|, x \in G_m\} \leq \epsilon_0\},$$

where  $G_m = \{m(t) : t \in [0, T]^d\}$ . Define also  $Q(t, v) = v^\top K_t^{-1} v$  for  $t \in [0, T]^d$  and  $v \in G_f$ . The function  $Q$  is continuous on the compact set  $[0, T]^d \times E(\epsilon_0)$ , therefore it is uniformly continuous. Hence, in particular, given  $\epsilon > 0$  there exists  $\delta > 0$ ,  $\delta < \epsilon_0$ , such that

$$\|v_1 - v_2\| < \delta \text{ implies } |Q(t, v_1) - Q(t, v_2)| < \epsilon, \quad \forall t, \quad (2.16)$$

Now observe that  $Q(t, m(t)) = \psi(t)$  and  $Q(t, \hat{m}(t)) = \hat{\psi}(t)$ . Using again Mourier's Strong Law of Large Numbers, we have that, for all  $t$ ,  $\|\hat{m}(t) - m(t)\| < \delta$ , a.s., for  $n$  large enough. Thus, from (2.16), we finally get for all  $t$ ,  $|\hat{\psi}(t) - \psi(t)| < \epsilon$ , a.s., for large enough  $n$ , that is, we have the uniform convergence of  $\hat{\psi}$  to  $\psi$ .  $\square$



# Chapter 3

## Maxima-Hunting

This chapter is devoted to a new intrinsic variable selection technique in the functional discrimination setting, the *maxima hunting* (MH) method. It is based on a direct use of the distance covariance and distance correlation measures proposed by Székely et al. (2007). These are flexible association measures with a handful of good properties. A brief review of these statistical tools is given in Section 3.1, and some useful alternative expressions for them in the case of binary classification are derived in Theorem 3.1.

The idea behind MH is as simple as selecting those points  $t$  in the functional variable  $X(t)$  that locally maximize the dependence with the response variable (measured in terms of the distance covariance/correlation). This methodology is easy to interpret, and has a sound functional motivation. Moreover, despite its simplicity, MH deals in a natural way with the redundancy problem removing automatically redundant variables around the maxima. The maxima hunting method is described in Section 3.2. Section 3.3 provides some theoretical support for this methodology. In particular, we present a few explicit models in which the procedure works, in the sense that the variables to be selected for an optimal classification are in fact maxima of the distance correlation function. The optimal rules are derived, for both homoscedastic (Prop. 3.1-3.3) and heteroscedastic (Thm. 3.3) cases, using techniques different from those in Chapter 2. Many other models of this sort can be constructed by a simple, easy-to-interpret, mixture mechanism.

The maxima hunting method is in fact defined in population terms, from the distance correlation function. Hence, the practical implementation of the method, for a given data set, arises as a result of the estimation of such function. This is backed by a consistency result (Thm. 3.2). MH performance is empirically assessed by means of extensive experiments (with both simulations and real data

sets) in Section 3.4. Section 3.5 presents some final conclusions as well as a ranking of all considered methods. Finally, all the proofs together with some additional results are included in the last section.

### 3.1. An auxiliary tool: the distance covariance

The problem of finding appropriate association measures between random variables (beyond the standard linear correlation coefficient) has received increasing attention in recent years. For example, the journal *Science* has published recently a new association measure (illustrated with examples in gene expression and microbiology, among other fields) by Reshef et al. (2011). In the accompanying perspective paper, Speed (2011) describes this proposal as “a correlation measure for the 21st century”. Another “generalized correlation association measure”, illustrated also with genetic microarray examples, has been proposed by Hall and Miller (2011).

Nevertheless, we will use here a third association measure proposed by Székely et al. (2007), see also Székely and Rizzo (2009, 2012, 2013). It is called *distance covariance* (*dcov*) or *distance correlation* (*dcor*) in the standardized version. It has a number of valuable properties: first, it can be used to define the association between two random variables  $X$  and  $Y$  of arbitrary (possibly different) dimensions; second, it characterizes independence in the sense that the distance covariance between  $X$  and  $Y$  is zero if and only if  $X$  and  $Y$  are independent; third, the distance correlation can be easily estimated in a natural plug-in way, with no need of smoothing or discretization.

**Definition 3.1** (Distance covariance). Given two random variables  $X$  and  $Y$  taking values in  $\mathbb{R}^p$  and  $\mathbb{R}^q$ , respectively, let  $\varphi_{X,Y}$ ,  $\varphi_X$ ,  $\varphi_Y$  be the characteristic functions of  $(X, Y)$ ,  $X$  and  $Y$ , respectively. Assume that the components of  $X$  and  $Y$  have finite first-order moments. The distance covariance between  $X$  and  $Y$ , is the non-negative number  $\mathcal{V}(X, Y)$  defined by

$$\mathcal{V}^2(X, Y) = \int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(u, v) - \varphi_X(u)\varphi_Y(v)|^2 w(u, v) du dv, \quad (3.1)$$

with  $w(u, v) = (c_p c_q |u|_p^{1+p} |v|_q^{1+q})^{-1}$ , where  $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$  is half the surface area of the unit sphere in  $\mathbb{R}^{d+1}$  and  $|\cdot|_d$  stands for the Euclidean norm in  $\mathbb{R}^d$ . Finally, denoting  $\mathcal{V}^2(X) = \mathcal{V}^2(X, X)$ , the (square) distance correlation is defined by

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

Of course the main idea is to define the distance between  $X$  and  $Y$  in terms of the weighted  $L^2$  distance between the characteristic function (cf)  $\varphi_{X,Y}$  of the joint distribution  $(X, Y)$  and the corresponding cf for the case of independence (i.e.,  $\varphi_X\varphi_Y$ ). Note that these definitions make sense even if  $X$  and  $Y$  have different dimensions (i.e.,  $p \neq q$ ).

The motivation of the chosen weight function  $w(u, v)$  is not that obvious. However, as proved in Székely and Rizzo (2012, Thm. 1), this is the most suitable choice for  $w$  in order to get equivariance properties for  $\mathcal{V}^2$ . In addition, the association measure  $\mathcal{V}^2(X, Y)$  can be consistently estimated through a relatively simple average of products calculated in terms of the mutual pairwise distances  $|X_i - X_j|_p$  and  $|Y_i - Y_j|_q$  between the sample values  $X_i$  and the  $Y_j$ .

**Definition 3.2** (Estimator of  $\mathcal{V}^2$ ). Let  $X, Y$  be the random vectors defined above and  $\{(X_i, Y_i)\}_{i=1}^n$  an observed random sample from their joint distribution. The empirical distance covariance is defined by

$$\mathcal{V}_n^2 = \frac{1}{n^2} \sum_{i,j=1}^n A_{ij}B_{ij}, \quad (3.2)$$

where  $A_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}$  with  $a_{ij} = |X_i - X_j|_p$ . The other elements stand for the empirical averages of rows ( $\bar{a}_{i.}$ ), columns ( $\bar{a}_{.j}$ ) and the global average ( $\bar{a}$ ) of the matrix  $(a_{ij})$ .  $B$  is the analogous matrix of distances for  $Y$  ( $b_{kl} = |Y_k - Y_l|_q$ ). Again, denoting  $\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X)$ , the empirical distance correlation is defined by

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

The almost surely convergence of  $\mathcal{V}_n^2$  is proved in (Székely et al., 2007, Thm. 2) and implementation in the language R of  $\mathcal{V}_n^2$  and  $\mathcal{R}_n^2$  can be found in the R-package *energy* by Székely and Rizzo.

Finally, let us recall that the powerful idea enclosed in these measures has motivated an increasing number of papers which explore extensions and propose new applications of *dcov* and *dcor*. In this vein, Székely and Rizzo (2013) extend the distance correlation to the problem of testing the independence of high-dimensional random vectors. The same authors define a *partial distance correlation* in Székely and Rizzo (2014). Dueck et al. (2014) propose an affinely invariant version of the *dcor* and Wang et al. (2015) adapt the measure to capture conditional dependencies. On the other hand, Lyons (2013) extends this association measures, *dcov* and *dcor*, from Euclidean to general metric spaces.

## 3.2. Variable selection based on maxima hunting

The interesting properties of  $dcov$  (or  $dcor$ ) have neither gone unnoticed in the variable selection setting. For example, an intrinsic variable selection method is given in Li et al. (2012), based on the idea of “sure independence screening” introduced by Fan and Lv (2008). This proposal (DC-SIS) is developed in the multivariate regression framework and the authors use  $dcor$  for ranking the individual variables. Note that our approach here is quite different since we are not primarily concerned with sure screening (capturing all variables related with the class) but with the idea of selecting sets (as small as possible) of non redundant variables that can achieve good predictive performances. Indeed, that procedure is not primarily designed to deal with functional data, as the correlations among the explanatory variables are not taken into account. The paper by Kong et al. (2015) proposes a modification of DC-SIS including a elimination step in terms of the distance covariance of the selected subset and the response variable. Another version of DC-SIS is given in Zhong and Zhu (2015) where an iterative procedure is used to detect important variables with a low rank score. On the other hand, Yenigün and Rizzo (2015) provide two novel variable selection methods for linear and nonlinear regression models, one of them based in the use of  $dcor$ . Nevertheless, in spite of some features in common, our approximation here is quite different: first, note that all of these works are focused in the regression framework and above all, there is no other reference (as far as we know) tackling the use of  $dcor$  as a variable selection tool in a functional context.

Our proposal is as follows: if we are assuming a sort of functional structure in the data, a high correlation between close variables is to be expected. This must be considered in the variable selection methodology in order to avoid redundancy. Our proposal is based on a direct use of the distance covariance association measure in a “functional” way. We just suggest to select the values of  $t$  corresponding to local maxima of the distance-covariance function  $\mathcal{V}_X^2 = \mathcal{V}^2(X_t, Y)$  or, alternatively, of the distance correlation function  $\mathcal{R}_X^2 = \mathcal{R}^2(X_t, Y)$ . This method has a sound intuitive basis as it provides a simple natural way to deal with the relevance vs. redundancy trade-off: the selected values must carry a large amount of information on  $Y$ , which takes into account the *relevance* of the selected variables. In addition, the fact of considering local maxima automatically takes care of the *redundancy* problem, since the highly relevant points close to the local maxima are automatically excluded from consideration. The MH procedure is also able to detect and incorporate to the model representative variables with small marginal scores. Low relevant areas are usually forgotten but they often provide complementary information (see, e.g. Zhong and Zhu (2015)), here we capture this supplementary information via the local maxima of these areas. These intuitions are

empirically confirmed by the results of Section 3.4, where the practical performance of the maxima-hunting method is quite satisfactory. Figure 3.1 shows how the function  $\mathcal{V}_X^2$  looks like in two different examples.

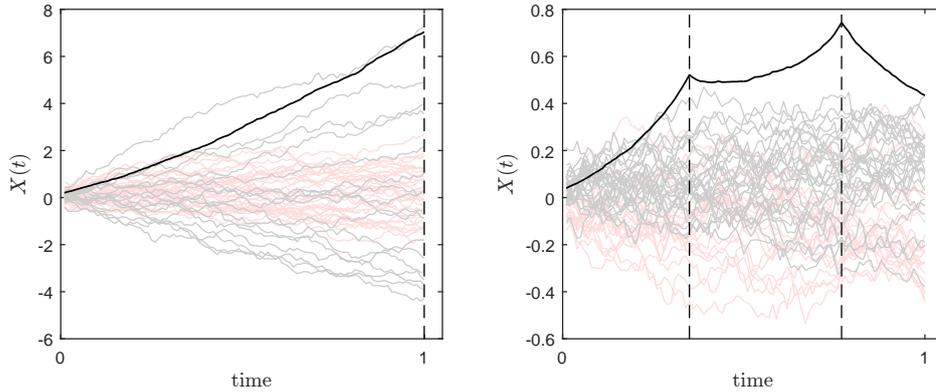


Figure 3.1: Left: 50 trajectories of model in Proposition 3.1. Right: Logistic model L11 (explained in Chapter 5) with 50 Ornstein-Uhlenbeck trajectories.  $\mathcal{V}^2(X_t, Y)$  (scaled) is in black and the relevant variables are marked by vertical dashed lines .

Let us also recall that the maxima hunting methodology provides a natural answer for the unsolved question of the stopping criterion (see Subsection 1.4.2). One could simply select all the local maxima. Although this is a promising starting point, further research is required since, unfortunately, some problems present redundant maxima belonging to different subintervals. Moreover, criteria to define what is a maximum are not always easy to establish in practice when working with discretized functions.

Otherwise, the extreme flexibility of these association measures allow us to consider the case of a multivariate response  $Y$ . So there is no conceptual restriction to apply the same ideas for multiple classification or even to a regression problem. However, we will limit ourselves here to the important problem of binary classification. In this case we can derive simplified expressions for  $\mathcal{V}_X^2$  which are particularly convenient in order to get empirical approximations. This is next shown.

For the sake of generality, throughout this subsection,  $d$  will denote a natural number and  $t$  will stand for a vector  $t = (t_1, \dots, t_d) \in [0, 1]^d$ . Also, for a given process  $X$ , we abbreviate  $X(t) = (X(t_1), \dots, X(t_d))$  by  $X_t$  and  $Z'$  will denote an independent copy of a random variable  $Z$ . We write  $u^\top$  and  $|u|_d$  to denote

the transposed and the Euclidean norm of a vector  $u \in \mathbb{R}^d$ . Let  $\eta(x) = \mathbb{P}(Y = 1|X = x)$  so that  $Y|X \sim \text{Binomial}(1, \eta(X))$  where the symbol  $\sim$  stands for “is distributed as”. Observe that  $p = \mathbb{P}(Y = 1) = \mathbb{E}(\mathbb{P}(Y = 1|X)) = \mathbb{E}(\eta(X))$ .

Our variable selection methodology will heavily depend on  $\mathcal{V}_X^2$ , the function giving the distance covariance dependence measure between the marginal vector  $X(t) = X_t$ , for  $t \in [0, 1]^d$  and  $d \in \mathbb{N}$ , and the class variable  $Y$ . The following theorem gives three alternative expressions for this function. The third one will be particularly useful in what follows.

**Theorem 3.1** (Expressions for  $\mathcal{V}^2$ ). *In the setting of the functional classification problem above stated, the function  $\mathcal{V}^2(X_t, Y)$  defined in (3.1) can be alternatively calculated with the following expressions,*

$$(a) \quad \mathcal{V}^2(X_t, Y) = \frac{2}{c_d} \int_{\mathbb{R}^d} \frac{|\zeta(u, t)|^2}{|u|_d^{d+1}} du, \quad (3.3)$$

where  $\zeta(u, t) = \mathbb{E} \left[ (\eta(X) - p) e^{iu^\top X_t} \right]$  and  $c_d$  is given in Definition 3.1.

$$(b) \quad \begin{aligned} \mathcal{V}^2(X_t, Y) &= -2\mathbb{E} [(\eta(X) - p)(\eta(X') - p)|X_t - X'_t|_d] \\ &= -2\mathbb{E} [(Y - p)(Y' - p)|X_t - X'_t|_d], \end{aligned} \quad (3.4)$$

where  $(X', Y')$  denotes an independent copy of  $(X, Y)$ , respectively.

$$(c) \quad \mathcal{V}^2(X_t, Y) = 4p^2(1 - p)^2 \left[ I_{01}(t) - \frac{I_{00}(t) + I_{11}(t)}{2} \right], \quad (3.5)$$

where  $I_{ij}(t) = \mathbb{E} (|X_t - X'_t|_d | Y = i, Y' = j)$ .

In a training sample  $\{(X_i, Y_i), i = 1, \dots, n\}$  denote by  $X_1^{(0)}, \dots, X_{n_0}^{(0)}$  and  $X_1^{(1)}, \dots, X_{n_1}^{(1)}$  the  $X$ -observations corresponding to values  $Y_i = 0$  and  $Y_i = 1$ , respectively. In this section, we use these data to obtain an estimator of  $\mathcal{V}_X^2$ , which is uniformly consistent in  $t$ . As a consequence, we can estimate the local maxima of  $\mathcal{V}_X^2$ : using part (c) of Theorem 3.1, a natural estimator for  $\mathcal{V}^2(X_t, Y)$  is

$$\tilde{\mathcal{V}}_n^2(X_t, Y) = 4\hat{p}^2(1 - \hat{p})^2 \left[ \hat{I}_{01}(t) - \frac{\hat{I}_{00}(t) + \hat{I}_{11}(t)}{2} \right], \quad (3.6)$$

where  $\hat{p} = n_1/(n_0 + n_1)$ ,  $\hat{I}_{rr}(t) = \frac{2}{n_r(n_r-1)} \sum_{i < j} |X_i^{(r)}(t) - X_j^{(r)}(t)|_d$ , for  $r = 0, 1$ , and  $\hat{I}_{01}(t) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} |X_i^{(0)}(t) - X_j^{(1)}(t)|_d$ . The uniform strong consistency of  $\tilde{\mathcal{V}}_n^2(X_t, Y)$  is established in Theorem 3.2 below.

**Theorem 3.2** (Uniform convergence of  $\tilde{\mathcal{V}}_n^2$ ). *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,  $\tilde{\mathcal{V}}_n^2(X_t, Y)$  is continuous in  $t$  and*

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

*Hence, if we assume that  $\mathcal{V}^2(X_t, Y)$  has exactly  $m$  local maxima at  $t_1, \dots, t_m$ , then  $\tilde{\mathcal{V}}_n^2(X_t, Y)$  has also eventually at least  $m$  maxima at  $t_{1n}, \dots, t_{mn}$  with  $t_{jn} \rightarrow t_j$ , as  $n \rightarrow \infty$ , a.s., for  $j = 1, \dots, m$ .*

In our numerical experiments we use the estimator of  $\mathcal{V}^2(X_t, Y)$  proposed in Székely et al. (2007) instead of the estimator (3.6) we use in Theorem 3.2. In the following lemma we show that both estimators are in fact equivalent.

**Lemma 3.1** (Asymptotic equivalence of estimators). *Let  $\{(X_i, Y_i), i = 1, \dots, n\}$  be a training sample from the joint distribution  $(X, Y)$  with  $X \in \mathbb{R}^d$ ,  $Y \in \{0, 1\}$ . Then the empirical estimators  $f_n$  and  $g_n$  of  $\mathcal{V}^2(X_t, Y)$  given by (3.2) and (3.6) are asymptotically equivalent, in the sense that  $\|f_n - g_n\|_\infty \rightarrow 0$  a.s., as  $n \rightarrow \infty$ . Also, the conclusions of Theorem 3.2 remains valid for the estimator  $f_n$ .*

### 3.3. Some theoretical, model-oriented motivation for variable selection and maxima-hunting

The variable selection methods we are considering here for the binary functional classification problem are aimed at selecting *a finite number of variables*. One might think that this is a “too coarse” approach for functional data. Nevertheless, we provide here some theoretical motivation by showing that, in some relevant cases, variable selection is “the best we can do” in the sense that, in some relevant models, the Bayes rule (i.e., the optimal classifier) has an expression of type  $g^*(X) = h(X(t_1), \dots, X(t_d))$ , so that it depends only on a finite (typically small) number of variables. In fact, in many situations, a proper variable selection leads to an improvement in efficiency (with respect to the baseline option of using the full sample curves), due to the gains associated with a smaller noise level.

The distribution of  $X(t)|Y = i$ , will be denoted by  $P_i$  for  $i = 0, 1$ . In all the examples below the considered processes are Gaussian, i.e., for all  $t_1, \dots, t_m \in [0, 1]$ , with  $m \in \mathbb{N}$ , the finite-dimensional marginal  $(X(t_1), \dots, X(t_m))|Y = i$

has a normal distribution in  $\mathbb{R}^m$  for  $i = 0, 1$ . Many considered models have non-smooth, Brownian-like trajectories. These models play a very relevant role in statistical applications, in particular to the classification problem; see, e.g., [Lindquist and McKeague \(2009\)](#).

We will follow the same strategy as in the previous chapter, that is, to obtain some specific explicit expressions of optimal rules via the Radon-Nykodim derivatives and the Expression (1.2). Although the same RKHS-based results of Chapter 2 could be used in order to get some of these RN-derivatives, here we use other classical tools aiming at illustrating different approaches. In particular, we will focus on the Cameron-Martin Theorem (see [MörTERS and Peres \(2010, p. 24\)](#)) and some results in [Shepp \(1966\)](#) which allow us to tackle heteroscedastic cases. Then, for the sake of clarity let us now recall some basic notions and results to be used throughout, even though they have already been commented before (see, e.g., [Athreya and Lahiri \(2006, ch. 4\)](#), for further details):  $P_0$  is said to be *absolutely continuous with respect to*  $P_1$  (which is denoted by  $P_0 \ll P_1$ ) if and only if  $P_1(A) = 0$  entails  $P_0(A) = 0$ ,  $A$  being a Borel set in  $\mathcal{C}[0, 1]$ . Two probability measures  $P_0$  and  $P_1$  are said to be *equivalent* if  $P_0 \ll P_1$  and  $P_1 \ll P_0$ ; they are *mutually singular* when there exists a Borelian set  $A$  such that  $P_1(A) = 0$  and  $P_0(A) = 1$ . The so-called *Hajek-Feldman dichotomy* (see [Feldman \(1958\)](#)) states that if  $P_0$  and  $P_1$  are Gaussian, then they are either equivalent or mutually singular. The *Radon-Nikodym Theorem* establishes that  $P_1 \ll P_0$  if and only if there exists a measurable function  $f$  such that  $P_1(A) = \int_A f dP_0$  for all Borel set  $A$ . The function  $f$  (which is unique  $P_0$ -almost surely) is called *Radon-Nikodym derivative of  $P_1$  which respect to  $P_0$* . It is usually represented by  $f = \frac{dP_1}{dP_0}$ .

Finally, in order to obtain the results in this section we need to recall (see [Baíllo et al. \(2011, Thm. 1\)](#)) that

$$\eta(x) = \left[ \frac{1-p}{p} \frac{dP_0}{dP_1}(x) + 1 \right]^{-1}, \quad \text{for } x \in \mathcal{S}, \quad (3.7)$$

where  $\mathcal{S}$  is the common support of  $P_0$  and  $P_1$ , and  $p = \mathbb{P}(Y = 1)$ . This equation provides the expression for the optimal rule  $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$  in some important cases where the Radon-Nikodym derivative is explicitly known.

### *Some examples*

Two non-trivial situations in which the Radon-Nikodym derivatives can be explicitly calculated are those problems where  $P_0$  is the standard Brownian motion  $B(t)$ , and  $P_1$  corresponds to  $B(t)$  plus a stochastic or a linear trend. In both cases the Bayes rule  $g^*$  turns out to depend just on one value of  $t$ . To be more precise,

it has the form  $g^*(X) = h(X(1))$ . This is formally stated in the following results. Proofs can be found in the Appendix.

**Proposition 3.1** (Bayes rule stochastic trend). *Let us assume that  $P_0$  is the distribution of a standard Brownian motion  $B(t)$ ,  $t \in [0, 1]$  and  $P_1$  is the distribution of  $B(t) + \theta t$ , where  $\theta$  is a random variable with distribution  $N(0, 1)$ , independent from  $B$ . Then, the Bayes rule is given by  $g^*(x) = \mathbb{I}_{\{x_1^2 > 4 \log(\frac{\sqrt{2}(1-p)}{p})\}}(x)$ , for all  $x \in \mathcal{C}[0, 1]$ .*

As a particular case, when the prior probabilities of both groups are equal,  $p = 1/2$ , we get  $g^*(x) = 1$  if and only if  $|x_1| > 2\sqrt{\log \sqrt{2}} \approx 1.77$ .

**Proposition 3.2** (Bayes rule linear trend). *Let us assume that  $P_0$  is the distribution of a standard Brownian motion  $B(t)$ ,  $t \in [0, 1]$  and  $P_1$  is the distribution of  $B(t) + ct$ , where  $c \neq 0$  is a constant. Then, for  $x \in \mathcal{C}[0, 1]$  the Bayes rule is given by  $g^*(x) = \mathbb{I}_{\{x_1 > \frac{c}{2} - \frac{1}{c} \log(\frac{p}{1-p})\}}(x)$ , if  $c > 0$ , and  $g^*(x) = \mathbb{I}_{\{x_1 < \frac{c}{2} - \frac{1}{c} \log(\frac{p}{1-p})\}}(x)$ , if  $c < 0$ .*

Before presenting our third example we need some additional notation. Let us now recall the countable family of *Haar functions*,  $\varphi_{m,k} = \sqrt{2^{m-1}} \left[ \mathbb{I}_{(\frac{2k-2}{2^m}, \frac{2k-1}{2^m})} - \mathbb{I}_{(\frac{2k-1}{2^m}, \frac{2k}{2^m})} \right]$ , for  $m, k \in \mathbb{N}$ ,  $1 \leq k \leq 2^{m-1}$ . The family  $\{\varphi_{m,k}\}$  is known to be an orthonormal basis in  $L^2[0, 1]$ . Moreover, define the “peak” functions  $\Phi_{m,k}$  by

$$\Phi_{m,k}(t) = \int_0^t \varphi_{m,k}(s) ds. \quad (3.8)$$

We want to use these peak functions to define the trend of the  $P_1$  distribution in another model of type “Brownian versus Brownian plus trend”. In this case the Bayes rule depends just on three points.

**Proposition 3.3** (Bayes rule “peak” trend). *Let us assume that  $P_0$  is the distribution of a standard Brownian motion  $B(t)$ ,  $t \in [0, 1]$  and  $P_1$  is the distribution of  $B(t) + \Phi_{m,k}(t)$ , where  $\Phi_{m,k}$  is one of the peak functions defined above. Then, for  $x \in \mathcal{C}[0, 1]$  the regression function  $\eta(x) = \mathbb{E}(Y|X = x)$  is*

$$\eta(x) = \left\{ \frac{1-p}{p} \exp \left( \frac{1}{2} - 2^{\frac{m-1}{2}} \left[ \left( x \frac{2k-1}{2^m} - x \frac{2k-2}{2^m} \right) + \left( x \frac{2k-1}{2^m} - x \frac{2k}{2^m} \right) \right] \right) + 1 \right\}^{-1} \quad (3.9)$$

and the Bayes rule  $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$  fulfils  $g^*(x) = 1$  if and only if

$$\left( x \frac{2k-1}{2^m} - x \frac{2k-2}{2^m} \right) + \left( x \frac{2k-1}{2^m} - x \frac{2k}{2^m} \right) > \frac{1}{\sqrt{2^{m+1}}} - \frac{1}{\sqrt{2^{m-1}}} \log \left( \frac{p}{1-p} \right). \quad (3.10)$$

It can be seen (MörTERS and Peres (2010, p. 28)) that  $\{\Phi_{m,k}\}$  is an orthonormal basis for the Dirichlet space  $\mathcal{D}[0, 1]$ . Let us recall that, according to Cameron-Martin Theorem, in order to get the equivalence of  $P_1$  and  $P_0$  the trend function is required to belong to that Dirichlet (which is exactly the same condition required in Theorem 2.1 since  $\mathcal{D}$  is the  $\mathcal{H}_K$  associated to the Brownian motion).

*A heteroskedastic case. Shepp's approach.*

The purpose of this paragraph is to show that some results in Shepp (1966) can be also used to give explicit expressions for the optimal classification rule in some significant particular cases of the general problem (2.3), which include discrimination between non-homoscedastic models.

**Theorem 3.3** (Bayes rule under heteroscedasticity). *Let us consider the classification problem (2.3). Let us denote by  $g(x) = \mathbb{I}_{\{\eta^*(x) > 0\}}$  the Bayes rule.*

- (a) *If  $m_0 \equiv 0$ ,  $m_1$  satisfies (2.2),  $\epsilon_0$  is the standard Brownian motion on  $[0, T]$ , with  $T < 1$ , and  $\epsilon_1$  is the standard Brownian bridge on  $[0, T]$ , then*

$$\eta^*(X) = -\frac{1}{2} \log(1-T) - \frac{TX(T)^2 + m_1(T)^2 - 2m_1(T)X(T)}{2T(1-T)} - \log\left(\frac{1-p}{p}\right). \quad (3.11)$$

- (b) *If the noise processes  $\epsilon_0, \epsilon_1$  are both standard Brownian bridges on  $[0, T]$  with  $T < 1$ , and both  $m_0$  and  $m_1$  satisfy (2.2), then*

$$\eta^*(X) = \frac{(X(T) - m_0(T))^2 - (X(T) - m_1(T))^2}{2T(1-T)} - \log\left(\frac{1-p}{p}\right). \quad (3.12)$$

Notice that when  $p = 1/2$ , the rule  $\mathbb{I}_{\{\eta^*(x) > 0\}}$  for (b) reduces to the indicator of

$$|X(T) - m_0(T)| - |X(T) - m_1(T)| > 0.$$

In addition, if  $m_1 \equiv 0$  (that is, no trend in the Brownian bridge), the Bayes rule in (a) reduces to just the indicator of

$$X(T)^2 < T(T-1)\log(1-T).$$

*Remark 3.1* (Additional examples). Analogous calculations can be performed (still obtaining explicit expressions for the Bayes rule of type  $g^*(x) = g(x(t_1), \dots, x(t_d))$ ), using a rescaled Brownian motion  $\sigma B(t)$  or a piecewise linear trend instead of these (see Remark 2.1). Likewise, other models could be obtained by linear combinations in the trend functions or by finite mixtures of other simpler models. Many of them have been included in the simulation study of Section 3.4.

Next, we will provide some theoretical support for the maxima-hunting method, by showing that in some specific useful models the optimal classification rule depends on the maxima of the distance covariance function  $\mathcal{V}^2(X_t, Y)$ , although in some particular examples, other points (closely linked to the maxima) are also relevant.

**Proposition 3.4** (Maxima of  $\mathcal{V}^2$ ). *Under the models assumed in Propositions 3.1 and 3.2, the corresponding distance covariance functions  $\mathcal{V}^2(X_t, Y)$  have both a unique relative maximum at the point  $t = 1$ .*

The model considered in Proposition 3.1 provides a clear example of the advantages of using the distance covariance measure  $\mathcal{V}^2(X_t, Y)$  rather than the ordinary covariance  $\text{Cov}^2(X_t, Y)$  in the maxima-hunting procedure. Indeed, note that in this case,  $\text{Cov}^2(X_t, Y) = p^2(1-p)^2(\mathbb{E}(X(t)|Y=0) - \mathbb{E}(X(t)|Y=1))^2 = 0$ , for all  $t \in [0, 1]$ , so that the ordinary covariance is useless to detect any difference between the values of  $t$ .

*Remark 3.2* (Other examples). Other similar results could be obtained for the models considered in Proposition 3.3 and Theorem 3.3.

Let us finally show a simple useful result valid for those cases in which there is only one relevant point. This means that the Bayes rule only depends on the trajectory  $\{X_t : t \in [0, 1]\}$  through the value of  $X(t^*)$ . The following result shows that under fairly general conditions  $\mathcal{V}^2(X_t, Y) < \mathcal{V}^2(X_{t^*}, Y)$ , for all  $t > t^*$ . Hence, if we use the global maximum of  $\mathcal{V}_X^2$  as a criterion to select the relevant point, we will never choose any point greater than  $t^*$ . Of course, it would be desirable to find mild conditions under which  $\mathcal{V}^2(X_t, Y) < \mathcal{V}^2(X_{t^*}, Y)$ , for all  $t < t^*$ . However, as far as we know, this is still an open problem.

**Proposition 3.5** (Global maximum of  $\mathcal{V}^2$ ). *Assume the process  $X_t$  has independent and non-degenerate increments. Assume also that there exists a function  $h : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\eta(X) = h(X_{t^*})$ . Then,  $\mathcal{V}^2(X_t, Y) < \mathcal{V}^2(X_{t^*}, Y)$ , for all  $t > t^*$ .*

Note that this result would apply, for example, to cases similar to those considered in Propositions 1 and 2 in the paper, provided that the argument  $t$  is replaced with  $1-t$ ; in those cases one would have  $t^* = 0$  and this would be the global maximum. Another possible example of this situation of unique maximum is given by some logistic models.

### 3.4. Empirical study

The goal of this section is to assess the performance of the maxima hunting methodology when compared with other reliable competitors. This is carried out by means of a extensive simulation study plus three selected real data examples. The study includes some models such as M2,...,M6 and G5,...,G8 for which some relevant variables do not correspond to maxima. Also, there is no reason to think that the many logistic-type models (and the real data examples) included in our experiments, are especially favorable to our proposals.

Let us recall again that common elements (models, methods, data sets) and methodological details are fully explained in Chapter 5, and the full list of simulation models is in Appendix A. However, all of the common elements involved in these experiments are briefly described for self-contained and clarity purposes.

#### 3.4.1. The variable selection methods under study. Criteria for comparisons

These are the methods, and their corresponding notations as they appear in the tables and figures below. The implementation details are given in Section 5.1.

1. **Maxima-hunting.** The methods based on the estimation of the maxima of  $\mathcal{R}_X^2$  and  $\mathcal{V}_X^2$  are implemented as follows. The functional data  $x(t)$ ,  $t \in [0, 1]$  are discretized to  $(x(t_1), \dots, x(t_N))$ , so a non-trivial practical problem is to decide which points in the grid are the local maxima: a point  $t_i$  is declared to be a local maximum when it is the highest local maximum on the sub-grid  $\{t_j\}$ ,  $j = i - h \dots, i + h$ . The proper choice of  $h$  depends on the nature and discretization pattern of the data at hand. Thus,  $h$  could be considered as a smoothing parameter to be selected in an approximately optimal way. In our experiments  $h$  is chosen by a validation step explained in next section.

Then, we sort the maxima  $t_i$  by **relevance** (the value of the function at  $t_i$ ). This seems to be the natural order and it produces better results than other simple sorting strategies. We denote these maxima-hunting methods by **MHR** and **MHV** depending on the use of  $\mathcal{R}_X^2$  or  $\mathcal{V}_X^2$ . This relevance criterion and an alternative domain criterion (sorting by the length of the interval where the maximum is global maximum) are illustrated in Figure 3.2. Our empirical results (not included in this study) show that the use of this domain criterion does not lead, on average, to any improvement with respect to the relevance ordering.

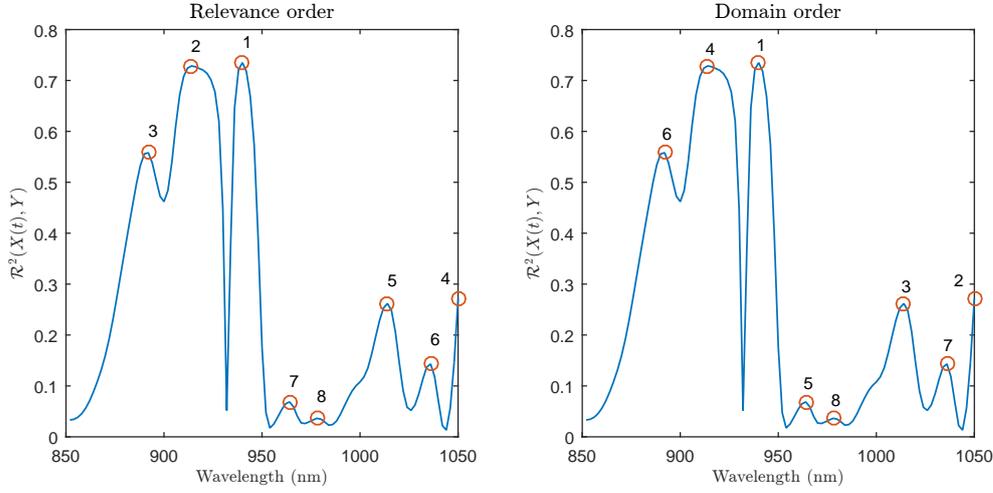


Figure 3.2: Blue line stands for  $\mathcal{R}^2(X(t), Y)$  for the first derivative of the Tecator data. The maxima are marked in red and the selection order is indicated by the number beside each maximum. On the left picture it is used the relevance criterion while the domain one is applied on the right graph. In this noiseless case identification by relevance is preferable.

2. **Univariate  $t$ -ranking method**, denoted by **T**, is frequently used when selecting relevant variables (see e.g. the review by [Fan and Lv \(2010\)](#)). It is based on the simple idea of selecting the variables  $X_t$  with highest Student's  $t$  two-sample scores  $T(X_t) = |\bar{X}_{1t} - \bar{X}_{0t}| / \sqrt{s_{1t}^2/n_1 + s_{0t}^2/n_0}$ . We include this ranking method in order to evaluate in practice the supposed disadvantage of univariate methods.

3. **mRMR**. The minimum Redundancy Maximum Relevance algorithm, is a relevant intrinsic variable selection method that will be widely explained in Chapter 4. We have considered mRMR as a natural competitor for our maxima-hunting approximation. We have computed both Fisher-Correlation and Mutual Information approaches given in the former paper by [Ding and Peng \(2005\)](#). We have also considered both difference and quotient criteria. For the sake of clarity we only show here the results of **FCQ** (Fisher Correlation Quotient) and **MID** (Mutual Information Difference) which outperform on average their corresponding counterparts.

4. **PLS**. Partial least squares is a well-known dimension reduction technique based on linear projections; see e.g. [Delaigle and Hall \(2012b\)](#) and references therein.

5. **Base**. The  $k$ -NN classifier is applied to the entire curves. The Base performance can be seen as a reference to assess the usefulness of dimension reduction

methods. Somewhat surprisingly, Base is often outperformed.

The **classifiers** used in all cases are either  $k$ -NN, based on the Euclidean distance or LDA (applied to the selected variables). Note that the Base method cannot be implemented with LDA since this classifier typically fails with infinite or high-dimensional data (see Section 1.2). Similar comparisons could be done with other classifiers, since the considered methods do not depend on the classifier. For comparing the different methods we use the natural accuracy measure, defined by the percentage of correct classification.

### 3.4.2. The simulation study

In this study we consider all the 100 models in Appendix A which cover all examples along this Chapter and incorporate logistic-type experiments and mixtures. Although these functional models are fully described in Subsection 5.2.1, let us briefly point out some of their basic characteristics. Trajectories are discretized in 100 equispaced points in the interval  $[0,1]$  and training sample sizes of  $n = 30, 50, 100, 200$  are considered for each model. Classification accuracy is assessed by means of an independent test sample of 200 observations. The number of variables and the classification parameters (if needed) are set through another independent validation sample of 200 curves.

The complete simulation outputs can be downloaded from [www.uam.es/antonio.cuevas/exp/outputs.xlsx](http://www.uam.es/antonio.cuevas/exp/outputs.xlsx). A summary of the 400 experiments (100 models  $\times$  4 samples sizes) grouped by sample size is presented in Tables 3.1 (for  $k$ -NN outputs) and 3.2 (for LDA outputs). The methods under study are in columns and each row contains the averages on 100 models (averaged, in turn, over 200 independent runs) with a specific sample size and classifier. Different measures are presented in rows and methods in columns. The row entries ‘Average accuracy’ provide the average percentage of correct classification. The rows ‘Average dim. red.’ stand for the average number of selected variables. The number of models in which each method beats the ‘Base’ benchmark procedure is given in ‘Victories over Base’ rows. This last measures are not shown in Table 3.2 since “Base” method cannot be computed with LDA. Additionally, in order to give an insight of what happen in the concrete models we have selected (with no particular criterion in mind) a sampling of just a few examples among the 400 experiments. The reader can consult the Excel tables available online with the entire results, if interested on some particular model. Table 3.3 provides the performance (averaged on 200 runs) measured in terms of classification accuracy. Models are presented in rows and methods in columns. The marked outputs in all the three tables correspond to the winner and second best method in each row.

Table 3.1: Performance outputs for the considered methods, using  $k$ -NN and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output ( $k$ -NN)	$n$	Methods						
		FCQ	MID	T	PLS	MHR	MHV	Base
Average accuracy	30	79.65	80.09	79.16	81.42	81.87	81.53	78.98
	50	80.40	81.43	79.84	82.48	82.89	82.59	80.34
	100	81.34	83.01	80.71	83.79	84.21	83.87	81.99
	200	82.09	84.28	81.27	84.84	85.37	84.96	83.38
Average dim. red	30	9.5	9.2	9.9	4.3	6.2	6.3	100
	50	9.6	9.38	10.1	4.8	6.2	6.2	100
	100	9.9	9.6	10.3	5.5	6.1	6.1	100
	200	10.1	9.8	10.4	6.2	5.8	5.8	100
Victories over Base	30	58	71	51	77	95	89	-
	50	53	71	46	76	91	89	-
	100	49	71	38	77	86	81	-
	200	42	73	33	72	80	75	-

This summary of the complete results allow us to draw some general considerations about the performance of the methods. Also, outputs of Table 3.3 are more or less representative of the overall conclusions of the entire study. For instance, MHR appears as the overall winner on average with a slight advantage. PLS and the maxima-hunting methods (MHR and MHV) obtain similar scores and clearly outperform the other benchmark methods. Note that they also beat (often very clearly) the Base method in almost all cases using just a few variables. This shows that dimension reduction is, in fact, “mandatory” in many cases. Note that these methods obtain improvements close to 2% of the total accuracy with just the 5-6% of original variables.

Regarding the comparison of  $k$ -NN and LDA in the second stage (after dimension reduction) the results show a slight advantage for  $k$ -NN (on average). The complete failure of LDA in models G1 and G3 was to be expected since in these cases the mean functions are identical in both populations. In terms of number of variables, when  $k$ -NN is used, MHR and MHV need less variables to achieve better results than the rest of variable selection methods. When LDA is used, the number of required variables is quite similar in all methods. Table 3.1 also shows that the benefits of reducing the dimension (compared to the Base approach) are higher when lower sample sizes are considered. This is a relevant fact since in

Table 3.2: Performance outputs for the considered methods, using LDA and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output LDA	$n$	Methods						
		FCQ	MID	T	PLS	MHR	MHV	Base
Average accuracy	30	77.58	78.72	76.77	81.04	80.66	80.71	-
	50	78.53	80.28	77.77	81.86	81.81	81.73	-
	100	79.62	81.85	78.93	82.71	82.99	82.81	-
	200	80.47	82.96	79.83	83.39	83.83	83.53	-
Average dim. red	30	4.7	5.6	4.9	2.7	5.5	5.4	-
	50	5.7	6.5	5.9	3.0	6.1	6.1	-
	100	7.1	7.9	7.4	3.5	7.0	7.0	-
	200	8.3	9.0	8.9	4.2	7.5	7.5	-

many practical cases (e.g. in biomedical studies) only small samples are available.

### 3.4.3. Real data examples

We have chosen three examples due to their popularity in FDA. There are many references on these datasets so we will just give brief descriptions of them; additional details can be found in Section 5.3.

*Berkeley Growth Data.* The heights of 54 girls and 39 boys measured at 31 non equidistant time points. See, e.g., [Ramsay and Silverman \(2005\)](#).

*Tecator.* 215 near-infrared absorbance spectra (100 grid points each) of finely chopped meat, obtained using a Tecator Infratec Food & Feed Analyzer. The sample is separated in two classes according to the fat content (smaller or larger than 20%). Tecator curves are often used in a differentiated version. We use here the second derivatives. See [Ferraty and Vieu \(2006\)](#) for details.

*Phoneme.* As usually we use the “binary” version of these data corresponding to log-periodograms constructed from 32 ms long recordings of males pronouncing the phonemes “aa” and “ao”. The sample size is  $n = 1717$  (695 from “aa” and 1022 from “ao”). Each curve was observed at 256 equispaced points (distinct from the previous chapter, here we use the entire curves).

In the comparisons with real data sets we have incorporated the method re-

Table 3.3: Average correct classification outputs, over 200 runs, with  $n = 50$ .

<i>k</i> -NN outputs							
Models	FCQ	MID	T	PLS	MHR	MHV	Base
L2.OUt	82.47	82.11	81.68	83.27	83.22	83.23	82.60
L6.OU	88.41	89.81	86.19	90.93	90.75	90.83	90.56
L10.B	81.09	85.02	81.13	85.90	87.27	87.42	85.46
L11.ssB	82.31	80.85	82.28	78.81	83.10	82.81	79.89
L12.sB	77.24	75.83	77.41	74.92	78.57	76.62	74.78
G1	65.86	70.70	65.57	66.95	71.59	71.80	70.10
G3	63.09	73.39	60.57	60.56	77.47	77.06	65.26
G6	84.27	91.95	84.14	93.67	93.38	93.71	92.19
M2	70.77	69.82	69.16	78.16	74.76	75.68	71.14
M6	81.15	83.08	79.73	83.47	83.32	83.35	80.99
M10	64.93	68.33	64.58	68.25	70.66	70.94	68.95

LDA outputs							
Models	FCQ	MID	T	PLS	MHR	MHV	Base
L2.OUt	79.80	78.95	78.23	80.07	80.24	80.14	-
L6.OU	87.79	88.91	84.46	91.01	89.44	89.35	-
L10.B	75.97	75.44	76.04	77.60	77.63	77.76	-
L11.ssB	80.95	80.09	80.81	79.39	81.88	81.63	-
L12.sB	76.39	75.20	76.40	75.02	77.38	75.96	-
G1	51.27	51.24	51.20	51.44	51.55	51.70	-
G3	51.09	52.26	50.96	50.35	52.95	52.69	-
G6	87.72	95.28	87.80	97.77	96.54	96.85	-
M2	67.44	76.51	66.81	84.38	82.24	83.06	-
M6	79.99	79.92	79.63	81.39	81.08	81.38	-
M10	60.03	65.61	59.24	67.49	67.25	67.99	-

Table 3.4: Classification accuracy (in %) for the real data with both classifiers.

<i>k</i> -NN outputs								
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	83.87	95.70	83.87	94.62	95.70	94.62	-	96.77
Tecator	99.07	99.07	99.07	97.21	99.53	99.53	-	98.60
Phoneme	80.43	79.62	80.43	82.53	80.20	78.86	-	78.97

LDA outputs								
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	91.40	94.62	91.40	95.70	95.70	96.77	96.77	-
Tecator	94.42	95.81	94.42	94.42	95.35	94.88	95.35	-
Phoneme	79.38	80.37	79.09	80.60	80.20	78.92	77.34	-

cently proposed by [Delaigle et al. \(2012\)](#). We denote it by DHB. Given a classifier, the DHB 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. So the authors implement a slightly modified version, which we have closely followed. It is based on a sort of trade-off between full and sequential search, together with some additional computational savings. Let us note, as an important difference with our maxima-hunting method, that the DHB procedure is a “wrapper” method, in the sense that it depends on the chosen classifier. Following [Delaigle et al. \(2012\)](#), we have only implemented the DHB method with the LDA classifier.

Apart from that, we proceed as in the simulation study except for the generation of the training, validation and test samples. Here we consider the usual cross-validation procedure which avoids splitting the sample (sometimes small) into three different sets. Each output is obtained by standard leave-one-out cross-validation. The only exception is the phoneme data set for which this procedure is extremely time-consuming (due to the large sample size); so we use instead ten-fold cross-validation (10CV). The respective validation steps are done with the same resampling schemes within the training samples. This is a usual way to proceed when working with real data; see [Hastie et al. \(2009, Subsection 7.10\)](#). Several outputs are given in Tables 2 (accuracy) and 3 (number of variables) below. The complete results can be found in [www.uam.es/antonio.cuevas/exp/outputs.xlsx](http://www.uam.es/antonio.cuevas/exp/outputs.xlsx).

Table 3.5: Average number of variables (or components) selected for the real data sets.

<i>k</i> -NN outputs								
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	1.0	3.5	1.0	2.8	4.0	4.0	-	31
Tecator	3.0	5.7	3.0	2.7	1.0	1.0	-	100
Phoneme	10.7	15.3	12.3	12.9	10.2	12.3	-	256

LDA outputs								
Data	FCQ	MID	T	PLS	MHR	MHV	DHB	Base
Growth	5.0	3.4	5.0	2.0	4.0	4.0	2.3	-
Tecator	8.4	2.6	3.1	9.7	1.7	1.8	3.0	-
Phoneme	8.5	17.1	7.9	15.5	16.1	11.0	2.0	-

These results are similar to those obtained in the simulation study. While (as expected) there is no clear global winner, maxima-hunting method looks like a very competitive choice. In particular, Tecator outputs are striking, since MHR and MHV achieve (with *k*-NN) a near perfect classification with just one variable. Note also that maxima-hunting methods (particularly MHR) outperform or are very close to the Base outputs (which uses the entire curves). PLS is overcome by our methods in two of the three problems but it is the clear winner in phoneme example. In any case, it should be kept in mind, as a counterpart, the ease of interpretability of the variable selection methods.

The DHB method performs well in the two first considered examples but relatively fails in the phoneme case. There is maybe some room for improvement in the stopping criterion (recall that we have used the same parameters as in [De-laigle et al. \(2012\)](#)). Recall also that, by construction, this is (in the machine learning terminology) a “wrapper” method. This means that the variables selected by DHB are specific for the LDA classifier (and might dramatically change with other classification rules). Also note that the use of the LDA classifier didn’t lead to any significant gain; in fact, the results are globally worse than those of *k*-NN except for a few particular cases.

Although our methodology is not primarily targeted to the best classification rate, but to the choice of the most representative variables, we can conclude that MH procedures combined with the simple *k*-NN are competitive when compared with PLS and other successful and sophisticated methods in literature: see [Galeano et al. \(2014\)](#) for Tecator data, [Mosler and Mozharovskiy \(2014\)](#) for

growth data and [Delaigle et al. \(2012\)](#) for phoneme data.

### 3.5. Overall conclusions: a tentative global ranking of methods

We have summarized the conclusions of our 400 simulation experiments in three rankings, prepared with different criteria, according to classification accuracy. With the *relative ranking* criterion, the winner method (with performance  $W$ ) in each of the 400 experiments gets 10 score points, and the method with the worst performance (say  $w$ ) gets 0 points. The score of any other method, with performance  $u$  is just assigned in a proportional way:  $10(u - w)/(W - w)$ . The *positional ranking* scoring criterion just gives 10 points to the winner in every experiment, 9 points to the second one, etc. Finally, the *F1 ranking* rewards strongly the winner. For each experiment, points are divided as in an F1 Grand Prix: the winner gets 25 points and the rest 18, 15, 10, 8, 6 and 4 successively. The final average scores are given in [Table 3.6](#) grouped by ranking type and sample size. The winner and the second best methods in each category appear marked. Also, a graphical representation of the relative ranking scores for all the 400 simulation experiments is shown in [Figure 3.3](#)

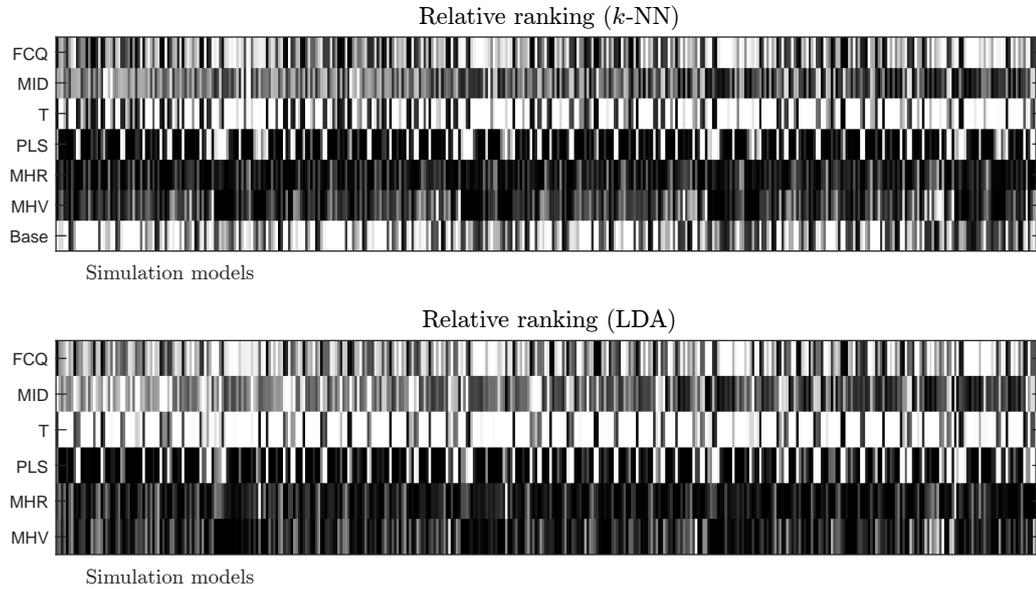
The results are self-explanatory and are in accordance with previous conclusions. Nevertheless, the following remarks might be of some interest for practitioners:

1. The maxima-hunting methods are the global winners (in particular when using the distance correlation measure), even if there is still room for improvement in the maxima identification. In fact, the maxima-hunting procedures result in accuracy improvements (with respect to the “base error”, i.e., using the whole trajectories) in 88.00% of the considered experiments. Overall, the gain of accuracy associated with MHR variable selection is relevant (2.41%).

2. While the univariate ranking methods, such as the  $t$  ranking, (which ignore the dependence between the involved variables) are still quite popular among practitioners, they are clearly outperformed by the “functional” procedures. It is quite remarkable the superiority of the maxima-hunting methods on the rest of variable selection procedures, requiring often a lesser number of variables.

Table 3.6: Final scores of the considered methods for the simulation experiments. The rankings correspond to the observed performances in classification accuracy. The individual scores are in turn combined according to three different ranking criteria (proportional, positional and F1).

<i>k</i> -NN rankings								
Ranking type	<i>n</i>	FCQ	MID	T	PLS	MHR	MHV	Base
Relative	30	4.66	4.79	3.61	6.94	8.64	7.64	2.68
	50	4.62	5.45	3.25	6.94	8.50	7.48	3.25
	100	4.37	6.23	2.71	7.06	8.35	7.21	3.97
	200	4.04	6.72	2.15	7.02	8.19	7.06	4.64
Positional	30	6.52	6.22	5.59	7.93	9.09	8.06	5.59
	50	6.55	6.50	5.64	7.90	8.72	7.95	5.74
	100	6.42	6.83	5.48	8.03	8.58	7.72	5.98
	200	6.26	7.30	5.27	7.96	8.34	7.62	6.25
F1	30	11.64	10.58	9.54	17.37	19.55	15.93	9.39
	50	12.01	11.27	9.77	17.29	18.12	15.80	9.74
	100	11.58	12.39	9.51	17.71	17.46	15.03	10.41
	200	11.24	13.90	9.01	17.19	16.71	14.89	11.06
LDA rankings								
Ranking type	<i>n</i>	FCQ	MID	T	PLS	MHR	MHV	Base
Relative	30	3.57	3.46	1.79	7.60	8.15	8.11	-
	50	3.74	4.61	1.89	7.20	8.60	8.16	-
	100	3.83	5.95	1.90	6.70	8.96	8.18	-
	200	3.89	6.74	2.27	6.09	8.78	7.83	-
Positional	30	6.75	6.51	5.71	8.54	8.75	8.74	-
	50	6.72	6.71	5.87	8.39	8.80	8.52	-
	100	6.72	7.15	5.92	7.95	8.79	8.47	-
	200	6.62	7.58	6.18	7.63	8.81	8.23	-
F1	30	11.96	11.12	9.58	19.08	17.95	18.31	-
	50	11.91	11.68	10.12	18.57	18.33	17.41	-
	100	12.20	12.92	10.24	16.64	18.58	17.42	-
	200	11.74	14.35	10.92	15.66	18.76	16.74	-



**Figure 3.3:** Display of relative ranking scores, the darker the better (black corresponds to 10 and white 0). Each column represents a simulation model and each file corresponds to a dimension reduction method. The ranking outputs are obtained with both  $k$ -NN (first display) and LDA (second display) classifiers. Maxima-hunting with  $\mathcal{R}$  is often the best and never the worst.

3. As an important overall conclusion, variable selection appears as a highly competitive alternative to PLS, which is so far the standard dimension reduction method in high-dimensional and functional statistics (whenever a response variable is involved). The results of the above rankings show that variable selection offers a better balance in terms of both accuracy and interpretability.

4. On average, the use of the classical Fisher’s discriminant rule LDA (after dimension reduction) provides worse results than the nonparametric  $k$ -NN rule. There is an apparent contradiction since examples of superiority of a linear classifier are shown in Chapter 2 and [Delaigle and Hall \(2012a\)](#) where asymptotic optimality results are provided. In addition, under some conditions, the proposed classifiers turns out to be “near-perfect” (in the sense that the probability of classification error can be made arbitrarily small) to discriminate between two Gaussian processes. However, it requires several conditions which are not fulfilled in most considered models.

*A final remark.* The present study shows that there are several quite natural models in which the maxima-hunting method is definitely to be recommended. The real data results are also encouraging. Our results suggest that, even when there is no clear, well-founded guess on the nature of the underlying model, the idea of

selecting the maxima of the distance correlation is a suitable choice, that always allows for a direct interpretation. Note that, even if some relevant variables didn't appear as maxima of the distance correlation function (such as in our benchmark), our MH procedure works, in the sense of providing a few meaningful variables, highly related with the response, and not redundant. It is also natural to ask what type of models would typically be less favourable for the maxima-hunting approach. As a rough, practical guide, we might say that some adverse situations might typically arise in those cases where the trajectories are extremely smooth, or when they are very wiggly, with many noisy abrupt peaks which tend to mislead the calculation of the maxima in the distance correlation function.

### 3.6. Some additional results and proofs

To prove Theorem 3.2 we need two lemmas dealing with the uniform strong consistency of one-sample and two-sample functional U-statistics, respectively.

**Lemma 3.2** (Uniform convergence of one-sample U-statistics). *Let  $X : T \rightarrow \mathbb{R}$  be a process with continuous trajectories a.s. defined on the compact rectangle  $T = \prod_{i=1}^d [a_i, b_i] \subset \mathbb{R}^d$ . Let  $X_1, \dots, X_n$  be a sample of  $n$  independent trajectories of  $X$ . Define the functional U-statistic*

$$U_n(t) = \frac{2}{n(n-1)} \sum_{i < j} k[X_i(t), X_j(t)],$$

where the kernel  $k$  is a real continuous, permutation symmetric function. Assume that

$$\mathbb{E} \left( \sup_{t \in T} |k[X(t), X'(t)]| \right) < \infty,$$

where  $X$  and  $X'$  denote two independent copies of the process. Then, as  $n \rightarrow \infty$ ,  $\|U_n - U\|_\infty \rightarrow 0$ , a.s., where  $U(t) = \mathbb{E}(k[X(t), X'(t)])$ .

*Proof.* First, we show that  $U(t)$  is continuous. Let  $t_n \subset T$  such that  $t_n \rightarrow t$ . Then, due to the continuity assumptions on the process and the kernel,  $k[X(t_n), X'(t_n)] \rightarrow k[X(t), X'(t)]$ , a.s. Using the assumption  $\mathbb{E} \left( \sup_{t \in T} |k[X(t), X'(t)]| \right) < \infty$ , Dominated Convergence Theorem (DCT) allows us to deduce  $U(t_n) \rightarrow U(t)$ .

Let  $M_\delta(t) = \sup_{s: |s-t|_d \leq \delta} |h(s) - h(t)|$  where, for the sake of simplicity, we denote  $h(t) = k[X(t), X'(t)]$ . The next step is to prove that, as  $\delta \downarrow 0$ ,

$$\sup_{t \in T} \mathbb{E}(M_\delta(t)) \rightarrow 0. \quad (3.13)$$

Both  $M_\delta(t)$  and  $\lambda_\delta(t) = \mathbb{E}(M_\delta(t))$  are continuous functions. Since  $h(t)$  is uniformly continuous on  $\{s : |s - t|_d \leq \delta\}$ ,  $M_\delta(t)$  is also continuous. The fact that  $\lambda_\delta(t)$  is continuous follows directly from DCT since  $|M_\delta(t)| \leq 2 \sup_{t \in T} |h(t)|$  and, by assumption,  $\mathbb{E}(\sup_{t \in T} |h(t)|) < \infty$ . By continuity,  $M_\delta(t) \rightarrow 0$  and  $\lambda_\delta(t) \rightarrow 0$ , as  $\delta \downarrow 0$ . Now, since  $\delta > \delta'$  implies  $\lambda_\delta(t) \geq \lambda_{\delta'}(t)$ , for all  $t \in T$ , we can apply Dini's Theorem to deduce that  $\lambda_\delta(t)$  converges uniformly to 0, that is,  $\sup_{t \in T} \lambda_\delta(t) \rightarrow 0$ , as  $\delta \downarrow 0$ .

The last step is to show  $\|U_n - U\|_\infty \rightarrow 0$  a.s., as  $n \rightarrow \infty$ . For  $i \neq j$ , denote  $M_{ij,\delta}(t) = \sup_{s: |s-t|_d < \delta} |h_{ij}(s) - h_{ij}(t)|$ , where  $h_{ij}(t) = k[X_i(t), X_j(t)]$ , and  $\lambda_\delta(t) = \mathbb{E}(M_{ij,\delta}(t))$ . Fix  $\epsilon > 0$ . By (3.13), there exists  $\delta > 0$  such that  $\lambda_\delta(t) < \epsilon$ , for all  $t \in T$ . Now, since  $T$  is compact, there exist  $t_1, \dots, t_m$  in  $T$  such that  $T = \cup_{k=1}^m B_k$ , where  $B_k = \{t : |t - t_k|_d \leq \delta\} \cap T$ . Then,

$$\begin{aligned} \|U_n - U\|_\infty &= \max_{1 \leq k \leq m} \sup_{t \in B_k} |U_n(t) - U(t)| \\ &\leq \max_{1 \leq k \leq m} \sup_{t \in B_k} [|U_n(t) - U_n(t_k)| + |U_n(t_k) - U(t_k)| + |U(t_k) - U(t)|] \\ &\leq \max_{1 \leq k \leq m} \sup_{t \in B_k} |U_n(t) - U_n(t_k)| + \max_{k=1, \dots, m} |U_n(t_k) - U(t_k)| + \epsilon, \end{aligned}$$

since  $|s - t|_d \leq \delta$  implies  $|U(s) - U(t)| = |\mathbb{E}[h(s) - h(t)]| \leq \mathbb{E}|h(s) - h(t)| \leq \lambda_\delta(t) < \epsilon$ .

For the second term, we have  $\max_{k=1, \dots, m} |U_n(t_k) - U(t_k)| \rightarrow 0$  a.s., as  $n \rightarrow \infty$ , applying SLLN for U-statistics (see e.g. [DsGupta \(2008, Theorem 15.3\(b\), p. 230\)](#)). As for the first term, observe that using again SLLN for U-statistics,

$$\begin{aligned} \sup_{t \in B_k} |U_n(t) - U_n(t_k)| &\leq \frac{2}{n(n-1)} \sum_{i < j} \sup_{t \in B_k} |h_{ij}(t_k) - h_{ij}(t)| \\ &= \frac{2}{n(n-1)} \sum_{i < j} M_{ij,\delta}(t_k) \rightarrow \lambda_\delta(t_k), \quad \text{a.s.,} \end{aligned}$$

where  $\lambda_\delta(t_k) < \epsilon$ . Therefore,

$$\begin{aligned} \limsup_n \|U_n - U\|_\infty &\leq \limsup_n \max_{k=1, \dots, m} \sup_{t \in B_k} |U_n(t) - U_n(t_k)| \\ &\quad + \limsup_n \max_{k=1, \dots, m} |U_n(t_k) - U(t_k)| + \epsilon \leq 2\epsilon. \end{aligned}$$

□

**Lemma 3.3** (Uniform convergence of two-sample U-statistics). *Let  $X^{(0)} : T \rightarrow \mathbb{R}$  and  $X^{(1)} : T \rightarrow \mathbb{R}$  be a pair of independent processes with continuous trajectories a.s. defined on the compact rectangle  $T = \prod_{i=1}^d [a_i, b_i] \subset \mathbb{R}^d$ . Let  $X_1^{(0)}, \dots, X_{n_0}^{(0)}$  and  $X_1^{(1)}, \dots, X_{n_1}^{(1)}$  be samples of  $n_0$  and  $n_1$  independent trajectories of  $X^{(0)}$  and  $X^{(1)}$ , respectively. Define the functional two-sample U-statistic*

$$U_{n_0, n_1}(t) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} k[X_i^{(0)}(t), X_j^{(1)}(t)],$$

where the kernel  $k$  is a continuous, permutation symmetric function. Assume that

$$\mathbb{E}\left(\sup_{t \in T} |h(t)| \log^+ |h(t)|\right) < \infty,$$

with  $h(t) = k[X^{(0)}(t), X^{(1)}(t)]$ . Then, as  $\min(n_0, n_1) \rightarrow \infty$ ,

$$\|U_{n_0, n_1} - U\|_\infty \rightarrow 0, \quad \text{a.s.},$$

where  $U(t) = \mathbb{E}(k[X^{(0)}(t), X^{(1)}(t)])$ .

*Proof.* It is analogous to the proof of Lemma 3.2 so it is omitted. We need to apply a strong law of large numbers for two-sample U-statistics. This result can be guaranteed under slightly stronger conditions on the moments of the kernel; see Sen (1977, Thm.1). Hence the condition  $\mathbb{E}\left(\sup_{t \in T} |h(t)| \log^+ |h(t)|\right) < \infty$  in the statement of the lemma.  $\square$

## Proofs of the main results

*Proof of Theorem 3.1.*

(a) From (3.1), as  $X_t$  is  $d$ -dimensional and  $Y$  is one-dimensional, taking into account  $c_1 = \pi$ , we have

$$\begin{aligned} \mathcal{V}^2(X_t, Y) &= \|\varphi_{X_t, Y}(u, v) - \varphi_{X_t}(u)\varphi_Y(v)\|_w^2 \\ &= \frac{1}{\pi c_d} \int_{\mathbb{R}} \int_{\mathbb{R}^d} |\varphi_{X_t, Y}(u, v) - \varphi_{X_t}(u)\varphi_Y(v)|^2 \frac{1}{|u|_d^{d+1} v^2} dudv. \end{aligned}$$

Let's analyze the integrand,

$$\begin{aligned}
\varphi_{X_t, Y}(u, v) - \varphi_{X_t}(u)\varphi_Y(v) &= \mathbb{E} \left[ e^{iu^\top X_t} e^{ivY} \right] - \mathbb{E} \left[ e^{iu^\top X_t} \right] \mathbb{E} \left[ e^{ivY} \right] \\
&= \mathbb{E} \left[ (e^{iu^\top X_t} - \varphi_{X_t}(u))(e^{ivY} - \varphi_Y(v)) \right] \\
&= \mathbb{E} \left[ \mathbb{E} \left[ (e^{iu^\top X_t} - \varphi_{X_t}(u))(e^{ivY} - \varphi_Y(v)) \mid X \right] \right] \\
&= \mathbb{E} \left[ (e^{iu^\top X_t} - \varphi_{X_t}(u)) \mathbb{E} \left[ (e^{ivY} - \varphi_Y(v)) \mid X \right] \right] \\
&\stackrel{(*)}{=} \mathbb{E} \left[ (e^{iu^\top X_t} - \varphi_{X_t}(u))(e^{iv} - 1)(\eta(X) - p) \right] \\
&= (e^{iv} - 1) \mathbb{E} \left[ (e^{iu^\top X_t} - \varphi_{X_t}(u))(\eta(X) - p) \right] \\
&= (e^{iv} - 1) \mathbb{E} \left[ e^{iu^\top X_t}(\eta(X) - p) \right] = (e^{iv} - 1)\zeta(u, t).
\end{aligned}$$

Step (\*) in the above chain of equalities is motivated as follows:

$$\begin{aligned}
\mathbb{E} \left[ (e^{ivY} - \varphi_Y(v)) \mid X \right] &= \mathbb{E} \left[ e^{ivY} \mid X \right] - \varphi_Y(v) = (e^{iv} - 1)\eta(X) - (e^{iv} - 1)p \\
&= (e^{iv} - 1)(\eta(X) - p).
\end{aligned}$$

Therefore, since  $\int_{\mathbb{R}} \frac{|e^{iv} - 1|^2}{\pi v^2} dv = 2$ ,

$$\mathcal{V}^2(X_t, Y) = \int_{\mathbb{R}} \frac{|e^{iv} - 1|^2}{\pi v^2} dv \int_{\mathbb{R}^d} \frac{|\zeta(u, t)|^2}{c_d |u|_d^{d+1}} du = \frac{2}{c_d} \int_{\mathbb{R}^d} \frac{|\zeta(u, t)|^2}{|u|_d^{d+1}} du.$$

(b) Since  $\zeta(u, t) = \mathbb{E} \left[ (\eta(X) - p) e^{iu^\top X_t} \right]$ ,

$$\begin{aligned}
|\zeta(u, t)|^2 &= \mathbb{E} \left[ (\eta(X) - p) e^{iu^\top X_t} \right] \mathbb{E} \left[ (\eta(X') - p) e^{-iu^\top X'_t} \right] \\
&= \mathbb{E} \left[ (\eta(X) - p)(\eta(X') - p) e^{iu^\top (X_t - X'_t)} \right] \\
&= \mathbb{E} \left[ (\eta(X) - p)(\eta(X') - p) \cos(u^\top (X_t - X'_t)) \right] \\
&= -\mathbb{E} \left[ (\eta(X) - p)(\eta(X') - p)(1 - \cos(u^\top (X_t - X'_t))) \right],
\end{aligned}$$

where we have used  $|\zeta(u, t)|^2 \in \mathbb{R}$  and  $\mathbb{E}[(\eta(X) - p)(\eta(X') - p)] = 0$ . Now, using expression (3.3),

$$\begin{aligned}
\mathcal{V}^2(X_t, Y) &= -2\mathbb{E} \left[ (\eta(X) - p)(\eta(X') - p) \int_{\mathbb{R}^d} \frac{1 - \cos(u^\top (X_t - X'_t))}{c_d |u|_d^{d+1}} du \right] \\
&= -2\mathbb{E} \left[ (\eta(X) - p)(\eta(X') - p) |X_t - X'_t|_d \right] \\
&= -2\mathbb{E} \left[ (Y - p)(Y' - p) |X_t - X'_t|_d \right],
\end{aligned}$$

since [see e.g. Lemma 1 in Székely et al. (2007)],

$$\int_{\mathbb{R}^d} \frac{1 - \cos(u^\top x)}{c_d |u|_d^{d+1}} du = |x|_d, \quad \text{for all } x \in \mathbb{R}^d.$$

(c) By conditioning on  $Y$  and  $Y'$  we have

$$\begin{aligned} \mathbb{E}[(Y - p)(Y' - p)|X_t - X'_t|_d] &= p^2 I_{00}(t)(1 - p)^2 - p(1 - p)I_{01}(t)2p(1 - p) \\ &\quad + (1 - p)^2 I_{11}(t)p^2 = p^2(1 - p)^2(I_{00}(t) + I_{11}(t) - 2I_{01}(t)). \end{aligned}$$

Now, using (3.4),  $\mathcal{V}^2(X_t, Y) = 4p^2(1 - p)^2 \left[ I_{01}(t) - \frac{I_{00}(t) + I_{11}(t)}{2} \right]$ .  $\square$

*Proof of Theorem 3.2.* Continuity of  $\mathcal{V}_n^2(X_t, Y)$  is straightforward from DCT. It suffices to prove the result for sequences of samples  $X_1^{(0)}, \dots, X_{n_0}^{(0)}$ , and  $X_1^{(1)}, \dots, X_{n_1}^{(1)}$ , drawn from  $X|Y = 0$  and  $X|Y = 1$ , respectively, such that  $n_1/(n_0 + n_1) \rightarrow p = \mathbb{P}(Y = 1)$ .

From the triangle inequality it is enough to prove the uniform convergence of  $\hat{I}_{00}(t)$ ,  $\hat{I}_{11}(t)$  and  $\hat{I}_{01}(t)$  to  $I_{00}(t)$ ,  $I_{11}(t)$  and  $I_{01}(t)$ , respectively. For the first two quantities we apply Lemma 3.2 to the kernel  $k(x, x') = |x - x'|$ . For the last one we apply Lemma 3.3 to the same kernel. Observe that  $\mathbb{E}\|X\|_\infty < \infty$  implies the moment condition of Lemma 3.2 whereas  $\mathbb{E}(\|X\|_\infty \log^+ \|X\|_\infty) < \infty$  implies the moment condition of Lemma 3.3. The last statement readily follows from the uniform convergence and the compactness of  $[0, 1]^d$ .  $\square$

*Proof of Lemma 3.1.* Denote the expressions (3.2) and (3.6) by  $f_n$  and  $g_n$  respectively. Suppose there are  $n_c$  elements of class  $c$ ,  $c = 0, 1$ , we first order the elements of the sample grouping those of the same class. Hence, matrices  $(a_{ij})$  and  $(b_{ij})$  involved in  $f_n$  have this form,

$$(a_{ij}) = \left( \begin{array}{c|c} |X_i^{(0)} - X_j^{(0)}| & |X_i^{(0)} - X_j^{(1)}| \\ \hline |X_i^{(1)} - X_j^{(0)}| & |X_i^{(1)} - X_j^{(1)}| \end{array} \right), \quad (b_{ij}) = \left( \begin{array}{c|c} 0 & 1 \\ \hline 1 & 0 \end{array} \right),$$

where  $X^{(c)}$  represents an element of class  $c$ . Then, the matrices are divided in four homogeneous submatrices we denote by  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 0)$  and  $(1, 1)$ . Now, the computation of matrix  $B_{ij}$  is straight forward:

$$B_{ij}^{(0,0)} = -2\frac{n_1^2}{n^2}; B_{ij}^{(0,1)} = B_{ij}^{(1,0)} = 2\frac{n_0 n_1}{n^2}; B_{ij}^{(1,1)} = -2\frac{n_0^2}{n^2}.$$

Therefore,

$$\begin{aligned} f_n &= \frac{1}{n^2} \sum_{i,j=1}^n A_{ij} B_{ij} \\ &= \frac{1}{n^2} \left( +4 \frac{n_0 n_1}{n^2} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} A_{ij} - 2 \frac{n_1^2}{n^2} \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} A_{ij} - 2 \frac{n_0^2}{n^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} A_{ij} \right) \end{aligned}$$

Now, we operate term by term. Note that  $a_{i\cdot}^\top = a_{\cdot i}$ ,  $\sum_{i=1}^n \bar{a}_i = n\bar{a}$  and  $a_{ii} = 0$  for all  $1 \leq i \leq n$ . We also denote by  $a_0$  and  $a_1$  the matrices formed by the first  $n_0$  and the last  $n_1$  files of  $(a_{ij})$  respectively.

$$\begin{aligned} \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} A_{ij} &= \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} \left[ |X_i^{(0)} - X_j^{(0)}| - \bar{a}_i - \bar{a}_j + \bar{a} \right] \\ &= \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} |X_i^{(0)} - X_j^{(0)}| - 2 \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} \bar{a}_i + \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} \bar{a} \\ &= 2 \sum_{i < j}^{n_0} |X_i^{(0)} - X_j^{(0)}| - 2n_0^2 \bar{a}_0 + n_0^2 \bar{a} \\ &= n_0(n_0 - 1) \hat{I}_{0,0} - n_0^2(2\bar{a}_0 - \bar{a}). \end{aligned}$$

Analogously,

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} A_{ij} = n_1(n_1 - 1) \hat{I}_{1,1} - n_1^2(2\bar{a}_1 - \bar{a}),$$

and

$$\sum_{i=1}^{n_0} \sum_{j=1}^{n_1} A_{ij} = n_0 n_1 \hat{I}_{0,1} - n_0 n_1 (\bar{a}_0 + \bar{a}_1 - \bar{a}).$$

Finally, replacing in (3.2),

$$\begin{aligned} f_n &= \frac{1}{n^2} \sum_{i,j=1}^n A_{ij} B_{ij} \\ &= 4 \frac{n_0 n_1}{n^4} \left( n_0 n_1 \hat{I}_{0,1} - \frac{n_1(n_0 - 1) \hat{I}_{0,0} + n_0(n_1 - 1) \hat{I}_{1,1}}{2} \right). \end{aligned} \quad (3.14)$$

Now, it is readily seen that

$$\begin{aligned} \|f_n - g_n\|_\infty &\leq 2 \frac{n_0 n_1^2}{n^4} \left( \|\hat{I}_{0,0} - I_{0,0}\|_\infty + \|\hat{I}_{0,0}\|_\infty \right) \\ &\quad + 2 \frac{n_0^2 n_1}{n^4} \left( \|\hat{I}_{1,1} - I_{1,1}\|_\infty + \|\hat{I}_{1,1}\|_\infty \right) \xrightarrow{n \rightarrow \infty} 0 \quad a.s. \end{aligned}$$

Then, the result follows from this and Theorem 3.2.  $\square$

*Proof of Proposition 3.1.* We know  $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$ . Then, we use equation (3.7), which provides  $\eta(x)$  in terms of the Radon-Nikodym derivative  $dP_0/dP_1$ , and the expression for  $dP_0/dP_1$  given in Liptser and Shiryaev (2013, p. 239). This gives

$$\eta(x) = \left[ \frac{1-p}{p} \sqrt{2} e^{-x_1^2/4} + 1 \right]^{-1}.$$

Now, from  $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$ , we get  $g^*(x) = 1 \Leftrightarrow x_1^2 > 4 \log \left( \frac{\sqrt{2}(1-p)}{p} \right)$ .  $\square$

*Proof of Proposition 3.2.* Again, we use expression (3.7) to derive the expression of the optimal rule  $g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}$ . In this case the calculation is made possible using the expression of the Radon-Nikodym derivative for the distribution of a Brownian process with trend,  $F(t) + B(t)$ , with respect to that of a standard Brownian:

$$\frac{dP_1}{dP_0}(B) = \exp \left\{ -\frac{1}{2} \int_0^1 F'(s)^2 ds + \int_0^1 F' dB \right\}, \quad (3.15)$$

for  $P_0$ -almost all  $B \in \mathcal{C}[0, 1]$ ; see, Mörters and Peres (2010, Thm. 1.38 and Remark 1.43), for further details. Observe that in this case we have  $F(t) = ct$ . Thus, from (3.7), we finally get  $\eta(x) = \left[ \frac{1-p}{p} \exp \left( \frac{c^2}{2} - cx_1 \right) + 1 \right]^{-1}$ , which again only depends on  $x$  through  $x(1) = x_1$ . The result follows easily from this expression.  $\square$

*Proof of Proposition 3.3.* In this case, the trend function is  $F(t) = \Phi_{m,k}(t)$ . So  $F'(t) = \varphi_{m,k}$  and  $F''(t) = 0$ . From equations (3.7) and (3.15), we readily get (3.9) and (3.10).  $\square$

*Proof of Theorem 3.3.* We will use the following result

**Theorem 3.4** (Shepp 1966, Thm. 1). *Let  $P_0, P_1$  be the distributions corresponding to the standard Brownian Motion  $\{B(t), t \in [0, T]\}$  and to a Gaussian process  $\{X(t), t \in [0, T]\}$  with mean function  $m_1$  in the Dirichlet space  $\mathcal{D}[0, T]$  and covariance function  $K$ . Then  $P_1 \sim P_0$  if and only if there exists a function  $K_1 \in L^2([0, T] \times [0, T])$  such that*

$$K(s, t) = \min\{s, t\} - \int_0^s \int_0^t K^{(1)}(u, v) dudv, \quad (3.16)$$

with  $1 \notin \sigma(K^{(1)})$ , the spectrum of  $K^{(1)}$ . In this case, the function  $K^{(1)}$  is given by  $K^{(1)}(s, t) = -\frac{\partial^2}{\partial s \partial t} K(s, t)$ .

We will also need Lemmas 1 and 2 in Shepp (1966), p. 334-335 which give the expression of the Radon-Nikodym derivative  $dP_1/dP_0$  in the case  $P_1 \ll P_0$  under the conditions of Theorem 3.4.

Now, to prove (a) Let  $\lambda_1, \lambda_2, \dots$  and  $\varphi_1, \varphi_2, \dots$  be the eigenvalues and the corresponding unit (with respect to the  $L^2$  norm) eigenfunctions of the integral operator defined by the kernel  $K^{(1)}$  in (3.16). Define  $X_j = \int_0^T \varphi_j(t) dX(t)$  and  $\xi_j = \int_0^T m'_1(t) \varphi_j(t) dt$ , where  $m'_1$  is defined in Equation (2.2). According to the above mentioned Lemmas 1 and 2 in Shepp (1966), we have

$$\frac{dP_1}{dP_0}(X) = \exp \left\{ -\frac{1}{2} \sum_{j=1}^{\infty} \left[ \log(1 - \lambda_j) + \frac{\lambda_j X_j^2 + \xi_j^2 - 2\xi_j X_j}{1 - \lambda_j} \right] \right\}. \quad (3.17)$$

For the Brownian Bridge, we have  $K(s, t) = \min\{s, t\} - st$  and, therefore,  $K^{(1)} \equiv 1$ . It is not difficult to show that in this case  $\lambda = T$  is the only non-zero eigenvalue, and  $\varphi(t) \equiv 1/\sqrt{T}$  is its corresponding unit eigenfunction. From Theorem 3.4,  $P_0 \sim P_1$  if and only if  $T < 1$ . Moreover, if  $m_1(0) = 0$  we have  $\xi = \int_0^T m'_1(t) dt / \sqrt{T} = m_1(T) / \sqrt{T}$ . Then, from (3.17),

$$\frac{dP_1}{dP_0}(X) = \exp \left\{ -\frac{1}{2} \left[ \log(1 - T) + \frac{TX(T)^2 + m_1(T)^2 - 2m_1(T)X(T)}{T(1 - T)} \right] \right\}. \quad (3.18)$$

Equation (3.11) follows from this expression and (1.2).

(b) Let  $P_B$  the probability measure corresponding to Brownian Motion. Particularizing (3.18) for the mean functions  $m_0$  and  $m_1$  we get expressions for  $dP_0/dP_B$  and  $dP_1/dP_B$ . Using the chain rule we have

$$\frac{dP_1}{dP_0}(X) = \frac{dP_1/dP_B(X)}{dP_0/dP_B(X)} = \exp \left\{ \frac{m_0(T)^2 - m_1(T)^2 - 2X(T)(m_0(T) - m_1(T))}{2T(1 - T)} \right\}.$$

The result follows from this expression and (1.2).  $\square$

*Proof of Proposition 3.4.* Let us first consider the model in Proposition 3.1 (i.e., Brownian vs. Brownian with a stochastic trend). Such model entails that  $X_t|Y = 0 \sim N(0, \sqrt{t})$  and  $X_t|Y = 1 \sim N(0, \sqrt{t^2 + t})$ . Now, recall that if  $\xi \sim N(m, \sigma)$ , then,

$$\mathbb{E}|\xi| = \sigma \sqrt{\frac{2}{\pi}} e^{-\frac{m^2}{\sigma^2}} + m \left( 2\Phi\left(\frac{m}{\sigma}\right) - 1 \right), \quad (3.19)$$

where  $\Phi(z)$  denotes the distribution function of the standard normal.

Now, using (3.5) and (3.19) we have the following expressions,

$$I_{01}(t) = \mathbb{E}|\sqrt{t}Z - \sqrt{t^2 + t}Z'| = \sqrt{\frac{2(t^2 + 2t)}{\pi}},$$

$$I_{00}(t) = \mathbb{E}|\sqrt{t}Z - \sqrt{t}Z'| = \sqrt{\frac{4t}{\pi}},$$

$$I_{11}(t) = \mathbb{E}|\sqrt{t^2 + t}Z - \sqrt{t^2 + t}Z'| = \sqrt{\frac{4(t^2 + t)}{\pi}},$$

where  $Z$  and  $Z'$  are independent  $N(0, 1)$  random variables.

Then, the function  $\mathcal{V}^2(X_t, Y) = 4p^2(1-p)^2 \left( I_{01}(t) - \frac{I_{00}(t) + I_{11}(t)}{2} \right)$  grows with  $t$  so it is maximized at  $t^* = 1$ , which is the only point that has an influence on the Bayes rule.

Let us now consider the model in Proposition 3.2 (i.e., Brownian vs. Brownian with a linear trend). Again, from (3.19) we have in this case,

$$I_{01}(t) = \mathbb{E}|ct + \sqrt{t}Z - \sqrt{t}Z'| = 2\sqrt{\frac{t}{\pi}}e^{-\frac{c^2t}{2}} + ct \left( 2\Phi \left( c\sqrt{\frac{t}{2}} \right) - 1 \right),$$

$$I_{00}(t) = I_{11}(t) = \mathbb{E}|\sqrt{t}Z - \sqrt{t}Z'| = \sqrt{\frac{4t}{\pi}},$$

where  $Z$  and  $Z'$  are iid standard Gaussian variables. Therefore using (3.5),

$$\mathcal{V}^2(X_t, Y) = C \left[ 2\sqrt{\frac{t}{\pi}} \left( e^{-\frac{c^2t}{2}} - 1 \right) + ct \left( 2\Phi \left( c\sqrt{\frac{t}{2}} \right) - 1 \right) \right],$$

where  $C = 4p^2(1-p)^2$ . We can check numerically that this an increasing function which reaches its only maximum at  $t^* = 1$ . According to Proposition 3.2 this is the only relevant point for the Bayes rule.  $\square$

*Proof of Proposition 3.5.* Using the notation in Theorem 3.1, for all  $t \in [0, 1]$ ,

$$\zeta(u, t) = \mathbb{E}[(\eta(X) - p)e^{iuX_t}] = \mathbb{E}[(h(X_{t^*}) - p)e^{iuX_{t^*}} e^{iu(X_t - X_{t^*})}].$$

When  $t > t^*$ , the third factor within the expectation above is independent of the other two factors. Hence,

$$\zeta(u, t) = \mathbb{E}[(h(X_{t^*}) - p)e^{iuX_{t^*}}] \cdot \mathbb{E}[e^{iu(X_t - X_{t^*})}] = \zeta(u, t^*) \cdot \varphi_{X_t - X_{t^*}}(u).$$

As a consequence, for all  $t > t^*$ ,  $|\zeta(u, t)|^2 = |\zeta(u, t^*)|^2 \cdot |\varphi_{X_t - X_{t^*}}(u)|^2 \leq |\zeta(u, t^*)|^2$ . Observe that there exists some  $u$  for which strict inequality holds (since by assumption, the increments are non-degenerate).

Finally, by the last inequality and Theorem 3.1 (a), for all  $t > t^*$ ,

$$\mathcal{V}^2(X_t, Y) = \frac{2}{c_1} \int_{\mathbb{R}} \frac{|\zeta(u, t)|^2}{u^2} du < \frac{2}{c_1} \int_{\mathbb{R}} \frac{|\zeta(u, t^*)|^2}{u^2} du = \mathcal{V}^2(X_{t^*}, Y).$$

□

# Chapter 4

## mRMR

In this chapter we study the application to the functional case of a well-known multivariate variable selection method, and we propose some modifications in order to achieve better results in the new setup.

As mentioned above, functional data are discretized in practice so, in principle, one might think that any multivariate dimension reduction method is potentially applicable to these “vectorial”, discretized functional data. We have seen however, that this is not the case with many popular techniques (see e.g. ranking methods which can be easily adapted but with poor results). In other cases, the FDA adaptations of multivariate techniques have been much more successful: two clear examples are PLS (Preda et al., 2007) and PCA (Ramsay and Silverman, 2005) methodologies. To our knowledge, intrinsic variable selection methods have not been incorporated yet to the FDA literature even though they are very popular, especially in the machine learning literature. In this chapter we explore the adaptation to the FDA setup of the so-called *minimum Redundancy Maximum Relevance* (mRMR) method by Ding and Peng (2005).

Overall, we believe the mRMR procedure is a very natural way to tackle the variable selection problem if one wants to make completely explicit the trade-off relevance/redundancy. The method relies on the use of an association measure to assess the relevance and redundancy of the considered variables. In the original papers (Ding and Peng, 2005; Peng et al., 2005) the so-called ‘mutual information’ measure was used for this purpose. The aim of the present work is to propose other alternatives for the association measure, still keeping the main idea behind the mRMR procedure. In fact, most mRMR researchers admit that there is considerable room for improvement. We quote from the discussion in Peng et al. (2005): *‘The mRMR paradigm can be better viewed as a general framework to effectively*

*select features and allow all possibilities for more sophisticated or more powerful implementation schemes*'. In this vein, we consider several versions of the mRMR and compare them by an extensive empirical study. Two of these versions are new: they are based on the 'distance covariance' and 'distance correlation' association measures proposed by Székely et al. (2007). Our results suggest that mRMR is a suitable variable selection algorithm for functional data and that the new version based on the distance correlation measure represents a clear improvement of the mRMR methodology.

The mRMR method is presented in Section 4.1: the considered association measures and the variable selection algorithm are described in Subsections 4.1.1 and 4.1.2 respectively. The different versions of mRMR are tested in Section 4.2. The empirical study includes an extensive simulation study (Subsection 4.2.1), together with different rankings of the considered methods (Subsection 4.2.2) and three real data sets (Subsection 4.2.3). In Section 4.3 we study the application of the new proposals to a real problem of spectral classification and metabolite detection. Finally, some conclusions are given.

## 4.1. The trade-off relevance/redundancy. The mRMR criterion

When faced with the problem of variable selection methods in high-dimensional (or functional) data sets, a natural idea arises at once: obviously, one should select the variables according to their relevance (representativeness). However, at the same time, one should avoid the redundancy which appears when two highly relevant variables are closely related. In that case, one might expect that both variables essentially carry the same information, so that choosing just one of them should suffice.

The mRMR variable selection method, as proposed in Ding and Peng (2005); Peng et al. (2005), provides a formal implementation of a variable selection procedure which explicitly takes into account this trade-off relevance/redundancy. It is extremely popular and, in fact, it has motivated thousands of citations in the machine learning community.

### 4.1.1. Association measures

As we will see in next Subsection, the mRMR criterion relies on the use of an association measure  $I(X, Y)$  between random variables. The choice of the

association measure  $I$  is a critical aspect in the mRMR methodology. In fact, this is the central point of the present work. Furthermore, the choice of appropriate association measures is a classical issue in mathematical statistics. Many different proposals are available and, in several aspects, this topic is still open for further research, especially in connection with the use of high-dimensional data sets (arising, e.g., in genetic microarray examples, [Reshef et al. \(2011\)](#); [Hall and Miller \(2011\)](#)).

A complete review of the main association measures for random variables is clearly beyond the scope of this paper. So, we will limit ourselves to present here the measures  $I(X, Y)$  we have used in this work:

**The ordinary correlation coefficient** between  $X$  and  $Y$  (in absolute value). This is the first obvious choice for the association measure  $I(X, Y)$ . It clearly presents some drawbacks (it does not characterize independence and it is unsuitable to capture non-linear association) but still, it does a good job in many practical situations.

**The Mutual Information Measure,**  $MI(X, Y)$  is defined by

$$MI(X, Y) = \int \log \frac{p(x, y)}{p_1(x)p_2(y)} p(x, y) d\mu(x, y), \quad (4.1)$$

where  $X, Y$  are two random variables with respective  $\mu$ -densities  $p_1$  and  $p_2$ ; in the standard, absolutely continuous case,  $\mu$  would be the product Lebesgue measure. In the discrete case,  $\mu$  would be a counting measure on a countable support. The joint density of  $(X, Y)$  is denoted by  $p(x, y)$ .

This is the association measure used in the original version of the mRMR procedure ([Ding and Peng, 2005](#); [Peng et al., 2005](#)). In fact, the opportunities MI offers for variable selection have been widely exploited resulting in a field within machine learning, the so-called information theoretic feature selection. A comprehensive review of intrinsic methods based on MI and some considerations about the measure are given in [Vergara and Estévez \(2014\)](#). Likewise, the interesting paper by [Brown et al. \(2012\)](#) provides a theoretical framework for information theoretic feature selection in terms of an optimization on the conditional likelihood (instead of the usual heuristic approaches). In this framework, some popular variable selection algorithms (including mRMR) can be seen as approximations of a general paradigm.

It is clear that  $MI(X, Y)$  measures how far is  $p(x, y)$  from the independence situation  $p(x, y) = p_1(x)p_2(y)$ . It is easily seen that  $MI(X, Y) = MI(Y, X)$  and  $MI(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent. Some other favourable properties of this measure for variable selection are described, e.g. in [Frénay et al. \(2013\)](#), including theoretical bounds that relates the Bayes error with the entropy.

In practice,  $MI(X, Y)$  must be approximated by considering, if necessary, ‘discretized versions’ of  $X$  and  $Y$ , obtained by grouping their values on intervals represented by suitable label marks,  $a_i, b_j$ . This leads to approximate expressions of type

$$\widehat{MI}(X, Y) = \sum_{i,j} \log \frac{\mathbb{P}(X = a_i, Y = b_j)}{\mathbb{P}(X = a_i)\mathbb{P}(Y = b_j)} \mathbb{P}(X = a_i, Y = b_j), \quad (4.2)$$

where, in turn, the probabilities can be empirically estimated by the corresponding relative frequencies. In [Ding and Peng \(2005\)](#) the authors suggest a threefold discretization pattern, i.e., the range of values of the variable is discretized in three classes. The limits of the discretization intervals are defined by the mean of the corresponding variable  $\pm\sigma/2$  (where  $\sigma$  is the standard deviation). We will explore this criterion in our empirical study below.

However, the estimation of MI for continuous variables is the main drawback of this measure. This is pointed out in several papers (see e.g. [Walters-Williams and Li \(2009\)](#); [Frénay et al. \(2013\)](#); [Vergara and Estévez \(2014\)](#)) which comment on the need of new approaches. In fact, [Seth and Principe \(2010\)](#) conclude that it is impossible to get a good MI estimator for small sample sizes and continuous variables. There are two typical strategies to face this problem. The first one (used for example in [Battiti \(1994\)](#), [Ding and Peng \(2005\)](#) and [Peng et al. \(2005\)](#)) is the estimation via histograms; this is a simple and reasonably effective method. The second alternative is the kernel based estimation (sometimes called “Parzen windows” in the literature) considered, for instance in [Peng et al. \(2005\)](#) and [Estévez et al. \(2009\)](#). This methodology can lead to better results but it suffers from the typical drawbacks of nonparametric procedures (choice of the smoothing parameter, need of large sample sizes). Many other methodologies have been proposed in order to overcome the estimation problem; see [Walters-Williams and Li \(2009\)](#) for a survey and some additional references. However, none of them have been widely accepted to replace the first two approaches mentioned above.

**The Fisher-Correlation ( $FC$ ) criterion:** It is a combination of the  $F$ -statistic,

$$F(X, Y) = \frac{\sum_k n_k (\bar{X}_k - \bar{X})^2 / (K - 1)}{\sum_k (n_k - 1) \sigma_k^2 / (n - K)}, \quad (4.3)$$

used as the relevance measure (4.4), and the ordinary correlation,  $C$ , used as the redundancy measure (4.5). In the expression (4.3),  $K$  denotes the number of classes (so  $K = 2$  in our binary classification problem),  $\bar{X}$  denotes the mean of  $X$ ,  $\bar{X}_k$  is the mean value of  $X$  of the elements belonging the  $k$ -th class, for  $k = 0, 1$ , and  $n_k$  and  $\sigma_k^2$  are the sample size and the variance of the  $k$ -th class, respectively.

Ding and Peng (2005) suggest that, in principle, this criterion might look more useful than  $\widehat{MI}$  when dealing with continuous variables but their empirical results do not support that idea. Such results are confirmed by our study so that, in general terms, we conclude that the mutual information (4.2) is a better choice even in the continuous setting.

**Distance covariance:** this association measure recently proposed by Székely et al. (2007) is largely described in Section 3.1. Let us still recall here that while definition (3.1) has a rather technical appearance, the resulting association measure has a number of interesting properties. Apart from the fact that (3.1) allows for the case where  $X$  and  $Y$  have different dimensions, we have  $\mathcal{V}^2(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent. Moreover, the indicated choice for the weights  $w(u, v)$  provides valuable equivariance properties for  $\mathcal{V}^2(X, Y)$  and the quantity can be consistently estimated (and no discretization is needed) from the mutual pairwise distances  $|X_i - X_j|_p$  and  $|Y_i - Y_j|_q$  between the sample values  $X_i$  and  $Y_j$ .

**Distance correlation:** this is just a sort of standardized version of the distance covariance. If we denote  $\mathcal{V}^2(X) = \mathcal{V}^2(X, X)$ , the (square) distance correlation between  $X$  and  $Y$  is defined by  $\mathcal{R}^2(X, Y) = \frac{\mathcal{V}^2(X, Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}$  if  $\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0$ ,  $\mathcal{R}^2(X, Y) = 0$  otherwise.

In fact, distance correlation fulfils most of the desirable properties of MI according to Frénay et al. (2013) and Vergara and Estévez (2014). It also (partially) satisfies and the postulates of Rényi (1959) for a suitable dependence measure. As a conclusion, we might say that  $\mathcal{R}^2$  is a more suitable choice than MI to be used in the mRMR procedure.

Of course, other association measures might be considered. However, in order to get an affordable comparative study, we have limited our study to the main association measures previously used in the mRMR literature. We have only added the new measures  $\mathcal{V}^2$  and  $\mathcal{R}^2$ , which we have tested as possible improvements of the method.

Let us finally note that all the association measures we are considering take positive values. So, the phenomena associated with the the negative association values analyzed in Demler et al. (2013) do not apply in this case.

### 4.1.2. Methodology

The mRMR method was proposed by Ding and Peng (2005) and Peng et al. (2005) as a tool to select the most discriminant subset of variables in the context of some relevant bioinformatics problems. Its good performance is assessed in many research works, specially in genetic problems; see for example Brown et al. (2012) for an extensive comparative study. In our functional binary classification problem, the description of the mRMR method is as follows: the functional explanatory variable  $X(t)$ ,  $t \in [0, 1]$  will be used in a discretized version  $(X(t_1), \dots, X(t_N))$ . When convenient, the notations  $X_t$  and  $X(t)$  will be used indistinctly. For any subset  $S$  of  $\mathbb{I} = \{t_1, \dots, t_N\}$ , the *relevance* and the *redundancy* of  $S$  are defined, respectively, by

$$\text{Rel}(S) = \frac{1}{\text{card}(S)} \sum_{t \in S} I(X_t, Y), \quad (4.4)$$

and

$$\text{Red}(S) = \frac{1}{\text{card}^2(S)} \sum_{s, t \in S} I(X_t, X_s), \quad (4.5)$$

where  $\text{card}(S)$  denotes the cardinality of  $S$  and  $I(\cdot, \cdot)$  is an ‘association measure’. The function  $I$  measures how much related are two variables. So, it is natural to think that the relevance of  $X_t$  is measured by how much related it is with the response variable  $Y$ , that is  $I(X_t, Y)$ , whereas the redundancy between  $X_t$  and  $X_s$  is given by  $I(X_s, X_t)$ . Now, in summary, the mRMR algorithm aims at maximizing the relevance avoiding an excess of redundancy. The use of a methodology of this type is especially important in the functional data problems, where those variables which are very close together are often strongly associated.

Now, in order to explain how the mRMR method works, let us assume that the measure  $I$  is given:

- (a) The procedure starts by selecting the most relevant variable, given by the value  $t_i$  such that the set  $S_i = \{t_i\}$  maximizes  $\text{Rel}(S)$  among all the singleton sets of type  $S_j = \{t_j\}$ .

- (b) Then, the variables are sequentially incorporated to the set  $S$  of previously selected variables, with the criterion of maximizing the difference  $\text{Rel}(S) - \text{Red}(S)$  (or alternatively the quotient  $\text{Rel}(S)/\text{Red}(S)$ ).
- (c) Finally, different stopping rules can be considered. We set the number of variables through a validation step (additional details can be found in next Section).

When using MI as an association measure, Peng et al. (2005) showed that the mRMR is equivalent to *Max-Dependency* (an exhaustive variable selection algorithm) for the first order incremental search.

#### *A comment on the mRMR literature*

The basic idea behind the mRMR method can be found in an early paper by Battiti (1994) who proposes a MI-based greedy algorithm (called MIFS) quite similar to mRMR. The main difference between the two methods is that in MIFS the relative influence of the relevance and the redundancy is addressed by a weighted average with a tuning parameter while in mRMR it is regulated through the cardinal of the subset (which varies during the execution). The mRMR balance of the relevancy and redundancy terms is extremely important and this is why mRMR outperforms MIFS in almost all experiments (Brown et al., 2012). A modification of MIFS with a kernel estimation of MI was proposed by Kwak and Choi (2002). However, Estévez et al. (2009) obtained better results with the original MIFS than with this newer version (perhaps because of the difficulties that entail the choice of the smoothing parameter). Closely related ideas tackling an explicit treatment of the relevance-redundancy trade off along with some theoretical background were also considered in Yu and Liu (2004).

Since the first paper by Ding and Peng (2005), many alternative versions of the mRMR procedure have been proposed in the literature. For instance, other weighting factors might be used instead of just  $\text{card}(S)$  in equation (4.5). In this line, Ponsa and López (2007) and Estévez et al. (2009) propose two different normalizations. Another source of variability is the association measure, either replacing it by a new one (as in this work) or changing the estimation of the MI. We have seen some examples in previous subsection, and there is a general agreement on the difficulty of estimating MI for continuous variables (Frénay et al., 2013; Vergara and Estévez, 2014). For instance, in the common case of kernel density estimation (Wand and Jones, 1994) the crucial issue of the optimal selection of the smoothing parameter (Cao et al., 1994) has not been, to our knowledge, explicitly addressed in this setup. Note that here ‘optimal’ should refer to the estimation of MI. Hence, following the suggestions of Vergara and Estévez (2014) and Seth and

Principe (2010) among others, it seems worthy to look for another suitable association measure keeping the advantageous properties of MI but with good enough estimators; distance correlation appears to be a good candidate. Also, mRMR is frequently used in two-stages algorithms where plays the role of both a first filter method (in Mundra and Rajapakse (2010); El Akadi et al. (2011) mRMR is used before SVM-CFE and a genetic algorithm respectively) and a second stage to remove redundancy (Zhang et al. (2008) applies ReliefF before mRMR). However, still the ‘original’ version of mRMR (with discretization-based MI estimation) seems to be the most popular standard; see e.g. Gao et al. (2013); Nguyen et al. (2014); Mandal and Mukhopadhyay (2015) for very recent examples.

## 4.2. The empirical study

We have checked five different versions of the mRMR variable selection methodology. They have been obtained by using different association measures (as indicated in the previous section) to assess relevance and redundancy. The association measures defined above, i.e, standard correlation (in absolute value), mutual information, Fisher-correlation criterion, distance covariance and distance correlation, will be denoted in the tables of our empirical study by **C**, **MI**, **FC**, **V** and **R**, respectively.

In all cases, the comparisons have been made in the context of problems of binary supervised classification, using 100 different models to generate the data  $(X, Y)$ . These models are defined in Subsection 5.2.1 and listed in Appendix A. All these models have been chosen in such a way that the optimal (Bayes) classification rule depends on just a finite number of variables. The processes considered include Brownian motion (with different mean functions), Brownian bridge and several other Gaussian models, in particular the Ornstein-Uhlenbeck process. Other mixture models based on them are also considered.

Our experiments essentially consist of performing variable selection for each model using the different versions of mRMR and evaluating the results in terms of the respective probabilities of correct classification when different classifiers are used on the selected variables.

For each considered model all the variable selection methods ( **C**, **MI**, etc.) are checked for four sample sizes,  $n = 30, 50, 100, 200$  and four classification methods ( $k$ -NN, **LDA**, **NB** and **SVM**). So, we have in total  $100 \times 4 \times 4 = 1600$  simulation experiments. All the functional simulated data are discretized to  $(x(t_1), \dots, x(t_{100}))$ , where  $t_i$  are equi-spaced points in  $[0, 1]$ .

We have used the four classifiers considered in the paper by [Ding and Peng \(2005\)](#), except that we have replaced the logistic regression classifier (which is closely related to the standard linear classifier) with the non-parametric  $k$ -NN method with the usual Euclidean distance. The other considered classification rules are the linear discriminant analysis (LDA), Naïve Bayes classifier (NB) and a support vector machine with linear kernel (SVM). All of them are widely known and details can be found, e.g. in [Hastie et al. \(2009\)](#).

As an objective reference, our simulation outputs include also the percentages of correct classification obtained with those classifiers based on the complete curves, i.e., when no variable selection is done at all (except for LDA whose functional version is not feasible; see Section 1.2. This reference method is called **Base**. A somewhat surprising conclusion of our study is that this Base method is often outperformed by the variable selection procedures. This could be due to the fact that the whole curves are globally more affected by noise than the selected variables. Thus, variable selection is beneficial not only in terms of simplicity but also in terms of accuracy.

The number  $k$  of nearest neighbours in the  $k$ -NN classifier, the cost parameter  $C$  of the linear SVM and the number of selected variables are chosen by standard validation procedures ([Guyon et al., 2006](#)). To this end, in the simulation study, we have generated independent validation and test samples of size 200. Each simulation output is based on 200 independent runs.

Let us finally recall that further details on the methodology, implementation, methods, etc. are given in Chapter 5.

### 4.2.1. A few numerical outputs from the simulations

We present here just a small sample of the entire simulation outputs, which can be downloaded from [www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx](http://www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx).

Tables 4.1 - 4.4 contain the results obtained with NB,  $k$ -NN, LDA and SVM respectively. The boxed outputs in these tables correspond to the winner and second best method in each row. The columns headings (MID, FCD, etc.) correspond to the different mRMR methods based on different association measures, as defined in Subsection 4.1.1. The added letter 'D' refers to the fact that global criterion to be maximized is just the difference between the measures (4.4) and (4.5) of relevance and redundancy, respectively. There are also other possibilities to combine

Table 4.1: Performance outputs for the considered methods, using NB and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (NB)	Sample size	MID	FCD	RD	VD	CD	Base
Average accuracy	$n = 30$	78.08	78.42	79.56	79.24	79.28	77.28
	$n = 50$	79.64	79.34	80.92	80.45	80.46	78.29
	$n = 100$	80.76	80.06	81.90	81.34	81.41	78.84
	$n = 200$	81.46	80.44	82.55	81.90	82.05	79.13
Average dim. red	$n = 30$	8.7	9.3	7.2	7.1	7.8	100
	$n = 50$	7.9	9.0	6.8	6.7	7.4	100
	$n = 100$	7.2	8.5	6.3	6.2	6.8	100
	$n = 200$	6.6	8.1	5.8	5.7	6.4	100
Victories over Base	$n = 30$	57	61	77	71	69	-
	$n = 50$	66	61	79	74	70	-
	$n = 100$	77	61	88	81	85	-
	$n = 200$	84	62	93	85	91	-

relevance and redundancy indices. One could take for instance the quotient and the corresponding outputs methods are denoted MIQ, FCQ, etc. in the online Excel file. However, these outputs are not given here for the sake of brevity. In any case, our results suggest that the difference-based methods are globally (although not uniformly) better than those based on quotients. The column ‘Base’ gives the results when no variable selection method is used (that is, the entire curves are considered). This column does not appear when the LDA classifier is used, since LDA cannot directly work on functional data.

The row entries ‘Average accuracy’ provide the average percentage of correct classification over the 100 considered model outputs; recall that every output is in turn obtained as an average over 200 independent runs. The rows ‘Average dim. red.’ provide the average numbers of selected variables. The number of models wherein every method beats the ‘Base’ benchmark procedure is given in ‘Victories over Base’.

It can be seen from these results that the global winner is the R-based mRMR method, with a especially good performance for small sample sizes. Note that the number of variables required by this method is also smaller, in general, than that of the remaining methods. Moreover, RD is the most frequent winner with respect to the Base method (with all classifiers) keeping, in addition, a more stable general performance when compared with the other variable selection methods. In

Table 4.2: Performance outputs for the considered methods, using  $k$ -NN and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output ( $k$ -NN)	Sample size	MID	FCD	RD	VD	CD	Base
Average accuracy	$n = 30$	80.09	79.26	81.30	80.54	80.40	78.98
	$n = 50$	81.43	79.91	82.44	81.47	81.33	80.34
	$n = 100$	83.01	80.76	83.82	82.54	82.32	81.99
	$n = 200$	84.28	81.34	84.89	83.37	83.15	83.38
Average dim. red	$n = 30$	9.2	9.8	7.7	8.3	8.0	100
	$n = 50$	9.3	9.9	7.9	8.5	8.1	100
	$n = 100$	9.6	10.2	8.2	8.7	8.3	100
	$n = 200$	9.8	10.4	8.5	8.8	8.7	100
Victories over Base	$n = 30$	71	51	83	72	69	-
	$n = 50$	71	45	81	70	68	-
	$n = 100$	71	38	78	60	65	-
	$n = 200$	73	33	82	56	58	-

Table 4.3: Performance outputs for the considered methods, using LDA and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (LDA)	Sample size	MID	FCD	RD	VD	CD	Base
Average accuracy	$n = 30$	78.72	76.87	79.35	78.23	78.37	-
	$n = 50$	80.28	77.84	80.59	79.15	79.36	-
	$n = 100$	81.85	78.97	81.88	80.22	80.47	-
	$n = 200$	82.96	79.83	82.87	81.02	81.30	-
Average dim. red	$n = 30$	5.6	4.9	5.0	4.6	5.2	-
	$n = 50$	6.5	5.9	5.9	5.5	6.1	-
	$n = 100$	7.9	7.5	7.1	6.8	7.4	-
	$n = 200$	9.0	8.9	8.0	8.0	8.3	-

Table 4.4: Performance outputs for the considered methods, using SVM and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (SVM)	Sample size	MID	FCD	RD	VD	CD	Base
Average accuracy	$n = 30$	81.53	79.41	81.50	80.35	80.51	81.91
	$n = 50$	82.61	80.01	82.45	81.00	81.20	82.99
	$n = 100$	83.75	80.75	83.45	81.77	82.00	84.11
	$n = 200$	84.55	81.27	84.22	82.38	82.61	84.91
Average dim. red	$n = 30$	10.5	11.0	9.2	9.7	9.4	100
	$n = 50$	10.5	11.1	9.3	9.7	9.6	100
	$n = 100$	10.7	11.3	9.6	10.0	9.9	100
	$n = 200$	10.9	11.5	9.7	10.1	9.9	100
Victories over Base	$n = 30$	37	39	49	43	42	-
	$n = 50$	42	34	56	44	46	-
	$n = 100$	49	32	57	41	47	-
	$n = 200$	48	29	59	42	49	-

this sense, R-based methods seem both efficient and reliable. While RD preforms well with all classifiers, MID results are clearly favoured by linear classification rules. In agreement with the results in [Ding and Peng \(2005\)](#), the performance of the FC-based method is relatively poor. Finally, note that the Base option (which uses the entire curves) is never the winner, with the partial exception of the SVM classifier.

### 4.2.2. Ranking the methods

It is not easy to draw general conclusions, and clear recommendations for practitioners, from a large simulation study. A natural idea is to give some kind of quantitative assessment summarizing the relative merits of the different procedures. Many different ranking criteria might be considered. As in the previous Chapter we have considered here the following ones:

- **Relative ranking:** for each considered model and sample size the winner method (in terms of classification accuracy) gets 10 score points and the method with the worst performance gets 0 points. The score of any other method, with performance  $u$ , is defined by  $10(u - w)/(W - w)$ , where  $W$  and  $w$  denote, respectively, the performances of the best and the worst method.
- **Positional ranking:** The winner gets 10 points, the second best gets 9, etc.

Table 4.5: Global scores of the considered methods under three different ranking criteria using NB. Each output is the average of 100 models

Ranking criterion (NB)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	2.43	5.10	8.67	7.08	8.10
	$n = 50$	3.04	4.31	9.16	6.97	7.86
	$n = 100$	3.38	3.92	9.28	6.84	7.82
	$n = 200$	3.84	3.57	9.20	6.56	7.59
Positional	$n = 30$	6.65	7.62	8.84	8.21	8.68
	$n = 50$	6.82	7.43	9.12	8.19	8.46
	$n = 100$	6.87	7.36	9.26	8.16	8.35
	$n = 200$	6.96	7.30	9.18	8.17	8.42
F1	$n = 30$	11.64	15.11	18.64	16.37	18.24
	$n = 50$	12.13	14.54	20.24	16.16	16.98
	$n = 100$	12.19	14.29	20.82	16.17	16.53
	$n = 200$	12.38	14.09	20.54	16.15	16.92

- F1 ranking:** the scores are assigned according to the current criteria in a Formula 1 Grand Prix: the winner gets 25 score points and the following ones get 18, 15, 10, 8, 6, and 4 points.

The summary results are shown in Tables 4.5 - 4.8 and a visual version of the complete (400 experiments) relative ranking outputs for the four classifiers are displayed in Figure 4.1 The conclusions are self-explanatory and quite robust with respect to the ranking criterion. The mRMR methods based on the distance correlation measure are the uniform global winners. The results confirm the relative stability of R, especially when compared with MI whose good performance is restricted to a few models. The problems estimating MI with smaller sample sizes can be also observed.

Of course, the criteria for defining these rankings, as well as the idea of averaging over different models, are questionable (although one might think of a sort of Bayesian interpretation for these averages). Anyway, this is the only way we have found to provide an understandable summary for such a large empirical study. On the other hand, since we have made available the whole outputs of our experiments, other different criteria might be used by interested readers.

### 4.2.3. Real data examples

We have chosen again three real-data examples on the basis of their popularity in the literature on Functional Data Analysis: we call them *Growth* (93 growth

Table 4.6: Global scores of the considered methods under three different ranking criteria using  $k$ -NN. Each output is the average of 100 models

Ranking criterion ( $k$ -NN)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	4.01	3.50	9.38	6.63	6.64
	$n = 50$	4.66	3.09	9.07	6.19	6.34
	$n = 100$	5.64	2.74	8.96	5.94	5.78
	$n = 200$	6.58	2.34	8.70	5.89	5.81
Positional	$n = 30$	7.24	7.14	9.43	8.17	8.02
	$n = 50$	7.42	7.08	9.39	8.14	7.97
	$n = 100$	7.71	7.04	9.26	8.25	7.74
	$n = 200$	8.02	6.95	9.13	8.21	7.69
F1	$n = 30$	13.37	13.59	21.69	16.17	15.18
	$n = 50$	13.98	13.39	21.33	16.22	15.08
	$n = 100$	15.05	13.16	20.46	17.03	14.30
	$n = 200$	16.33	12.67	19.71	16.82	14.47

Table 4.7: Global scores of the considered methods under three different ranking criteria using LDA. Each output is the average of 100 models

Ranking criterion (LDA)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	5.00	1.98	8.94	6.24	6.47
	$n = 50$	5.74	1.93	8.77	5.65	6.14
	$n = 100$	6.07	1.94	8.51	5.50	5.95
	$n = 200$	6.53	2.08	8.44	5.36	5.92
Positional	$n = 30$	7.57	6.68	9.31	8.17	8.27
	$n = 50$	7.78	6.78	9.28	8.00	8.16
	$n = 100$	7.85	6.90	9.14	8.02	8.09
	$n = 200$	7.99	6.86	9.11	8.01	8.03
F1	$n = 30$	14.69	11.81	20.86	16.51	16.13
	$n = 50$	15.56	12.13	20.60	15.72	15.99
	$n = 100$	15.81	12.39	19.86	16.07	15.87
	$n = 200$	16.29	12.25	20.11	15.79	15.56

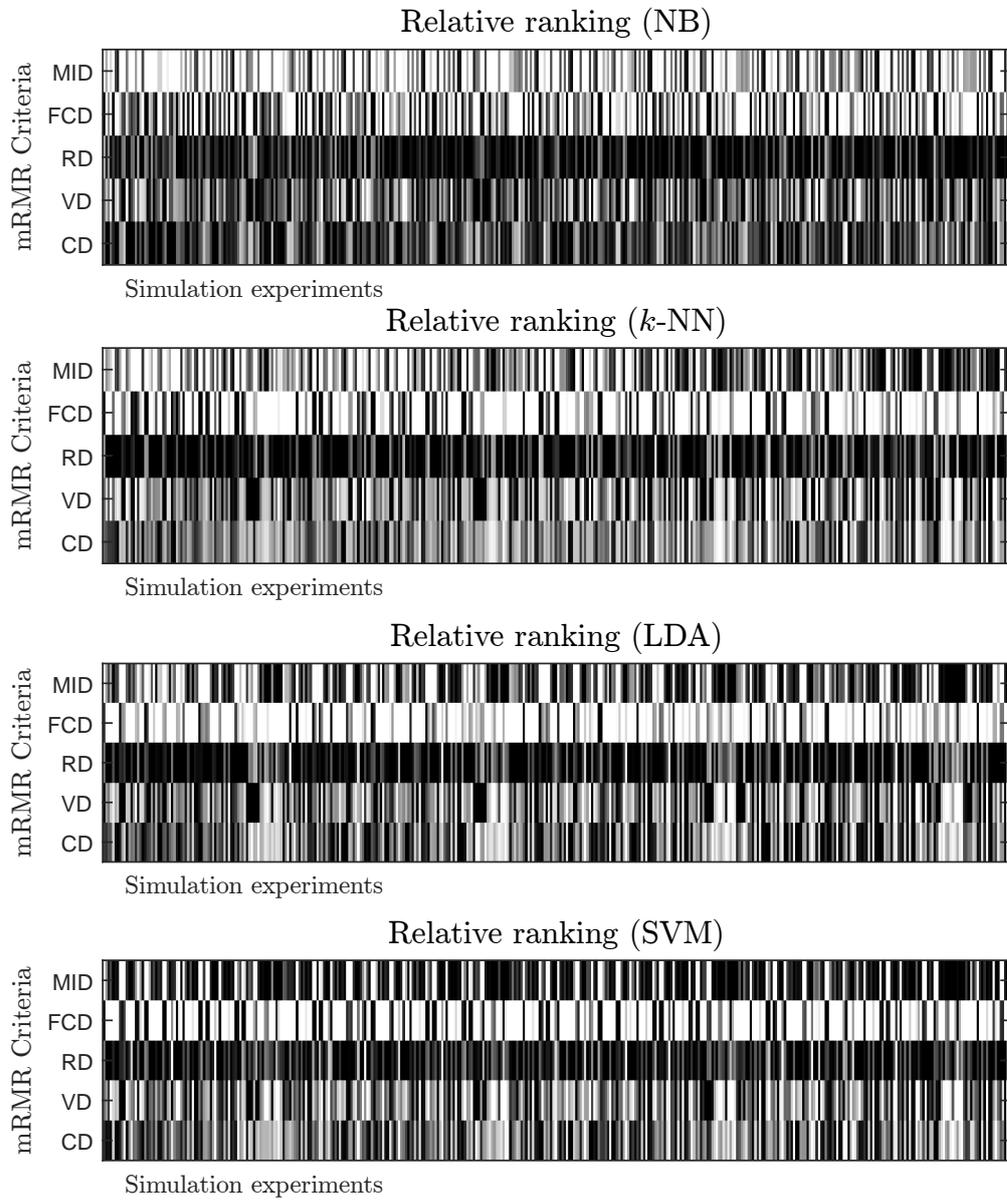


Figure 4.1: Chromatic version of the global relative ranking table taking into account the 400 considered experiments (columns) and the difference-based mRMR versions: the darker the better. From top to bottom displays correspond to with the NB,  $k$ -NN, LDA and SVM classifiers.

Table 4.8: Global scores of the considered methods under three different ranking criteria using SVM. Each output is the average of 100 models

Ranking criterion (SVM)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	6.32	2.99	8.10	5.34	5.57
	$n = 50$	6.63	3	8.28	5.07	5.70
	$n = 100$	6.82	2.87	8.13	4.97	5.59
	$n = 200$	7.19	2.45	8.24	5.06	5.28
Positional	$n = 30$	8.07	7.22	9.06	7.87	7.78
	$n = 50$	8.09	7.20	9.09	7.78	7.84
	$n = 100$	8.22	7.19	9.02	7.84	7.73
	$n = 200$	8.32	7.05	9.15	7.83	7.65
F1	$n = 30$	16.55	13.98	19.63	15.35	14.49
	$n = 50$	16.61	13.86	19.80	14.94	14.79
	$n = 100$	17.17	13.84	19.31	15.29	14.39
	$n = 200$	17.43	13.10	20.10	15.09	14.28

curves in boys and girls), *Tecator* (215, near-infrared absorbance spectra from finely chopped meat) and *Phoneme* (1717 log-periodograms corresponding to the pronunciation of the sounds ‘aa’ and ‘ao’). The respective dimensions of the considered discretizations for these data are 31, 100 and 256. The second derivatives are used for the *Tecator* data. More details on these data are given in Section 5.3.

The methodology followed in the treatment of these data sets is similar to that followed in the simulation study, with a few technical differences. As in the previous Chapter, for *Tecator* and *Growth* data sets, a standard leave-one-out cross-validation is used. Such a procedure turns out to be too expensive (in computational terms) for the *Phoneme* data set. So in this case we have carried out 50-fold cross validation; see, for example, (Hastie et al., 2009, Sec. 7.10) for related ideas.

A summary of the comparison outputs obtained for these data sets using the different mRMR criteria (as well as the benchmark ‘Base’ comparison, with no variable selection) is given in Table 4.9. Again, the letter D in MID, FCD, etc. indicates that the relevance and redundancy measures are combined by difference. The analogous outputs using the quotient (instead of the difference) can be found in [www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx](http://www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx).

The conclusions are perhaps less clear than those in the simulation study. The

Table 4.9: Performances of the different mRMR methods in three real data sets. From top to bottom tables stand for Naive Bayes,  $k$ -NN, LDA and linear SVM outputs respectively.

<b>NB outputs</b>							
Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	92.47	87.10	89.25	87.10	86.02	84.95
	Tecator	98.60	97.67	99.53	99.53	98.14	97.21
	Phoneme	79.03	80.27	80.49	79.39	80.14	74.08
Number of variables	Growth	2.0	1.1	2.2	1.0	1.3	31
	Tecator	2.0	5.9	1.0	1.0	3.3	100
	Phoneme	12.6	10.3	15.8	5.8	15.9	256
<b><math>k</math>-NN outputs</b>							
Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	95.70	83.87	94.62	91.40	84.95	96.77
	Tecator	99.07	99.07	99.53	99.53	99.07	98.60
	Phoneme	80.14	80.48	81.14	80.31	80.55	78.80
Number of variables	Growth	3.5	1.0	2.5	4.8	1.1	31
	Tecator	5.7	3.0	1.0	1.0	4.0	100
	Phoneme	15.4	13.3	17.7	16.5	10.7	256
<b>LDA outputs</b>							
Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	94.62	91.40	94.62	94.62	89.25	-
	Tecator	95.81	93.95	94.88	95.81	94.88	-
	Phoneme	79.50	79.34	79.21	79.39	79.98	-
Number of variables	Growth	3.4	5.0	3.1	4.2	5.0	-
	Tecator	2.6	8.8	5.6	5.0	5.0	-
	Phoneme	19.1	8.8	14.6	17.1	12.0	-
<b>SVM outputs</b>							
Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	94.62	87.10	94.62	95.70	86.02	95.70
	Tecator	98.14	99.07	99.53	99.53	98.60	99.07
	Phoneme	80.90	80.83	80.67	80.78	80.67	80.96
Number of variables	Growth	3.4	5.0	2.5	4.2	5.0	31
	Tecator	6.7	2.0	1.0	1.0	4.1	100
	Phoneme	18.5	8.6	16.2	16.7	16.0	256

lack of a uniform winner is apparent. However, the R-based method is clearly competitive and might even be considered as the global winner, taking into account both, accuracy and amount of dimension reduction. The *Tecator* outputs are particularly remarkable since RD and VD provide the best results (with three different classifiers) using just one variable. Again, variable selection methods beat here the ‘Base’ approach (except for the Growth example) in spite of the drastic dimension reduction provided by the mRMR methods.

### 4.3. A real application: NMR spectral fingerprints

In this section we include some results concerning a real application of the mRMR-RD methodology. This is a product of the collaboration with the Hospital Universitari Vall d’Hebron and Institut de Recerca (VHIR), Barcelona, Spain, and it is further developed in [Barba et al. \(2015\)](#).

The problem of interest here concerns the relation of a high fat diet (HFD) with cardiovascular diseases (in particular, ischemia) in mice, and the possible differences between sexes. The global effects of sex and diet on metabolism are studied by means of metabolomic techniques which consist on the measure of the metabolites in a bio fluid or tissue.  $^1\text{H}$ -NMR metabolic fingerprinting spectra, which are popular in metabolomic studies, are used for the data analysis. In this setting, NMR stands for nuclear magnetic resonance spectroscopy, and  $^1\text{H}$  indicates that hydrogen is used to absorb the electromagnetic radiation. NMR is a robust analytical approach that has been used in the field of metabolomics for years. Although it is less sensitive than other methodologies like mass spectrometry it is easy to automate and, thus, better suited for clinical applications. In a fingerprinting approach, NMR spectra are treated as curves in order to obtain classifiers able to differentiate between various conditions (e.g. cases and controls). Therefore, the application of our methods has full sense.

In summary, the objectives of this work are to evaluate the effects of short term HFD on myocardial metabolism and its interactions with sex using  $^1\text{H}$ -NMR based metabolomics. Our contribution to this study is twofold: to achieve a good classification accuracy in a difficult functional problem and to identify some relevant metabolites. In particular, our proposal is to use the simple LDA classifier after a suitable dimension reduction via mRMR-RD, and use the ranking of variables generated by this algorithm for metabolite identification and further research. The available sample sizes are quite reduced yet so the results are just preliminary and should be understood as a first exploratory approach to the subject.

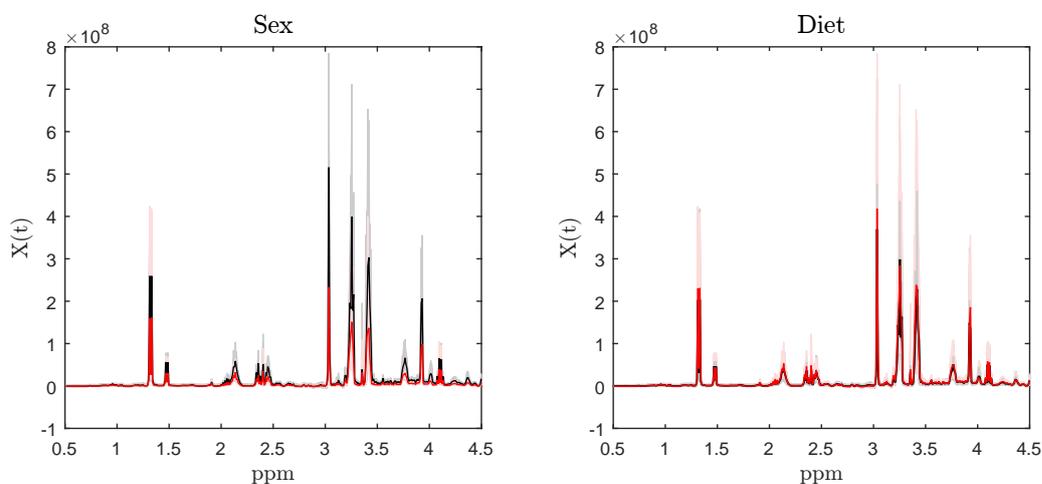


Figure 4.2: NMR spectral trajectories. Observations are coloured according to *sex* (left) and *diet* (right) labels. Colour black stands for male and HDF classes.

The experiments were performed on 23 mice C57BL6 of 16 to 28 weeks of age. Half of the animals (7 males, 5 females) were given a high fat diet (HFD) ad libitum for two weeks. The other half (6 males, 5 females) were given standard chow. NMR spectroscopy was performed on a 9.7 T vertical bore magnet interfaced with a Bruker Avance 400 spectrometer. Spectra from extracts consisted in the accumulation of 32 scans with a fully relaxed pulse-and-acquire sequence. All spectra were acquired at 30C. For the analysis we have used the aliphatic part of the spectra from cardiac tissues. This part, between 0.5 and 4.5 ppm, is discretized in 438 equispaced points. We study two different binary classification problems with these data: *sex* (male, female) and *diet* (HFD, control). Figure 4.2 shows the 23 trajectories and the class means for both problems with different colours for different classes. A first visual inspection reveals that the diet problem is more difficult since the mean functions of the classes are nearly overlapped.

It is worth mentioning that spectra normalization and variable scaling to unit variance (preprocessing techniques often used in NMR spectra) did not improve the classification results. Likewise, smoothing via splines was also tested with similar results. Therefore, the simple raw data (that is, the original spectra with no dimension reduction technique or scale transformation applied on them) are considered hereafter.

We have performed several supervised classification techniques, for both characteristics, sex and diet. As can be seen from Figure 4.2, discrimination in *sex*

Table 4.10: Classification matrices for *sex* (left) and *diet* (right) problems. Results are obtained with a 3-NN classifier over the entire curves.

	Sex		Diet		
	$\hat{M}ale$	$\hat{F}emale$	$\hat{H}DF$	$\hat{c}ontrol$	
<i>Male</i>	<b>9</b>	4	<i>HDF</i>	<b>5</b>	7
<i>Female</i>	1	<b>9</b>	<i>control</i>	5	<b>6</b>

problem seems easier than in *diet* (this is confirmed by the numerical outputs). Nevertheless, the roughness of the data make difficult to deal with them as they stand. This is shown in a first preliminary approach on the raw data, intended just as a benchmark reference. Since the standard linear classification method (LDA) cannot be directly used with the high-dimensional spectra data, we have employed the  $k$ -nearest neighbours ( $k$ -NN) classifier, with  $k=3$ . This is a "plain", assumption free, methodology with a minimal data processing. Table 4.10 shows the classification matrices (or "confusion matrices") corresponding to this preliminary spectra classification methodology. Columns correspond to predicted values while rows stand for the true ones. Correct classified items are marked in bold. In all cases, the classification errors have been obtained using a leave-one-out methodology.

These results are quite poor, in special for the diet problem. As a further alternative we propose performing classification on the result of applying a dimension reduction technique and then, in a second stage, using LDA. In the dimension reduction step we have in turn checked two methodologies: variable selection through the mRMR-RD method and PLS (which can be considered the standard). Again, the usual leave-one-out methodology is used to assess the proportion of correct classification. Table 4.11 shows the classification matrix obtained with mRMR-RD + LDA, while Table 4.12 corresponds to that resulting from using LDA after PLS. As before, columns and rows correspond to the predicted and the true values values respectively, and correct classified items are marked in bold.

In short, both alternative approaches, mRMR+LDA and PLS+LDA, resulted in a greatly increased classification success (with respect to classification based on the raw data with no dimension reduction). While there were no relevant differences in classification performance between mRMR and PLS, the use of mRMR for variable selection allows for an easier interpretation of the results. In this case, mRMR+LDA is able to correctly classify 21 out of 23 samples with just 2 variables in *sex* and 8 variables in *diet*, among the original 438 points.

Table 4.11: Classification matrices for *sex* (left) and *diet* (right) problems. Results are obtained with a mRMR-RD + LDA.

	Sex			Diet	
	$\hat{M}ale$	$\hat{F}emale$		$\hat{H}DF$	$\hat{c}ontrol$
<i>Male</i>	<b>11</b>	2	<i>HDF</i>	<b>12</b>	0
<i>Female</i>	0	<b>10</b>	<i>control</i>	2	<b>9</b>

Table 4.12: Classification matrices for *sex* (left) and *diet* (right) problems. Results are obtained with PLS + LDA.

	Sex			Diet	
	$\hat{M}ale$	$\hat{F}emale$		$\hat{H}DF$	$\hat{c}ontrol$
<i>Male</i>	<b>10</b>	3	<i>HDF</i>	<b>11</b>	1
<i>Female</i>	0	<b>10</b>	<i>control</i>	1	<b>10</b>

Let us recall that as a final outcome of PLS, a few linear combinations of the whole set of original variables are provided. These synthetic components are typically difficult to interpret. By contrast, mRMR selects a few “representative” variables from each spectrum; so the method provides a dimension reduction in terms of some selected original variables. Indeed, this fact can be exploited in a further research centred on the relevant metabolites. In the case of sex, the variables more frequently selected in the leave-one-out classification process were those corresponding to myo-inositol, taurine and glutamate. In the case of diet, selected variables showed a decrease in creatine, taurine and citrate in HFD fed mice as compared to their control fed counterparts. Moreover, on closer inspection, some of the selected variables could give us some insights about possible interactions between sex and diet. For example, it could be seen that the differences in diet arise mainly from male animals. Also, it seems that female hearts (both control and HFD) tend to cluster with hearts from male animals fed with HFD in some metabolites. Figure 4.3 shows the display of the data projected on two relevant variables (metabolites). The sex is indicated by the corresponding symbol and colours denote the type of diet (red for HFD and black for control). In this case, both variables separate males and females very well (which suggest different concentrations of the associated metabolites in both populations) but, more interestingly, the combination of both metabolites seems to form four clusters dividing the sample in the four possible cases.

In conclusion, the analysis of the NMR spectra via mRMR shows difference

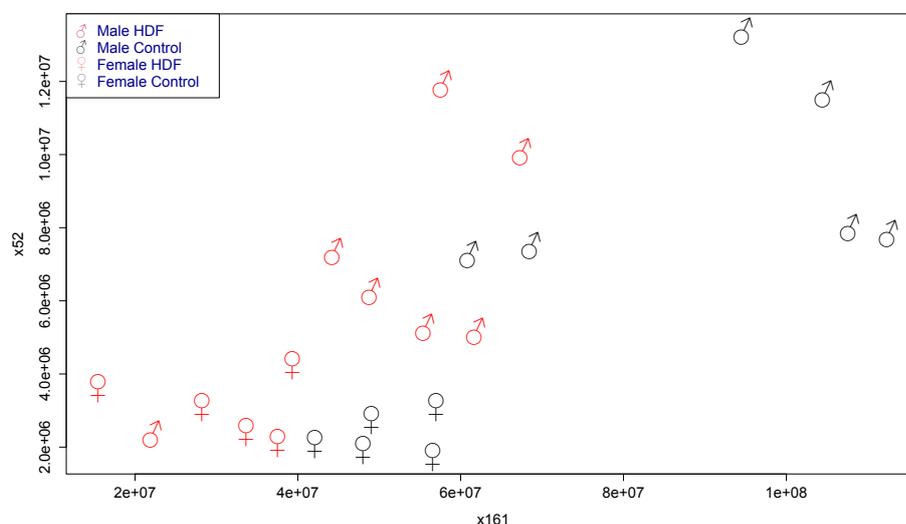


Figure 4.3: Projection on two relevant variables. Red and black indicates HDF and control diets respectively, while each sex is represented by its symbol.

between sex and diet in mice. In fact, the differences highlighted by the mRMR allow to achieve a very good classification performance. In addition, the variables selected by the algorithm make possible a further research that suggests probable interactions between sex and diet. However, the sample size is too small for stronger conclusions, even though we have used simple methods with few parameters and all leave-one-out cross validation. These results together with other experiments are used in [Barba et al. \(2015\)](#) to extract some clinical implications and biological conclusions about this problem.

Finally, from a statistical point of view we can conclude that the metabolomic analysis using variable selection combined with linear discrimination appears as a good strategy in terms of both, accuracy and interpretability. In particular mRMR-RD seems preferable to dimension reduction via PLS.

#### 4.4. Final conclusions and comments

The mRMR methodology has become an immensely popular tool in the machine learning and bioinformatics communities. For example, the papers by [Ding and Peng \(2005\)](#) and [Peng et al. \(2005\)](#) had 983 and 3047 citations, respectively on Google Scholar (by August 31, 2015). As we have mentioned, these authors explicitly pointed out the need of further research, in order to get improved versions

of the mRMR method. The idea would be to keep the basic mRMR paradigm but using other association measures (besides the mutual information). This work exactly follows such line of research, with a particular focus on the classification problems involving functional data.

We think that the results are quite convincing: our extensive simulation study (based on 1600 simulation experiments and real data) places the mRMR method based in the R association measure by Székely et al. (2007) globally above the original versions of the mRMR paradigm. This is perhaps the main conclusion of this Chapter. The good performance of the distance correlation in comparison with the other measures can be partially explained by the fact that this measure captures non-linear dependencies (unlike C and FC), has a simple smoothing-free empirical estimator (dissimilar to MI) and is normalized (different from V).

Furthermore, the results in previous section shows that the R-based mRMR is completely feasible in real applications with functional data. Indeed, the classification of NMR spectra after mRMR-RD selection achieves accuracy levels which are far from being accomplished classifying the entire curves. Likewise, the R-based selection is also better suited for the classification task than the PLS projection, and in addition, mRMR-RD identify relevant metabolites that can be used in further research.

There are, however, some other more specific comments to be made.

1. Once again we can see that intrinsic variable selection is worthwhile in functional data analysis. Accuracy can be kept (and often improved) using typically less than the 10% of the original variables, with the usual benefits of the dimension reduction. This phenomenon appears in all the considered classifiers.
2. The average number of selected variables with the R- or V-based methods is also smaller than that of MI and FC (that is, the standard mRMR procedures). This entails an interpretability gain: the fewer selected variables, the stronger case for interpreting the meaning of such selection in the context of the considered problem.
3. The advantage of the R-based methods over the remaining procedures is more remarkable for the case of small sample sizes. This looks as a promising conclusion since small samples are very common in real problems (e.g. in biomedical research).

4. In those problems involving continuous variables there is a case for using non-parametric kernel density estimators in the empirical approximation of the mutual information criterion. However, these estimators are known to be highly sensitive to the selection of the smoothing parameter, which can be seen as an additional unwelcome complication. On the other hand, the results reported so far (e.g. in Peng et al. (2005)) do not suggest that kernel estimators will lead to a substantial improvement over the simplest, much more popular discretization estimators (see e.g. Brown et al. (2012)).
5. Still in connection with the previous remark, it is worth noting the lack of smoothing parameters in the natural estimators of  $V$  and  $R$  (see Definition 3.2). This can be seen as an additional advantage of the  $R$ - or  $V$ -based mRMR method over the main drawback of MI (Vergara and Estévez, 2014).
6. The better performance of  $R$  when compared with  $V$  can be explained by the fact that  $R$  is normalized so that relevance (4.4) and redundancy (4.5) are always measured ‘in the same scale’. Otherwise, one of these two quantities could be overrated by the mRMR algorithm, specially when the difference criterion is used. It is related with the “balance” phenomenon explained in Brown et al. (2012), so bounded dependence measures must be recommended for mRMR.
7. The method FCD (sometimes suggested in the literature as a possible good choice) does not appear to be competitive. It is non-bounded and unable to detect non-linear dependencies. It is even defeated by the simple correlation-based method CD.
8. In general, the difference-based methods are preferable to their quotient-based counterparts. The quotient-based procedures are only slightly preferable when combined with methods (FC, V) where relevance and redundancy are expressed in different scales. The outputs for these quotient-based methods can be found in the complete list of results [www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx](http://www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx).
9. Finally, if we had to choose just one among the considered classification methods, we should probably take  $k$ -NN. The above commented advantages in terms of ease of implementation and interpretability do not entail any significant price in efficiency.

# Chapter 5

## On the empirical studies

In this chapter we explain all aspects related to the empirical studies we have carried out. The aim is to avoid the duplication of information, to improve readability and to ensure that our experiments are reproducible. More specifically, the present chapter aims at giving the interested reader some design considerations, implementation details and additional information on the data and methods under study. Also, some complementary empirical results are given.

All the methods considered somewhere in the present work (both dimension reduction procedures and classifiers) are described in [5.1](#), with special attention to the implementation details. Section [5.2](#) is devoted to the simulation study: simulation models are described and the methodology is explained. Finally, Section [5.3](#) contains a description of the real datasets considered in this work and the methodological differences with respect to the simulation framework.

### 5.1. Methods and implementation

Our empirical results cover a wide range of methodologies, for both dimension reduction and classification. Some of them have been described above, at the appropriate places, and others have been omitted or just appear in the on-line materials for the sake of brevity and readability. In this section we describe in more detail all of these methods with the exception of the new proposals which have been fully explained in the corresponding chapters. Finally, we give some computational details.

### 5.1.1. Dimension reduction methods

*RK-VS.* The RKHS-based variable selection method is described in Chapter 2. Note that this is one of the original proposals in this thesis.

*MH.* See Chapter 3 for the description of maxima-hunting methods (our second proposal for variable selection). Note that we have considered two versions, based on the maxima of both  $dcov$  (MHV) and  $dcor$  (MHR).

*mRMR.* See Chapter 4 for the description of the minimum Redundancy Maximum Relevance method. Let us recall that the nomenclature of this method depends on the relevance measure and the association criterion. The considered measures are: mutual information (MI),  $dcor$  (R),  $dcov$  (V), the Fisher-correlation criterion (FC), and the absolute value of the standard correlation (C). Our new proposals are based on the use of  $dcor$  and  $dcov$  (Székely et al., 2007) association measures. In order to combine the relevance and redundancy measures in the mRMR methodology we have used both the difference (suffix D) and quotient criterion (suffix Q) in the experiments, although the latter has been mostly relegated to the on-line material since, in general, it offers worse results.

*T.* The Univariate  $t$ -ranking method is frequently used when selecting relevant variables (see e.g. the review by Fan and Lv (2010)). It is based on the simple idea of selecting the variables  $X_t$  with highest Student's  $t$  two-sample scores

$$T(X_t) = \frac{|\bar{X}_{1t} - \bar{X}_{0t}|}{\sqrt{S_{1t}^2/n_1 + S_{0t}^2/n_0}},$$

where  $\bar{X}_{it}$  and  $S_{it}^2$  denotes the sample mean and variance of the variable  $X_t$  in the class  $i$ .  $T$  is related with the correlation measures and can present good results when few non-redundant variables are needed, However, it proves unsuitable in many functional problems.

*MaxRel.* Maximum relevance is the name given to the ranking method which uses the mutual information; see e.g. Peng et al. (2005). Thus, according to MaxRel, variables are sorted by the score  $MI(X_t, Y)$  and the top scored ones are selected. It involves some estimation problems due to the use of the mutual information (see Subsection 4.1.1). MaxRel outputs are not included since they are similar to those of  $T$  with a smaller classification success. Nevertheless, we can see an example of MaxRel performance in Figure 1.6.

*DHB.* We have denoted by DHB the variable selection method proposed in [Delaigle et al. \(2012\)](#). Let us note, as an important difference with our proposals, that the DHB procedure is a “wrapper” method, in the sense that it depends on the chosen classifier (see Subsection 1.4.2). Given a classifier, the DHB 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). These time limitations are the reason why we have only applied the DHB method in the real data sets. Moreover, following [Delaigle et al. \(2012\)](#), we have only used this method with the Fisher’s linear classifier (LDA) since the other classification rules considered in this thesis have not analyzed in the DHB paper. According to our experiments (see Subsection 3.4.3) the extra computational costs associated with the DHB methodology do not entail any a significant accuracy gain in return. However, our results are yet too limited in extent. So, further research should be done to draw general conclusions.

*PLS.* According to the available results ([Preda et al. \(2007\)](#); [Delaigle and Hall \(2012b\)](#)) PLS is the “method of choice” for dimension reduction in functional classification. This is due to the fact that the response variable is involved in the computation of the PLS projections. In particular, this procedure aims at maximizing the covariance between the new components in the reduced space and the class label  $Y$ . Note however that PLS is not a variable selection procedure; it does not provide a few selected variables but rather a number of linear combinations of the original variables. So, PLS lacks the interpretability of variable selection. In some sense, the motivation for including PLS is to check how much do we lose by restricting ourselves to variable selection methods, instead of considering other more general linear projections procedures (as PLS) for dimension reduction. In fact, our experiments show the good performance of PLS, but it is somewhat surprising that our proposals (based on the more restrictive criterion of just selecting a few original variables) often outperform PLS. This is particularly true for our methods MHR and RK-VS. In general terms, The relative performance of PLS tends to be better for complex problems but it is clearly outperformed by variable selection methods when few relevant variables are involved.

*PCA.* We have also tested the popular Principal Component Analysis (PCA). As PLS, this approach (briefly commented already in the introductory chapter) relies on the use of linear projections to reduce the dimension. In this case the objective is to preserve the variance of the original data in the reduced space. PCA is by far the most used method for dimensionality reduction; lots of references are available, see for instance [Ramsay and Silverman \(2005\)](#) for its adaptation to the

functional setting. As expected, our results show that PLS is preferable to PCA in the classification setting, so PCA results are not included in this thesis. However, it is worth mentioning that the distance between methodologies is drastically shortened when linear classifiers are involved.

*Oracle.* By “oracle” we mean a virtual, unfeasible procedure based on the use of all the “really relevant” variables, that is, those variables which are actually involved in the expression of the Bayes rule. Since these “truly relevant” variables are unknown in practice, we use this method just as a reference for illustrative purposes. In some sense, this selection is the optimal one, so the result of a classifier built over these variables is a suitable bound. The relevant points for each simulation model are known by construction or derived from theoretical results (see Sections 2.2 and 3.3). Such relevant variables are indicated in the list of models (Appendix A).

*Base.* This is another benchmark procedure. It is just the result of applying any given classifier to the entire functional data with no dimension reduction at all. In general, the Base performance can be seen as a reference to assess the usefulness of dimension reduction methods. Somewhat surprisingly, this Base procedure is often outperformed by variable selection methods. Among the evaluated classifiers, the best Base results were achieved by SVM. Note that the Base method cannot be implemented with LDA since this classifier typically fails with infinite or high-dimensional data.

*Random.* A uniformly random variable selection method was also implemented as a naive benchmark. In principle, variable selection methods outperformed by a Random selector would not make sense. Although Random is usually the worst method (specially in high-dimensional and complex problems), it surprisingly outperforms (on average) the ranking methods and the mRMR-FC approach, which highlighted the inappropriateness of these methods in FDA. On the other hand, the unexpected competitive results in some examples might open a door for further research in random selection. Random results cannot be found in this document but in the on-line material.

### 5.1.2. Classifiers

In order to compare the different methods above, we use the natural accuracy measure, defined by the percentage of correct classification. Hence, we need several classifiers with different strategies aiming at covering the widest possible

range of approaches. The classifiers used in our study are roughly those considered in [Ding and Peng \(2005\)](#) with the addition of  $k$ -NN. All of them are simple methods broadly used in the variable selection literature, which generally achieve good performances. We give next a brief description of these classifiers; further details can be found in standard references such as [Hastie et al. \(2009\)](#) or [Duda et al. \(2012\)](#). Similar comparisons could be done with other classifiers, since the considered methods do not depend on the classifier.

- **The  $k$ -Nearest Neighbors classifier ( $k$ -NN).** An all purposes and easy to interpret, non-parametric classifier. According to this method a new observation is assigned to the class of the majority of its  $k$  closest neighbours. We use the usual Euclidean distance (or  $L^2$ -distance when the method is used with the complete curves) to define the neighbours. The parameter  $k$  is fitted through the validation step, as explained below.
- **Linear Discriminant Analysis (LDA).** The classic Fisher's linear discriminant ([Fisher, 1936](#)) is, still today, the most popular classification method among practitioners. It is known to be optimal under gaussianity and homoscedasticity of the distributions in both populations but, even when these conditions are not fulfilled, LDA tends to show a good practical performance in many real data sets; see, e.g., [Hand \(2006\)](#). Finally, let us recall that LDA is only used over the reduced data since it is unfeasible for the complete curves.
- **Support Vector Machine (SVM).** This is one of the most popular classification methodologies in the last two decades. The basic idea is to look for the 'best hyperplane' in order to maximize the separation margin between the two classes. The use of different kernels (to send the observations to higher dimensional spaces where the separation is best achieved) is the most distinctive feature of this procedure. The most common kernels are linear and Gaussian. As in [Ding and Peng \(2005\)](#) we have used linear kernels, which are easier to both train interpret.
- **Naïve Bayes classifier (NB).** This method relies on the assumption that the selected variables are Gaussian and conditionally independent in each class. So a new observation is assigned according to its posterior probability calculated from the Bayes rule. Of course the independence assumption will often fail (especially in the case of functional data). However, as shown in [Ding and Peng \(2005\)](#); [Arauzo-Azofra et al. \(2011\)](#) among others, this rule works as an heuristics which offers sometimes a surprisingly good practical performance.

In general terms, our extensive simulation study shows that SVM and  $k$ -NN are preferable to LDA and NB. SVM achieves the highest accuracy rates but it is unfeasible in some “non-linear” problems (those in which the class means are very similar). On the other hand,  $k$ -NN is slightly outperformed by SVM, although it is feasible for all considered problems and it is easier to train and interpret. We have also considered other classifiers: the logistic regression and our RK-C (based on the RKHS theory and the sparsity assumption). The logistic regression is very similar to LDA so it is not included. RK-C is well explained in Section 2.4. It has the advantages and disadvantages of a linear classifier with the additional benefits derived from the variable selection. Let us also recall that this rule can achieve optimal results in several models. Finally, in Section 2.5 we have also compared RK-C outputs with those obtained in [Delaigle and Hall \(2012a\)](#) which consist of three versions of the centroid classifier defined in the paper and a classifier based on the nonparametric functional regression ([Ferraty and Vieu, 2006](#)).

### 5.1.3. Computational details

Our empirical study required the implementation of all methods described above, including both dimension reduction algorithms and classifiers. The code has been written in MATLAB. It is available upon request. It is also our intention to prepare an user-friendly R library or MATLAB toolbox. Here are some algorithmic details:

- We have implemented the minimum Redundancy Maximum Relevance algorithm in order to allow us to introduce different association measures (such as the distance correlation) in the definition of the method. The original version of mRMR (based on the mutual information measure) is available from <http://penglab.janelia.org/proj/mRMR/>. Also, a MATLAB/C++ function (not compatible with the current MATLAB versions) can be also downloaded from that URL.

Following [Ding and Peng \(2005\)](#), the criteria (4.4) and (4.5) are in fact replaced in practice by approximate expressions, numbered (6) and (7) in [Ding and Peng \(2005\)](#). Hence, the criterion we optimize in our experiments is

$$I(X_t, Y) - \frac{1}{|S|} \sum_{s \in S} I(X_t, X_s),$$

or alternatively the analogous quotient expression. As these authors point out, the first term is equivalent to the relevance criterion (4.4) while the second provides an approximation for the minimum redundancy criterion (4.5) when maximizing.

- We have implemented the original iterative PLS algorithm that can be found, e.g. in [Delaigle and Hall \(2012b\)](#). On the other hand, PCA uses the MATLAB function *pcacov*.
- We use the empirical estimators of distance correlation and distance covariance given in [Definition 3.2](#), which are also implemented in an efficient way by means of the function *pdist2*. We have also seen that this estimator is uniformly convergent ([Lemma 3.1](#)).
- The mutual information is computed in the discrete version given in [\(4.2\)](#). Following [Ding and Peng \(2005\)](#), the limits of the discretization intervals are defined by the mean of the corresponding variable  $\pm\sigma/2$  (where  $\sigma$  is the standard deviation). Other discretizations were proved with similar or worse results.
- The DHB algorithm has been implemented according to the instructions given in [Delaigle et al. \(2012\)](#). The authors implement a slightly modified version, which we have closely followed. It is based on a sort of trade-off between full and sequential search, together with some additional computational savings. We have also used the same parameters and the first stopping criterion proposed by these authors.
- Our  $k$ -NN implementation is built around the MATLAB function *pdist2* and allows for the use of different distances; we employ the usual Euclidean distance. Also, the computation for different  $k$ 's can be simultaneously made with no additional cost.
- Our LDA is a faster implementation of the MATLAB function *classify*.
- The Naïve Bayes classifier is based on the MATLAB functions *NaiveBayes.fit* and *predict*.
- The linear SVM has been performed with the MATLAB version of the LIBLINEAR library (see [Fan and Lv \(2008\)](#)) using the parameters *bias* and *solver type 2*. It obtains (with our data) very similar results to those of the default *solver type 1*, but faster. LIBLINEAR is much faster than the more popular LIBSVM library when using linear kernels.
- The number  $k$  of nearest neighbours in the  $k$ -NN rule, the cost parameter  $C$  of the SVM classifier and the number of selected variables are chosen by standard validation procedures ([Guyon et al., 2006](#); [Hastie et al., 2009](#)). The validation step is described in the next Sections. The derivatives (when needed) have been done via splines with the R package *fda.usc* (see [Febrero-Bande and Oviedo de la Fuente \(2012\)](#)).

## 5.2. Simulations

To our knowledge, this is the largest empirical study so far on variable selection. We have included 100 different models defined in terms of the most usual stochastic processes and variants of them. As we are interested in variable selection, a natural choice was to include in the study some models in which the optimal classification rule depended only on a finite number of variables. Note however that, the presence of "suitable" models would not favour necessarily our proposals against other dimension reduction methods. In fact, the study includes some models for which some relevant variables do not correspond to maxima, and only 7 examples fulfil all the assumptions of RK-VS model. Also, there is no reason to think that the many logistic-type models (and the real data examples) included in our experiments, are especially favourable to our proposals. Finally, one might expect that these "sparse" models (depending only on a finite number of variables) were always more suitable for variable selection methods than for partial least squares but, somewhat surprisingly, our empirical study shows that this is not exactly the case.

### 5.2.1. Models

Our simulation study consists of 400 experiments (100 models  $\times$  4 sample sizes), aimed at comparing the practical performances of several intrinsic variable selection methods (and other dimension reduction procedures) described in the previous Section. These experiments are obtained by considering 100 different underlying models and 4 sample sizes, where by "model" we mean either,

- (M1) a pair of distributions for  $X|Y = 0$  and  $X|Y = 1$  (corresponding to  $P_0$  and  $P_1$ , respectively); in all cases, we take  $p = \mathbb{P}(Y = 1) = 1/2$ .
- (M2) The marginal distribution of  $X$  plus the conditional distribution  $\eta(x) = \mathbb{P}(Y = 1|X = x)$ .

Models vary in difficulty and number of relevant variables. In all the considered models the optimal Bayes rule turns out to depend on a finite number of relevant variables, see Sections 2.2 and 3.3. The processes involved include also different levels of smoothing. The full list of considered models is in Appendix A. All of them belong to one of the following classes:

**Gaussian models:** they are denoted  $G1, G1b, \dots, G8$ . All of them are generated according to the general pattern (M1). In all cases the distributions of  $X(t)|Y = i$  are chosen among one of the following types: first, the **standard Brownian Motion**,  $B$ , in  $[0, 1]$ , i.e., a Gaussian process with  $\mathbb{E}(B(t)) = 0$  and

covariance function  $\gamma(s, t) = \min\{s, t\}$ . Second, **Brownian Motion,  $BT$ , with a trend**  $m(t)$ , i.e.,  $BT(t) = B(t) + m(t)$ ; we have considered several choices for  $m(t)$ : a linear trend,  $m(t) = ct$ , a linear trend with random slope, i.e.,  $m(t) = \theta t$ , where  $\theta$  is a Gaussian r.v., and different members of two parametric families: the *peak* functions  $\Phi_{m,k}$  and the *hillside* functions, defined by

$$\Phi_{m,k} = \int_0^t \varphi_{m,k}(s) ds \quad , \quad \text{hillside}_{t_0,b}(t) = b(t - t_0)\mathbb{I}_{[t_0, \infty)},$$

where,  $\varphi_{m,k}(t) = \sqrt{2^{m-1}} \left[ \mathbb{I}_{\left(\frac{2k-2}{2^m}, \frac{2k-1}{2^m}\right)} - \mathbb{I}_{\left(\frac{2k-1}{2^m}, \frac{2k}{2^m}\right)} \right]$  for  $m \in \mathbb{N}$ ,  $1 \leq k \leq 2^{m-1}$ .

Third, the **Brownian bridge**:  $BB(t) = B(t) - tB(1)$ . Fourth, is the **Ornstein-Uhlenbeck process**, with a covariance function of type  $\gamma(s, t) = a \exp(-b|s - t|)$  and zero mean (*OU*) or different mean functions  $m(t)$  (*OUT*). Finally smoother processes have been also computed by convolving Brownian trajectories with Gaussian kernels. We have considered two levels of smoothing denoted by sB and ssB.

**Logistic models:** they are defined through the general pattern (M2). The process  $X = X(t)$  follows one of the above mentioned distributions and  $Y \sim \text{Binom}(1, \eta(X))$  with

$$\eta(x) = \frac{1}{1 + e^{-\Psi(x(t_1), \dots, x(t_d))}},$$

a function of the relevant variables  $x(t_1), \dots, x(t_d)$ . We have considered 15 versions of this model and a few variants, denoted  $L1, L2, L3, L3b, \dots, L15$ . They correspond to different choices for the link function  $\Psi$  (both linear and nonlinear) and for the distribution of  $X$ . For example, in the models L2 and L8 we have  $\Psi(x) = 10x_{30} + 10x_{70}$  and  $\Psi(x) = 10x_{50}^4 + 50x_{80}^3 + 20x_{30}^2$ , respectively. All the link functions considered can be found in Appendix A.

**Mixtures:** they are obtained by combining (via mixtures) in several ways the above mentioned Gaussian distributions assumed for  $X|Y = 0$  and  $X|Y = 1$ . These models are denoted M1, ..., M11 in the output tables.

### 5.2.2. Methodology

For each model, all the selection methods are checked for four sample sizes ( $n = 30, 50, 100, 200$ ). The experiment is completed with a classifier (which acts on the selected variables) in order to assess the performance. In this way we get  $100 \times 4 = 400$  experiments for each classifier under study.

All the functional simulated data are discretized to  $(x(t_1), \dots, x(t_{100}))$ , where  $t_i$  are equispaced points in  $[0, 1]$ . In fact (to avoid the degeneracy  $x(t_0) = 0$  in the Brownian-like models) we take  $t_1 = 6/105$ . Similarly, for the case of the Brownian bridge, we truncate as well at the end of the interval.

In practice, all procedures are implemented in a sequential way: the variables are sequentially selected until some stopping criterion is fulfilled. In our case, the dimension of the reduced space (number of variables or components) is set by standard data-based validation procedures. Parameter validation can be carried out mainly through a validation set or by cross-validation on the training set [see e.g. [Guyon et al. \(2006\)](#)]. In the case of the simulation study, the validation and test samples are randomly generated. Other parameters involved such as the number  $k$  of nearest neighbours in the  $k$ -NN classifier, the cost parameter in SVM and the smoothing parameter  $h$  in maxima-hunting methods, are fixed in the same validation step.

In summary, the methodology used in the simulation study is as follows (see also the flowchart in [Figure 5.1](#)):

1. In each run of the simulation experiments three independent samples are generated: the training sample of size  $n$  ( $= 30, 50, 100, 200$ ), a validation sample of size 200 and a test sample of size 200.
2. The relevant variables are selected using the training sample (alternatively the PLS-PCA projections are computed).
3. The parameters are fitted through the validation sample.
4. The data are reduced according the result of the previous steps.
5. The “accuracy” outputs correspond to the percentages of correct classification obtained for the reduced test samples, that is, the samples obtaining by replacing the functional data with the corresponding multivariate data made of the selected variables. In all cases the classifier is built from the reduced training sample.
6. The final outputs are based on the average over 200 independent runs of the whole procedure.

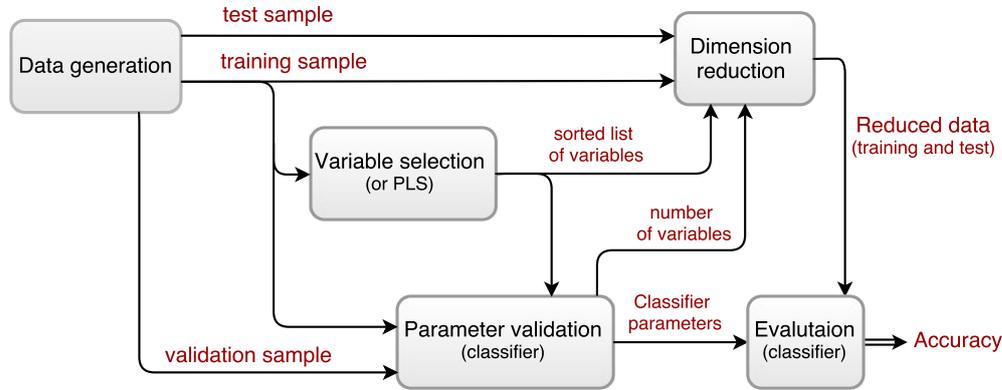


Figure 5.1: Methodology flowchart for simulations. This process is repeated 200 times for each experiment.

### 5.2.3. Additional results

Here we show some additional outputs of our simulation study with a twofold objective: to give a more detailed information about the different models involved in the benchmark (until now we have only divided the models by sample size), and to show all proposed methods in a single table. So, Table 5.1 shows the classification accuracy (percentage of correct classification) for different groups of models and methods. For clarity purposes we only present the results obtained with one classifier. We have chosen the  $k$ -NN rule since it is the best method which is suitable for all the 100 considered models (SVM is unfeasible in a few of them). Anyway, results from the other considered classifiers are quite similar in relative terms, which is just a consequence of the intrinsic approach.

The methods appear in columns; together with our new proposals we have included PLS and the Base approach for comparative purposes. The simulation outputs are grouped in different categories (in rows) by model type and sample size  $n$ . The rows are labelled by the general model type, that is, logistic, Gaussian and mixtures. The logistic models are also divided by the type of process involved according to the notation given in Subsection 5.2.1. RKHS denote the models that fulfil the hypotheses of RK-VS and “All models” include the outputs of all the 100 considered models for each  $n$ . We have followed the methodology described above and the outputs are averaged over 200 independent runs. The marked values correspond to the two best performances in each row. Analogously, Table 5.2 shows the results for the number of selected variables.

In view of Table 5.1 we can conclude that all dimension reduction methods (in-

Table 5.1: Average percentage of correct classification using  $k$ -NN

Output	$n$	mRMR-RD	MHR	RK-VS	PLS	Base
All models	30	81.30	81.87	81.39	81.42	78.98
	50	82.44	82.89	82.86	82.48	80.34
	100	83.82	84.21	84.70	83.79	81.99
	200	84.89	85.37	86.21	84.84	83.38
Logistic OU	30	78.71	79.20	78.58	79.22	75.63
	50	79.64	80.02	79.98	80.04	76.87
	100	80.96	81.26	81.66	81.13	78.44
	200	82.10	82.56	83.21	82.07	79.73
Logistic OU <sub>t</sub>	30	81.87	82.30	81.91	82.71	79.50
	50	82.83	83.18	83.13	83.52	80.62
	100	84.12	84.33	84.90	84.52	82.02
	200	85.00	85.30	86.23	85.31	83.14
Logistic B	30	83.29	83.94	83.94	84.01	81.10
	50	84.38	84.90	85.47	85.08	82.35
	100	85.68	86.31	87.40	86.30	83.92
	200	86.78	87.63	89.27	87.39	85.35
Logistic sB	30	84.00	84.55	84.40	84.48	81.90
	50	84.87	85.31	85.65	85.36	83.02
	100	86.09	86.62	87.51	86.61	84.44
	200	87.07	87.84	89.17	87.58	85.73
Logistic ssB	30	85.92	86.35	86.39	85.97	84.47
	50	86.86	87.11	87.49	86.78	85.41
	100	87.93	88.05	88.89	87.86	86.71
	200	88.89	88.75	90.24	88.81	87.91
Gaussian	30	81.09	82.47	81.03	79.68	79.18
	50	82.23	83.60	82.35	80.91	80.89
	100	83.20	84.50	83.76	82.31	82.52
	200	83.77	84.98	84.37	83.33	83.80
Mixtures	30	73.13	73.32	72.09	71.59	70.27
	50	75.51	75.56	74.86	73.80	72.86
	100	78.20	77.95	77.76	76.38	75.84
	200	80.15	80.02	79.82	78.16	78.22
RKHS	30	83.96	85.79	86.16	85.35	83.20
	50	84.80	86.68	87.62	86.61	84.99
	100	85.69	87.58	88.91	87.85	86.61
	200	86.30	88.19	89.68	88.74	87.94

Table 5.2: Average number of selected variables (or PLS components) using  $k$ -NN

Output	$n$	mRMR-RD	MHR	RK-VS	PLS	Base
All models	30	7.7	6.2	7.8	4.3	100
	50	7.9	6.2	7.6	4.8	100
	100	8.2	6.1	7.0	5.5	100
	200	8.5	5.8	6.6	6.2	100
Logistic OU	30	7.8	6.8	7.8	4.2	100
	50	8.2	6.9	7.6	4.8	100
	100	8.4	7.0	7.0	5.5	100
	200	8.4	6.6	6.6	6.3	100
Logistic OUt	30	8.2	7.1	8.1	3.9	100
	50	8.6	7.0	7.8	4.4	100
	100	8.7	6.9	7.3	5.1	100
	200	8.7	6.9	6.8	5.8	100
Logistic B	30	7.8	6.7	7.7	4.3	100
	50	7.9	6.7	7.3	4.8	100
	100	8.2	6.7	6.7	5.7	100
	200	8.4	6.3	6.0	6.6	100
Logistic sB	30	7.8	6.7	7.8	4.2	100
	50	7.9	6.7	7.5	4.9	100
	100	8.2	6.5	6.8	5.5	100
	200	8.5	6.2	6.2	6.3	100
Logistic ssB	30	7.0	3.2	7.3	3.7	100
	50	7.2	3.2	7.0	4.1	100
	100	7.7	2.9	6.5	5.0	100
	200	8.2	2.5	6.0	5.5	100
Gaussian	30	6.8	5.5	7.3	4.8	100
	50	6.9	5.5	7.2	4.9	100
	100	7.5	5.4	7.3	5.4	100
	200	8.3	5.3	7.5	6.0	100
Mixtures	30	8.1	6.7	8.5	5.4	100
	50	8.6	6.4	8.5	5.8	100
	100	8.8	6.3	8.0	6.6	100
	200	9.2	6.0	8.0	6.9	100
RKHS	30	7.2	5.9	5.6	5.0	100
	50	7.5	6.2	5.5	5.2	100
	100	8.5	5.9	5.3	5.5	100
	200	9.5	5.6	5.5	6.0	100

cluding PLS) have a good overall performance since the Base approach is beaten in all sections. Although mRMR-RD outperforms the original versions of mRMR and Base, it is surpassed by the functional-oriented proposals (MHR and RK-VS). This is encouraging since the latter are constructed from a sound functional motivation. PLS is also outperformed by MHR and RK-VS, and behaves much like mRMR-RD on average. However, PLS is more unstable, obtaining very good results in some settings (e.g. with OUt) and being very close to Base in others (Gaussian and mixtures). In addition, let us recall that the use of PLS components entails a loss in interpretability with respect to variable selection.

Overall, the two preferable methods are MHR and RK-VS. The maxima hunting procedure seems to be more stable along the different models since it is completely model free. On the other hand, RK-VS is based on some assumptions which leads to the highest accuracy rates when the model fulfils these assumptions, and to a partial accuracy loss as we move away from them. Nevertheless, RK-VS is quite robust and even in the less favourable considered setting (mixtures) it is better than Base and PLS. Note also that RK-VS improves its relative results with bigger samples sizes while for the smallest sets it is often outperformed by MHR. This reveals some difficulties to estimate the pooled covariance matrix with very few observations. A solution could be to include some extra information in the model as in  $\text{RK-VS}_B$  (see Section 2.4).

Regarding the number of variables, MHR uses less features, followed by RK-VS, and finally mRMR-RD. PLS uses less components but they are usually hard to interpret. Curiously, MHR applied to logistic models with smoother processes (ssb) gets outperforms the other classifiers using less than a half of selected variables. Thus, in this kind of (rough) models the smoothing seems to be appropriate (specially when using MH). However, further research is needed for verifying these partial findings and drawing more general conclusions (remember that in other cases, e.g. Section 4.3, smoothing is counter-indicated).

Finally, a practical recommendation would be the use of RK-VS where the required assumptions are approximately fulfilled, and MHR when we are far from the RK-VS hypotheses or the samples are rather small.

## 5.3. Real data

### 5.3.1. Data sets

We have chosen three examples (on the basis of their popularity in the FDA literature) as well as an example of near perfect classification given in [Delaigle and Hall \(2012a\)](#). While these data sets have been already mentioned in previous chapters, we give here a broader description. We start with a summary of some basic features in [Table 5.3](#). Here, Phoneme stands for the smoothed version with the first 150 variables. The Base column refers to the accuracy level of the Base method defined above, i.e., the average success of a certain functional classifier. We think that this is a suitable reference value for further comparisons. In this case we have computed the base accuracy as the average of 100 independent runs with a nested (or stratified) 10-fold cross-validation (10CV); more details are given in the next subsection. In addition, [Figure 5.2](#) shows the trajectories  $X(t)$  and mean functions for each set and each class.

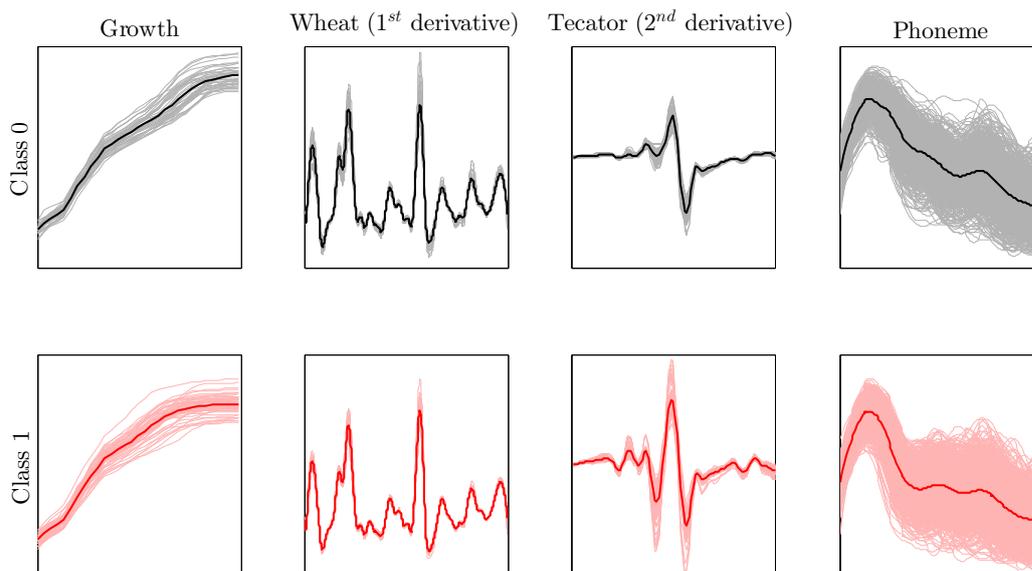


Figure 5.2: Data trajectories and mean functions from class 0 (first row) and class 1 (second row). Columns correspond to growth, Tecator and phoneme data from left to right.

**Growth** These are the popular growth data of the *Berkeley Growth Study* ([Tuddenham and Snyder, 1954](#)). These data have been thoroughly analysed in the monograph by [Ramsay and Silverman \(2005\)](#) and are available in the *fda* package of R. It contains the heights of 54 girls and 39 boys measured at 31 non-equally

Table 5.3: Description of the real datasets:  $n$  is the number of observations of dimension  $N$ ; “Base” represents the percentage of 10CV accuracy (over 100 independent runs) obtained with the complete curves using  $k$ -NN and linear SVM.

Dataset	$n$	$N$	Base $k$ -NN	Base SVM	References
Growth	93	31	96.16	89.74	Ramsay and Silverman (2005)
Wheat (1 <sup>st</sup> der.)	100	701	96.67	100.00	Delaigle and Hall (2012a)
Tecator (2 <sup>nd</sup> der.)	215	100	98.25	98.53	Ferraty and Vieu (2006)
Phoneme	1717	150	79.47	82.45	Hastie et al. (2009)

distant time points from 1 to 18 years; the height was measured every three months from 1 to 2 years, annually from 2 to 8, and twice a year from 8 to 18. This data set has been used in many classification studies, see e.g. Mosler and Mozharovskiy (2014) for a recent summary.

**Wheat** Wheat data correspond to 100 near-infrared spectra of wheat samples measured from 1100nm to 2500nm in 2nm interval (701 variables); an extensive description is given in Kalivas (1997). Following Delaigle and Hall (2012a) we have divided the data in two populations of 59 and 41 observations according to the protein content (more or less than 15 respectively). A particularity of NIR datasets is the high homogeneity among the raw data, which makes the classification problem harder. For this reason, these data are often used in a differentiated version, that is, they are smoothed (e.g., via splines) and then the first or the second derivative of the smoothed curves is used (see e.g. the monograph Ferraty and Vieu (2006)). In this case we have considered the derivative curves obtained with splines as in Delaigle and Hall (2012a). For this wheat data the near-perfect classification is achieved.

**Tecator** This is another well-known data set used many times as a benchmark for comparisons in FDA studies. It is available, for example, via the *fda.usc* R package. It consists of 215 near-infrared absorbance spectra of finely chopped meat, obtained using a Tecator Infratec Food & Feed Analyzer. Thus the final data set is made of 215 curves, observed at 100 equispaced points, ranging from 850 to 1050 nm with associated values of moisture, fats and protein contents. Following Ferraty and Vieu (2006), the sample is separated in two classes according to the fat content (smaller or larger than 20%). As in the previous example, Tecator data are highly homogeneous so we have use a differentiated version (via splines). We show here the results corresponding to the second derivatives (which turn out to provide a higher discrimination power than the raw data or the first derivative). A recent review of classification performances for different methods is given in

Galeano et al. (2014).

**Phoneme** These are data of speech recognition originally discussed in Hastie et al. (1995). They can be downloaded from [www-stat.stanford.edu/ElemStatLearn](http://www-stat.stanford.edu/ElemStatLearn) and are analyzed in Hastie et al. (2009) and Ferraty and Vieu (2006) among others. The original sample has 4509 curves, corresponding to log-periodograms constructed from 32 ms long recordings of males pronouncing five phonemes: “aa” as in “dark”, “ao” as in “water”, “sh” as in “she”, “iy” as in “she”, and “dcl” as in “dark”. Each curve was observed at 256 equispaced points. This five-classes discrimination problem is adapted to our binary setup by taking just (as in Delaigle and Hall (2012a)) the curves corresponding to the phonemes “aa” and “ao”. The sample size is  $n = 1717$  (695 from “aa” and 1022 from “ao”). Different versions of this dataset have been used in the literature varying the smoothing degree and the truncation point of the log-periodograms (Ferraty and Vieu, 2006; Delaigle and Hall, 2012a; Galeano et al., 2014). We have considered in this thesis the raw data (see Sections 3.4 and 4.2) and the truncated version (the first 50 variables) used in Delaigle and Hall (2012a) smoothed with a local linear method (see Subsection 2.5.3).

### 5.3.2. Methodology

Although we have attempted to follow a similar methodology to that in Subsection 2.5.1, dealing with real data entails some differences with respect to the simulations. These are a direct consequence of the limited (and often small) number of available observations. Unlike the case of the simulation studies, we cannot here generate new samples for validation and test, so the data at hand must be carefully re-used for obtaining these samples. There are several techniques for generating samples good enough for assessing the classification accuracy with a low risk of over-fitting. In this thesis we have considered the popular cross-validation methodology to tackle this problem, even though other approaches such as resampling procedures could also be used. Note, however that in Section 2.5 we have followed a resampling methodology aiming at making a fairer comparison with the results in Delaigle and Hall (2012a). Nevertheless, this cross-validation strategy is able to use more observations in the model estimation and its general performance is better.

Cross-validation (CV) is a well known validation model which is frequently used in practice in the variable selection and classification literature (Guyon et al., 2006; Hastie et al., 2009). It is based on averaging the evaluation measures (in our case the classification accuracy) over different partitions of the sample. These

partitions are defined in such a way that all observations are evaluated only once. We have considered two different variants: the leave-one-out cross-validation (LOOCV) and the  $k$ -fold cross-validation ( $k$ CV). The former relies on the evaluation of just one observation at each iteration so we have  $n - 1$  examples to estimate the model. LOOCV reduces the variance of the estimation but is much more time consuming than other approaches, so it is adequate for fairly small problems. On the other hand,  $k$ CV consists on randomly dividing the data in  $k$  groups of the same size. Then  $k - 1$  subsamples are used for training and the last one as test sample. The test sample is replaced at each iteration with a different (“untested”) one. When  $k$  is relatively small,  $k$ CV is affordable for big samples (note that  $k = n$  leads to LOOCV) and the variance can be reduced averaging over several  $k$ CV runs or increasing  $k$  (with the additional cost derived). We have used, in general,  $k = 10$ , which is a typical choice in practice. This produces training samples with  $9n/10$  observations and test samples of size  $n/10$ .

Finally, note that we need to generate both a validation and a test sample. Thus, we follow a nested (or stratified) CV strategy. First, training and test samples are produced in an usual CV iteration. Then, validation is carried out through another identical CV procedure over the training sample. Remember that the parameters involved in the validation stage are the number of variables (or components), the smoothing parameter  $h$  of MH, and those required for the classifiers.

In summary, the general methodology used in the real data study is as follows (see also the flowchart in Figure 5.3):

1. In each run of the real data experiments, a CV partition generates different pairs of training and test samples. The size of these subsamples depends on the CV model.
2. The relevant variables are selected using the training sample (alternatively the PLS-PCA projections are computed).
3. For each training sample an internal CV partition generates different pairs of training-b and validation samples. The parameters are fitted through these training-b and validation samples.
4. The data are reduced (i.e., the variables are selected or the PLS projections are recalculated) according to the result of the previous steps.
5. The “accuracy” outputs correspond to the percentages of correct classification obtained for the reduced test samples. In all cases the classifier is built from the reduced training sample.

6. Final outputs are the average over the CV partitions (they can additionally be averaged over several independent runs of the whole process).

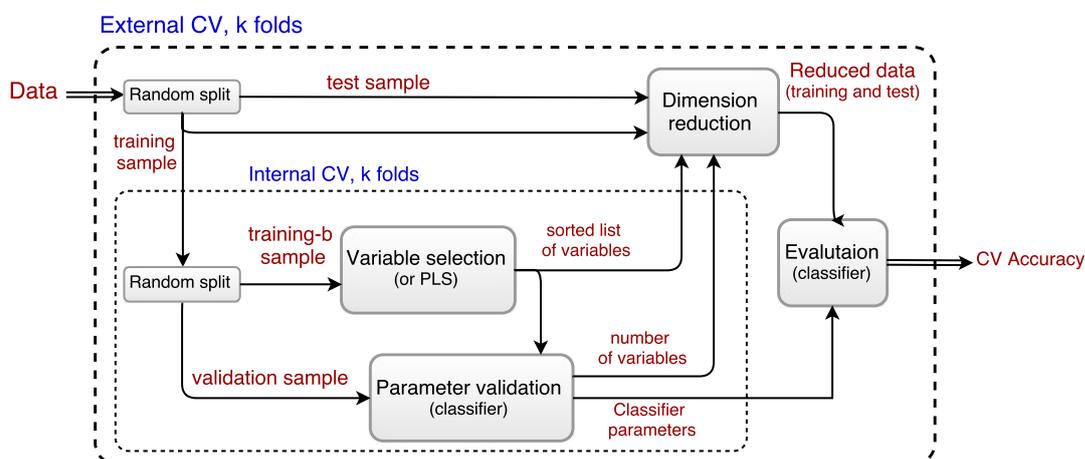


Figure 5.3: Methodology flowchart for real data.

### 5.3.3. Additional results

As in the previous Section, we show here some additional outputs aiming at presenting a comprehensive summary of the performance of our methods with the real datasets. Therefore, Table 5.4 shows the average classification accuracy of our new proposals (mRMR-RD, MHR and RK-VS) for all real datasets considered in this thesis (see Table 5.3). Together with our methods we have included PLS and the Base approach for comparative purposes. We have chosen the  $k$ -NN and the SVM (with linear kernel) classifiers because of their good performance. For the sake of clarity we have chosen the Phoneme version with the first 150 variables and smoothed (by splines) trajectories. This leads to better results than the other alternatives and it is perhaps the most used in the literature (see, e.g, Ferraty and Vieu (2006) or Galeano et al. (2014)). For illustrative purposes we have also included the Tecator, Wheat and Phoneme raw data.

The methods appear in columns and the datasets in rows. We have followed the methodology described above (with a nested 10CV) and the outputs are averaged over 100 independent runs in order to reduce the variability. The outputs in boldface correspond to the two best performances in each row. Values in parentheses stand for the standard deviation. Analogously, Table 5.5 shows the results for the number of selected variables.

Table 5.4: Percentage of classification accuracy (and standard deviation) for the real data with both classifiers.

<i>k</i> -NN outputs					
Data	mRMR-RD	MHR	RK-VS	PLS	Base
Growth	95.27 (7.04)	95.28 (6.78)	95.19 (7.23)	<b>95.96 (6.23)</b>	<b>96.16 (6.35)</b>
Wheat	81.99 (10.87)	81.57 (10.90)	<b>95.88 (6.31)</b>	<b>84.64 (10.33)</b>	83.65 (10.93)
Wheat (1 <sup>st</sup> der.)	<b>100.00 (0.00)</b>	<b>100.00 (0.00)</b>	<b>100.00 (0.00)</b>	99.37 (1.97)	92.05 (7.87)
Tecator	70.32 (9.03)	86.65 (7.19)	<b>90.83 (6.09)</b>	<b>86.80 (7.44)</b>	79.87 (8.22)
Tecator (2 <sup>nd</sup> der.)	<b>99.18 (2.00)</b>	<b>99.01 (2.25)</b>	98.21 (2.79)	97.49 (3.33)	98.25 (2.84)
Phoneme	80.50 (2.98)	79.36 (2.42)	<b>80.91 (2.83)</b>	<b>81.73 (2.62)</b>	79.27 (2.58)
Phoneme (smooth)	80.01 (2.81)	80.55 (2.89)	<b>81.51 (2.73)</b>	<b>81.29 (2.54)</b>	79.47 (2.61)

SVM outputs					
Data	mRMR-RD	MHR	RK-VS	PLS	Base
Growth	<b>95.31 (1.12)</b>	92.28 (1.46)	95.22 (1.25)	<b>95.61 (0.98)</b>	89.74 (1.42)
Wheat	82.23 (1.13)	98.63 (0.85)	<b>100.00</b>	99.44 (0.59)	<b>100.00</b>
Wheat (1 <sup>st</sup> der.)	<b>99.61 (0.49)</b>	<b>99.61 (0.49)</b>	99.57 (0.50)	99.52 (0.52)	<b>100.00 (.00)</b>
Tecator	97.53 (0.39)	96.19 (0.36)	<b>98.51 (0.29)</b>	97.44 (0.58)	<b>98.00 (0.66)</b>
Tecator (2 <sup>nd</sup> der.)	98.14 (0.42)	98.23 (0.34)	98.12 (0.19)	<b>99.37 (0.31)</b>	<b>98.53 (0.55)</b>
Phoneme	81.26 (0.34)	80.48 (0.38)	<b>81.98 (0.22)</b>	80.90 (0.26)	<b>81.94 (0.30)</b>
Phoneme (smooth)	81.89 (0.31)	81.52 (0.25)	<b>82.41 (0.20)</b>	82.30 (0.22)	<b>82.45 (0.24)</b>

Table 5.5: Average number of selected variables (and standard deviation) for the real data with both classifiers.

<i>k</i> -NN outputs					
Data	mRMR-RD	MHR	RK-VS	PLS	Base
Growth	3.36 (0.30)	3.79 (0.13)	<b>2.94 (0.36)</b>	<b>2.31 (0.21)</b>	31
Wheat	<b>1.91 (0.50)</b>	6.79 (0.80)	<b>2.09 (0.17)</b>	3.45 (0.21)	701
Wheat (1 <sup>st</sup> der.)	<b>1.00 (0.00)</b>	<b>1.00 (0.00)</b>	<b>1.00 (0.00)</b>	1.92 (0.07)	701
Tecator	<b>1.86 (0.58)</b>	3.00 (0.00)	<b>2.11 (0.16)</b>	4.08 (0.24)	100
Tecator (2 <sup>nd</sup> der.)	<b>1.11 (0.20)</b>	<b>1.18 (0.21)</b>	1.66 (0.60)	1.90 (0.25)	100
Phoneme	13.87 (1.31)	<b>10.49 (0.96)</b>	14.35 (0.88)	<b>9.29 (1.12)</b>	150
Phoneme (smooth)	9.82 (1.20)	<b>3.33 (0.41)</b>	<b>7.01 (0.74)</b>	8.41 (1.28)	150
SVM outputs					
Data	mRMR-RD	MHR	RK-VS	PLS	Base
Growth	2.83 (0.23)	3.42 (0.14)	<b>2.53 (0.32)</b>	<b>2.33 (0.22)</b>	31
Wheat	<b>2.59 (0.75)</b>	4.28 (0.22)	<b>2.00 (0.00)</b>	3.00 (0.01)	701
Wheat (1 <sup>st</sup> der.)	1.76 (0.16)	<b>1.71 (0.12)</b>	<b>1.73 (0.14)</b>	1.92 (0.06)	701
Tecator	8.55 (0.29)	<b>3.00 (0.00)</b>	<b>3.93 (0.68)</b>	6.17 (0.50)	100.00
Tecator (2 <sup>nd</sup> der.)	5.03 (0.78)	<b>1.51 (0.24)</b>	<b>1.47 (0.40)</b>	2.65 (0.34)	100
Phoneme	16.38 (0.79)	<b>13.68 (1.62)</b>	17.37 (0.86)	<b>10.60 (0.62)</b>	150
Phoneme (smooth)	16.74 (1.22)	<b>3.72 (0.15)</b>	<b>7.70 (0.89)</b>	9.72 (0.46)	150

The overall results are clearly positive for the variable selection methods. The slight losses in efficiency of some of them with respect to PLS are a small price to be paid for using a simpler dimension reduction methodology. The Phoneme data are by far the most complicated to handle, while the higher rates of accuracy are achieved with differentiated Wheat trajectories. RK-VS seems to have the better global performance though MHR uses less variables. Both classifiers exhibit a similar behaviour: SVM is the winner in Phoneme and  $k$ -NN in Growth.

Note that NIR data are very smooth and homogeneous, which entails some difficulties to classify the raw data. However, when we take derivatives our methods are able to achieve the near perfect classification using just one or two variables. On the other hand, smoothing Phoneme trajectories (which are extremely rough) leads, in general, to better classification results. In this case, it is quite remarkable the improvement of MHR in terms of both accuracy and number of variables.

Regarding the methodology, the nested 10CV appears as a suitable benchmark choice. It is easy to reproduce (the variability can be reduced averaging over independent runs), has a reasonable cost and produces reasonably good results in comparison with other methods in the literature (even with simple classifiers). Maybe LOOCV, (which uses more completely the available data and is fully replicable), could be another benchmark alternative, provided that it is computationally feasible (which is not the case for the Phoneme data). We have used this LOOCV approach in previous chapters for Growth and Tecator data with similar results to those of Table 5.4.

Finally, although in principle we were not primarily concerned with the best classification rate but with the best choice of variables, we can conclude that variable selection procedures combined with  $k$ -NN or a linear SVM, are competitive when compared with PLS and other successful and sophisticated methods in the literature: see Galeano et al. (2014) for Tecator, Mosler and Mozharovskyi (2014) for Growth and again Galeano et al. (2014) or Delaigle and Hall (2012a) for Phoneme.

# Chapter 6

## Conclusions

Functional data have grown in importance in the last decades thanks to their increasing presence in relevant areas and the technological improvements which allow for their processing. Throughout this thesis we have focused on the supervised classification problem with functional data, which have been studied using both standard and novel approaches.

Functional classification entails several challenges which are mostly due to the infinite dimension of the data spaces and the high collinearity between close variables. Most of this dissertation is devoted to tackle these problems by means of intrinsic variable selection techniques. As a major conclusion of our study we might say that these specific dimension reduction techniques are extremely useful, in terms of statistical efficiency. In addition, the use of variable selection procedures entails a gain in interpretability compared with other popular dimension reduction methods such as PCA and PLS, which provide not directly interpretable results in terms of the original variables. Last but not least, the intrinsic approach leads to significant time savings with respect to wrapper methodologies with apparently no loss in accuracy. In this vein, the variables selected according to our proposals are also independent of the classifier at hand.

From a practical point of view, we have proposed three intrinsic variable selection methods:

- **mRMR-RD.** It is a new version of mRMR, a popular and well-established variable selection method. Our proposal consists in replacing the original association measures (MI or FC) with the distance correlation measures proposed by [Székely et al. \(2007\)](#). We argue that this choice avoids the estimation problems related to the mutual information. Results in Section 4.2 show that this new version also leads to an improvement in accuracy with

respect to the original mRMR formulations. It is also feasible for functional data since mRMR-RD outperforms the Base approach (which uses the whole functions).

- **MHR.** The maxima-hunting methodology is purely functional-oriented and in fact, it is unfeasible in the multivariate setting. MHR relies on a direct use of the increasingly popular distance correlation measure  $\mathcal{R}^2$ . The simple idea of selecting the variables coinciding with the local maxima of  $\mathcal{R}^2(X_t, Y)$  has proven to be effective in all the considered examples. This good performance is partially due to the fact that (besides its simplicity) MHR intrinsically deals with the relevance-redundancy trade-off. The method is also backed by a uniform convergence result and some examples in which the variables involved in the optimal rule are those selected by MHR.
- **RK-VS.** This method appears as a direct consequence of looking at the functional classification problem from an RKHS point of view. In this context, under model (2.3) and a sparsity assumption, the optimal rules turn out to depend on a finite number of variables. These variables can be selected by maximizing the Mahalanobis distance between the finite-dimensional projections of the class means, which is a quite natural idea when dealing with Gaussian processes. Our RK-VS method is an iterative approximation to this. This is an easy-to-interpret and fast methodology which allows for easily adding extra information about the model. The empirical performance of RK-VS is extremely good when the considered problems fit the assumed model but it turns out to be also quite robust against partial departures from the hypotheses, typically leading to very good results in general problems.

All considered methods have been tested through the most extensive simulation study so far available (to our knowledge) in the literature. The experiments consisted of simulation models with different characteristics and several real datasets. In addition, the methods were also checked in a real metabolomic problem. The access to these data is a result of our collaboration with the Hospital Universitari Vall d'Hebron and Institut de Recerca (VHIR) in Barcelona, and especially with Ignasi Barba and David García-Dorado. While we have included only the outputs of mRMR-RD (as in the preliminary draft with [Barba et al. \(2015\)](#)), the methods MHR and RK-VS have been also tested with similar and even better results.

The overall performance of our new proposals has been quite satisfactory in the experiments. In general, our methods obtained high accuracy levels and out-

performed the Base approach with a few variables. This justifies the use of variable selection techniques in this framework since it not only improves the computational costs but the classification accuracy. Moreover, the new methods have shown to be competitive and often better than some standard alternatives. This is the case of PLS, the reference dimension reduction method in problems of this type, which is slightly beaten in almost all examples. In addition, we must also consider the gain in interpretability provided by the variable selection methodology, which is specially relevant in the real data applications.

Nevertheless, we cannot recommend a unique method for all situations. While the functional-oriented methods (MHR and RK-VS) have a better average performance than mRMR-RD, there are some exceptions. In fact, all the considered procedures involve the use of algorithms which suffer from different drawbacks. Thus, the MHR approach tends to fail when the maxima are hard to estimate. This typically happens for very smooth or very "wiggly" samples, or when many redundant maxima do appear. As for the RK-VS selection method, it is relatively weaker with very small sample sizes. Finally, mRMR suffers also from different shortcomings (although some of them are reduced with the use of *dcor*), ranging from the lack of a complete theoretical motivation to some other intrinsic problems (see [Brown et al. \(2012\)](#) and [Frénay et al. \(2013\)](#) for some additional details). In any case, almost any of this methods (which take into account the redundancy in different ways) is clearly preferable to the "ranking" procedures which essentially ignore the redundancy.

Regarding the comparison between the use of  $\mathcal{R}^2$  versus the unbounded  $\mathcal{V}^2$ , the first is in general preferable but still this is not a uniform pattern since  $\mathcal{V}^2$ -based methods are better in some cases. About the decision on whether or not to (moderately) smooth the data, it is in general advisable when the methods (especially MH) must be used with very rough data.

On the theoretical side, a major aim in this thesis was to contribute to the mathematical foundation of FDA as a statistical counterpart for the stochastic processes theory. So, in addition to our new proposals for variable selection, we have outlined a theoretical framework to motivate these proposals in population terms, that is, in terms of the underlying processes which generate the data. In this task, the Radon-Nikodym (RN) derivatives and the RKHS theory have been the basic tools. Thus, we have shown that the Radon-Nikodym derivatives can be used to provide explicit (not too complicate) expressions for the optimal rules in several important classification problems. These RN derivatives can be expressed (via an important theorem by [Parzen \(1961\)](#)) in terms of the RKHS space associated with the covariance operator of the underlying process. This suggests an RKHS-based

class of models for which variable selection is a natural aim. They are defined in terms of the sparsity assumption [SA]. In fact, these particular sparse models are “dense” within the more general model under study. As a consequence, RKHS appears as an appealing alternative to the classical  $L^2$  setup. In some sense, the RKHS approach amounts to a “change of coordinates” allowing for the presence of a sort of “Dirac deltas”, which are particularly useful to formalize sparse models depending just on a finite number of variables. Note that this cannot be easily made in the classical  $L^2$  setting, since the function  $\beta(t)$  in the linear model  $Y = \int_0^1 \beta(t)X(t)dt$  cannot be replaced by a linear combination of Dirac deltas (which do not belong to  $L^2$ ). However, as we have seen, this idea can be easily put in RKHS terms.

As a practical consequence of the RKHS approach, a method for variable selection (RK-VS) is defined. An associated classifier (RK-C) is also proposed. It might be seen as a functional-motivated version of the Fisher’s linear rule. It is fast and easy-to-interpret. It is asymptotically optimal under the general model (2.3) and the sparsity assumption [SA]. It is also quite flexible, in the sense that its performance can be improved by the inclusion of extra information about the model.

In addition, we have seen that the perfect classification of Gaussian processes (which cannot possibly happen in finite-dimensional problems) can appear if and only if the corresponding probability measures are mutually singular and otherwise (under equivalent distributions), the Bayes rule is linear. The perfect classification can be achieved through the optimal rules of a sequence of absolutely continuous problems which approximate the singular one.

## 6.1. Further work

These are, in our view, some interesting topics for future research:

*General problem of functional classification and near perfect classification:* Extension of our results to non Gaussian (e.g. mixtures) and non homoscedastic settings. It would be also desirable to derive further explicit expressions of Bayes rules in other models such as Ornstein-Uhlenbeck, fractional Brownian motion, etc.

*RKHS methodology:* This theory has a huge potential of applications in FDA. Some obvious fields (not necessarily involving variable selection) for future de-

velopment are the functional linear model with scalar response and the functional clustering methodology. Another natural goal is the study of new exploratory/visualization RKHS-motivated tools.

*Multiclass discrimination:* This is another obvious, extremely relevant, field for further study. In general, the extension from the binary to the multiple class setup is not trivial. However, the maxima hunting methodology seems promising in this regard, given the good properties of the distance covariance measure.

*How many variables to select?:* The automatic selection of the number of variables  $d$  is still an open problem. It would be desirable to find some feasible criterion independent of the classifier. Maxima-hunting procedures could represent a good starting point since they readily give a estimation of  $d$  (the number of local maxima). However, this approximation is too affected by noise and other features, hence some additional work is still needed.

*Variable selection targeted to other problems (different from supervised binary classification):* Again, the natural fields for development would be functional regression and clustering.

*Open problems in the maxima hunting methodology:* Here the goal would be to describe a wide class of explicit models where the relevant variables (from which the Bayes rule solely depends) correspond to the local maxima of the distance covariance (or distance correlation) function. We have so far a few examples but a broader collection of models would provide a better ground for the use of MH methods. On the other hand, we have detected (both theoretically and empirically) that although the local maxima capture the most relevant information, those variables corresponding to the non-differentiable points of  $\mathcal{R}^2(X_t, Y)$  are also important and should be taken into account. Finally, some maxima are redundant with each other, which suggests the use of some backward elimination procedure after the selection.

*Further applications of the distance correlation measure:* Distance correlation ( $dcor$ ) has proved to be a very useful tool for variable selection tasks. Indeed, it retains most good properties of the mutual information measure (which is by far the most used association measure in variable selection problems), together with some additional advantages, specially in estimation. It would be interesting to make further studies in this line, in order to build a comprehensive framework around  $dcor$ -based variable selection, similar to that outlined by Vergara

and Estévez (2014) around MI . This research includes, among other considerations, the use of *dcor* in different algorithms as well as theoretical developments (regarding, e.g., error bounds).

*Variable selection in "parametric" models:* We might also take advantage of the flexibility of RK-VS and RK-C by using different (parametric) models for different problems. This would lead to use "plug in" versions of the RK methodologies. From an algorithmic point o view, it would also be of some interest combining our proposals with some feasible wrapper methods in a two-stages algorithm.

*mRMR:* This extremely popular methodology (which has motivated thousands of citations in the machine learning community) is essentially based on a heuristically motivated algorithm, backed with a number of successful practical applications. In spite of some partial advances, a complete asymptotic theory (including consistency results for the identification of the relevant variables) is still lacking.

*Applications:* We plan to continue our work on the analysis of biomedical data for both improving functional discrimination rates and detecting relevant variables (gene, metabolites, etc). This would be a joint research with the team lead by Dr. David García-Dorado at Hospital Universitari Vall d'Hebron of Barcelona. In particular, we plan to use functional classification techniques (combined with variable selection) with serum spectra data from diabetic patients.

# Chapter 6

## Conclusiones

La importancia de los datos funcionales ha aumentado en las últimas décadas gracias a su creciente presencia en áreas relevantes y a los avances tecnológicos que hacen posible su procesamiento. A lo largo de esta tesis nos hemos centrado en el problema de clasificación supervisada con datos funcionales, estudiándolo desde distintos puntos de vista.

La clasificación funcional presenta algunas dificultades, debidas en su mayoría a la dimensión infinita de los espacios donde viven los datos y a la elevada colinealidad entre variables cercanas. La mayor parte de este trabajo se dedica a solventar estas dificultades mediante la utilización de métodos de selección de variables intrínsecos. Una primera conclusión de nuestro estudio es que estas técnicas de reducción de dimensión son extremadamente útiles en FDA (en términos de eficiencia estadística). Además, la selección de variables conlleva una ganancia en interpretabilidad cuando la comparamos con otros métodos populares de reducción de dimensión como PLC y PLS, cuyas proyecciones no son directamente interpretables en términos de las variables originales del problema. Por último pero no menos importante, la utilización de métodos intrínsecos supone un ahorro significativo en tiempo con respecto a las metodologías *wrapper*, sin acarrear (aparentemente) pérdidas en la precisión. Asimismo, la selección realizada por nuestras propuestas (intrínsecas) es independiente del clasificador que estemos usando.

Desde un punto de vista práctico, en esta tesis hemos propuesto tres métodos intrínsecos de selección de variables:

- **mRMR-RD**. Esta es una nueva versión de mRMR, un método de selección de variables contrastado y muy popular. Nuestra propuesta consiste en reemplazar la medida de asociación original (MI o FC) por la correlación de

distancias propuesta por Székely et al. (2007). Con este cambio se evitan los problemas de estimación inherentes a la información mutua. Los resultados de la Sección 4.2 muestran que la nueva versión obtiene mayores tasas de acierto que las formulaciones originales de mRMR. Además, mRMR-RD también supera el acierto *Base* (obtenido con las curvas completas), lo que sugiere que es una propuesta viable para datos funcionales.

- **MHR.** La metodología de la “caza de máximos” es genuinamente funcional, de hecho no puede usarse en el contexto multivariado. MHR se basa en una utilización directa de la cada vez más popular distancia de correlaciones  $\mathcal{R}^2$ . Pese a su sencillez, la idea de seleccionar las variables coincidiendo con los máximos locales de  $\mathcal{R}^2(X_t, Y)$  ha probado su efectividad en todos los ejemplos estudiados. Su buen funcionamiento se debe en parte a que MHR gestiona intrínsecamente el equilibrio entre relevancia y redundancia. El método está respaldado por un resultado de convergencia uniforme y una serie de ejemplos en los que las variables involucradas en la regla óptima son aquéllas seleccionadas por MHR.
- **RK-VS.** Este método es una consecuencia directa de observar el problema de clasificación funcional desde el punto de vista que ofrece el RKHS asociado. Así, bajo el modelo general (2.3) y una hipótesis sobre las funciones de medias de los procesos involucrados, la regla óptima resulta depender de un número finito de variables. Estas variables pueden seleccionarse al maximizar la distancia de Mahalanobis entre las proyecciones finito dimensionales de las medias de las clases (una idea bastante natural cuando se trabaja con procesos Gaussianos). Nuestro método RK-VS es una aproximación iterativa a esta estrategia. El resultado es un algoritmo rápido y fácilmente interpretable que permite añadir información extra sobre el modelo de manera sencilla. En la práctica, los resultados de RK-VS son extremadamente buenos cuando los problemas verifican las hipótesis necesarias para que se cumplan los resultados mencionados, pero el método ha resultado ser también bastante robusto ante desviaciones parciales de las hipótesis, obteniendo muy buenos resultados en problemas generales.

Todos estos métodos han sido puestos a prueba mediante el estudio de simulación más extenso (hasta donde sabemos) disponible en la literatura. Los experimentos han incluido modelos de simulación con distintas características y algunos conjuntos de datos reales. Además, los métodos también han sido probados en un problema metabólico real. El acceso a estos datos ha sido fruto de una colaboración con el Hospital Universitari Vall d’Hebron e Institut de Recerca (VHIR) de Barcelona, y especialmente con Ignasi Barba y David García-Dorado. Aunque en esta tesis sólo hemos incluido los resultados correspondientes a mRMR-RD

(al igual que en el manuscrito con Barba et al. (2015)), MHR y RK-VS también fueron probados en este problema con resultados similares o mejores.

El rendimiento global de nuestras nuevas propuestas ha sido muy satisfactorio en todos los experimentos realizados. En general, nuestros métodos han alcanzado altas tasas de acierto en la clasificación y han superado al método *Base* con unas pocas variables. Esto justifica plenamente la utilización de técnicas de selección de variables en este contexto, ya que no sólo mejoran los costes computacionales sino también el acierto. Además, estos nuevos métodos se han demostrado competitivos, y a menudo mejores, que otras alternativas previamente conocidas. Este es el caso de PLS, el método de reducción de dimensión de referencia en este tipo de problemas, que es ligeramente superado en casi todos los ejemplos. Asimismo, también debe ser tenida en cuenta la ganancia en interpretabilidad que aporta la selección de variables, y que es especialmente relevante en aplicaciones con datos reales.

Con todo, no podemos recomendar un único método para todas las situaciones. En general, los métodos con una orientación funcional (MHR y RK-VS) tienen un mejor rendimiento medio que mRMR-RD, pero hay algunas excepciones. De hecho, todos los métodos utilizados tienen algunos puntos débiles. Así, MHR tiene algunos problemas cuando los máximos son redundantes entre sí o difíciles de estimar; lo que ocurre típicamente cuando los datos tienen trayectorias muy suaves o muy abruptas. Por su parte, RK-VS es algo más débil ante tamaños muestrales muy pequeños. Finalmente, aunque algunos de los problemas de mRMR se solventan con la utilización de *dcor*, el método sigue padeciendo algunas limitaciones que van desde la ausencia de una motivación teórica completa hasta deficiencias intrínsecas (ver Brown et al. (2012) y Frénay et al. (2013) para detalles adicionales). En cualquier caso, todos estos métodos (que tienen en cuenta la redundancia de distintas maneras) son claramente preferibles a los que usan estrategias “ranking” que esencialmente ignoran la redundancia.

En la comparación entre  $\mathcal{R}^2$  y  $\mathcal{V}^2$ , vemos que la primera es, en general, preferible. Sin embargo, este no es un criterio uniforme ya que hay casos en que los métodos basados en  $\mathcal{V}^2$  obtienen mejores resultados que los que usan la versión normalizada. En cuanto al suavizado, parece recomendable una suavización moderada al trabajar con datos muy abruptos, si bien no es beneficioso en todos los casos ni para todos los métodos (es especialmente recomendable con MH).

Desde el punto de vista teórico, uno de los principales objetivos de esta tesis ha sido contribuir al fundamento matemático de FDA estableciendo vínculos con la teoría de procesos estocásticos. En este sentido, además de las nuevas propues-

tas para selección de variables, también hemos esbozado un marco teórico que motiva estas propuestas en términos poblacionales, es decir, en términos de los procesos subyacentes que generan los datos. Las herramientas básicas para ello han sido las derivadas de Radon-Nikodym (RN) y la teoría RKHS. En la tesis hemos visto que las derivadas RN nos permiten obtener expresiones explícitas (no demasiado complicadas) para las reglas óptimas de clasificación en algunos problemas importantes. Estas derivadas pueden expresarse (usando un teorema de Parzen (1961)) en términos del espacio RKHS asociado al operador de covarianzas de los procesos subyacentes. El resultado es una clase de modelos basados en el enfoque RKHS en los que la selección de variables es un objetivo natural. Estos modelos, definidos mediante la hipótesis [SA], son “densos” en el modelo general estudiado. Como consecuencia, el espacio RKHS se presenta como una alternativa interesante a la configuración clásica basada en el espacio  $L^2$ . Y es que, en cierto sentido, el enfoque RKHS origina un “cambio de coordenadas” donde aparecen un tipo de “deltas de Dirac”, lo que es especialmente útil para formalizar modelos dispersos que dependan de un número finito de variables. Esto no puede conseguirse en un marco  $L^2$  tradicional de manera sencilla ya que la función  $\beta(t)$  del modelo lineal  $Y = \int_0^1 \beta(t)X(t)dt$  no puede reemplazarse por una combinación lineal de deltas de Dirac (que no pertenecen a  $L^2$ ). Sin embargo, como hemos visto, esta idea sí puede llevarse a cabo fácilmente al poner el problema en términos del RKHS.

Una consecuencia práctica del enfoque basado en RKHS es la definición del método de selección de variables RK-VS. También se ha propuesto un clasificador asociado (RK-C) que puede verse como una versión de la regla lineal de Fisher con una motivación funcional. Se trata de un clasificador rápido y fácil de interpretar, que es asintóticamente óptimo bajo el modelo general (2.3) y la hipótesis [SA]. Además es un método flexible, en el sentido de que puede mejorar su rendimiento con la inclusión de información adicional sobre el modelo.

Asimismo, también hemos visto que la clasificación perfecta entre procesos Gaussianos (fenómeno imposible en problemas de dimensión finita) se da si y sólo si las respectivas medidas de probabilidad son mutuamente singulares. En caso contrario (cuando las distribuciones son equivalentes), la regla Bayes es lineal. La clasificación perfecta puede obtenerse mediante las reglas óptimas de una sucesión de problemas absolutamente continuos que aproximen el problema singular.

## 6.1. Trabajo futuro

En nuestra opinión, esto son algunos temas interesantes para futuras investigaciones:

*Problema de clasificación funcional general y clasificación casi perfecta:* Extender nuestros resultados a modelos heterocedásticos o no Gaussianos (por ejemplo, mixturas). También sería deseable la obtención de expresiones explícitas para la regla Bayes en otros casos: Ornstein-Uhlenbeck, movimiento Browniano fraccionario, etc.

*Metodología RKHS:* La teoría RKHS presenta un gran potencial para desarrollar aplicaciones en FDA. Las primeras áreas en las que continuar con esta línea de investigación (no necesariamente alrededor de la selección de variables) parecen los modelos lineales funcionales con respuesta escalar y el clustering con datos funcionales. El estudio de nuevas herramientas de exploración/visualización con una motivación RKHS sería otro objetivo natural.

*Clasificación multiclase:* Este problema, muy relevante en la práctica, es un claro objetivo para futuras investigaciones. En general, la extensión del caso binario al multiclase no es trivial. Sin embargo, gracias a las buenas propiedades de la distancia de covarianzas, la metodología basada en la caza de máximos parece prometedora a este respecto.

*¿Cuántas variables seleccionar?:* La selección automática del número de variables  $d$  es un problema abierto. Sería deseable encontrar algún criterio viable e independiente del clasificador. Los métodos de caza de máximos representan un buen punto de partida al dar de manera natural una estimación de  $d$  (el número de máximos locales). Sin embargo, este valor se ve demasiado afectado por el ruido y otros aspectos por lo que aún es necesario continuar la investigación.

*Nuevos enfoques para la selección de variables (distintos de la clasificación supervisada binaria):* De nuevo, las primeras alternativas naturales parecen la regresión funcional y el clustering.

*Problemas abiertos en la caza de máximos:* Aquí el objetivo sería describir una amplia clase de modelos en los que las variables relevantes (aquellas que aparecen en la regla Bayes) coincidan con los máximos locales de  $\mathcal{V}^2(X_t, Y)$  (o  $\mathcal{R}^2(X_t, Y)$ ). Actualmente disponemos de algunos ejemplos, pero una colección más extensa proporcionaría una base más sólida para la utilización de métodos

MH. Por otra parte, hemos detectado (tanto teórica como empíricamente) que, aunque los máximos locales capturan la mayor parte de la información relevante, aquellas variables asociadas con los puntos no diferenciables de  $\mathcal{V}^2(X_t, Y)$  (o  $\mathcal{R}^2(X_t, Y)$ ) también son importantes y deberían tenerse en cuenta. Finalmente, algunos máximos son redundantes entre sí, esto sugiere el uso de alguna técnica de eliminación *backward* tras la primera selección.

*Nuevas aplicaciones de la distancia de correlaciones:* La distancia de correlaciones (*dcor*) ha demostrado ser una herramienta muy valiosa en lo referente a selección de variables. De hecho, *dcor* mantiene la mayoría de las buenas propiedades de la información mutua (la medida de asociación más utilizada en problemas de selección de variables) y presenta algunas ventajas adicionales, especialmente en la estimación. Sería interesante profundizar en esta línea para construir un marco general alrededor de *dcor* similar al existente en torno a MI (Vergara and Estévez, 2014). Este trabajo incluiría, entre otros aspectos, la utilización de *dcor* en distintos algoritmos y desarrollos teóricos (por ejemplo, en relación con cotas de error).

*Selección de variables en modelos "paramétricos":* También se podría sacar ventaja de la flexibilidad de RK-VS y RK-C mediante el uso de diferentes modelos (paramétricos) según los problemas. Esto conllevaría la utilización de versiones "plug in" de los métodos RK. Desde el punto de vista algorítmico, también sería interesante combinar nuestras propuestas con métodos *wrapper* adecuados en algoritmos de dos etapas.

*mRMR:* Pese a ser extremadamente popular (ha motivado miles de citas en la comunidad de *machine learning*), esta metodología está esencialmente sustentada en un algoritmo de motivación heurística con un buen rendimiento en la práctica. A pesar de tímidos avances parciales, todavía no se ha obtenido una teoría asintótica completa para mRMR (incluyendo resultados de consistencia para la identificación de las variables relevantes).

*Aplicaciones:* Tenemos previsto continuar el análisis de datos biomédicos funcionales buscando tanto la mejora del acierto en clasificación como la detección de variables relevantes (genes, metabolitos, etc.). Se trata de una investigación conjunta con el equipo liderado por el doctor David García-Dorado en el Hospital Universitari Vall d'Hebron of Barcelona. En concreto, ya estamos trabajando en la aplicación de técnicas de clasificación funcional (combinadas con selección de variables) a datos espectrales obtenidos del suero de pacientes diabéticos.

# Appendix A

## Simulation models

We now list all the models included in the simulation study. The relevant variables are indicated in brackets (for Gaussian and mixture models) or in the expression of  $\psi(X)$  (for the logistic-type models). Variables in bold face had found to be specially relevant in terms of their influence in the error rate.

### 1. Gaussian models:

1. **G1**:  $\begin{cases} P_0 : & B(t) \\ P_1 : & B(t) + \theta t \end{cases}, \theta \sim N(0, 3)$   
*variables* =  $\{X_{100}\}$ .
2. **G1b**:  $\begin{cases} P_0 : & B(t) \\ P_1 : & B(t) + \theta t \end{cases}, \theta \sim N(0, 5)$   
*variables* =  $\{X_{100}\}$ .
3. **G2**:  $\begin{cases} P_0 : & B(t) + t \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_{100}\}$ .
4. **G2b**:  $\begin{cases} P_0 : & B(t) + 3t \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_{100}\}$ .
5. **G3**:  $\begin{cases} P_0 : & BB(t) \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_{100}\}$ .
6. **G4**:  $\begin{cases} P_0 : & B(t) + \mathit{hillside}_{0.5,4}(t) \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_{47}, \mathbf{X}_{100}\}$ .
7. **G5**:  $\begin{cases} P_0 : & B(t) + 3\Phi_{1,1}(t) \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_1, \mathbf{X}_{48}, X_{100}\}$ .
8. **G6**:  $\begin{cases} P_0 : & B(t) + 5\Phi_{2,2}(t) \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_{48}, \mathbf{X}_{75}, X_{100}\}$ .
9. **G7**:  $\begin{cases} P_0 : & B(t) + 5\Phi_{3,2}(t) + 5\Phi_{3,4}(t) \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_{22}, \mathbf{X}_{35}, X_{49}, X_{74}, \mathbf{X}_{88}, X_{100}\}$ .
10. **G8**:  $\begin{cases} P_0 : & B(t) + 3\Phi_{2,1.25}(t) + 3\Phi_{2,2}(t) \\ P_1 : & B(t) \end{cases}$   
*variables* =  $\{X_9, \mathbf{X}_{35}, X_{48}, X_{62}, \mathbf{X}_{75}, X_{100}\}$ .

**2. Logistic models:** These are the  $\psi$  functions used to define the models,

$$\mathbf{L1:} \psi(X) = 10X_{65}.$$

$$\mathbf{L2:} \psi(X) = 10X_{30} + 10X_{70}.$$

$$\mathbf{L3:} \psi(X) = 10X_{30} - 10X_{70}.$$

$$\mathbf{L4:} \psi(X) = 20X_{30} + 50X_{50}20X_{80}.$$

$$\mathbf{L5:} \psi(X) = 20X_{30} - 50X_{50} + 20X_{80}.$$

$$\mathbf{L6:} \psi(X) = 10X_{10} + 30X_{40} + 10X_{72} + 10X_{80} + 20X_{95}.$$

$$\mathbf{L7:} \psi(X) = \sum_{i=1}^{10} 10X_{10i}.$$

$$\mathbf{L8:} \psi(X) = 20X_{30}^2 + 10X_{50}^4 + 50X_{80}^3.$$

$$\mathbf{L9:} \psi(X) = 10X_{10} + 10|X_{50}| + 0X_{30}^2 X_{85}.$$

$$\mathbf{L10:} \psi(X) = 20X_{33} + 20|X_{68}|.$$

$$\mathbf{L11:} \psi(X) = \frac{20}{X_{35}} + \frac{30}{X_{77}}.$$

$$\mathbf{L12:} \psi(X) = \log X_{35} + \log X_{77}.$$

$$\mathbf{L13:} \psi(X) = 40X_{20} + 30X_{28} + 20X_{62} + 10X_{67}.$$

$$\mathbf{L14:} \psi(X) = 40X_{20} + 30X_{28} - 20X_{62} - 10X_{67}.$$

$$\mathbf{L15:} \psi(X) = 40X_{20} - 30X_{28} + 20X_{62} - 10X_{67}.$$

The variations included are,

$$\mathbf{L3b:} \psi(X) = 30X_{30} - 20X_{70}.$$

$$\mathbf{L4b:} \psi(X) = 30X_{30} + 20X_{50} + 10X_{80}.$$

$$\mathbf{L5b:} \psi(X) = 10X_{30} - 10X_{50} + 10X_{80}.$$

$$\mathbf{L6b:} \psi(X) = 20X_{10} + 20X_{40} + 20X_{72} + 20X_{80} + 20X_{95}.$$

$$\mathbf{L8b:} \psi(X) = 10X_{30}^2 + 10X_{50}^4 + 10X_{80}^3.$$

### 3. Mixture models:

$$1. \mathbf{M1} : \begin{cases} P_0 : \begin{cases} B(t) + 3t & , 1/2 \\ B(t) - 2t & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_{100}\}.$

$$2. \mathbf{M2} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{2,2}(t) & , 1/2 \\ B(t) + 5\Phi_{3,2}(t) & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_{22}, X_{35}, X_{48}, X_{75}, X_{100}\}.$

$$3. \mathbf{M3} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{2,2}(t) & , 1/10 \\ B(t) + 5\Phi_{3,2}(t) & , 9/10 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_{22}, X_{35}, X_{48}, X_{75}, X_{100}\}.$

$$4. \mathbf{M4} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{2,2}(t) & , 1/2 \\ B(t) + 5\Phi_{3,3}(t) & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_{48}, X_{62}, X_{75}, X_{100}\}.$

$$5. \mathbf{M5} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{2,1}(t) & , 1/3 \\ B(t) + 3\Phi_{2,2}(t) & , 1/3 \\ B(t) + 5\Phi_{3,2}(t) & , 1/3 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_1, X_{22}, X_{35}, X_{48}, X_{75}, X_{100}\}.$

$$6. \mathbf{M6} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{2,1}(t) & , 1/2 \\ B(t) + 3t & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_1, X_{22}, X_{49}, X_{100}\}.$

$$7. \mathbf{M7} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{1,1}(t) & , 1/2 \\ BB(t) & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_1, X_{48}, X_{100}\}.$

$$8. \mathbf{M8} : \begin{cases} P_0 : \begin{cases} B(t) + \theta t, \theta \sim N(0, 5) & , 1/2 \\ B(t) + hillside_{0.5,5}(t) & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_{47}, X_{100}\}.$

$$9. \mathbf{M9} : \begin{cases} P_0 : \begin{cases} B(t) + \theta t, \theta \sim N(0, 5) & , 1/2 \\ BB(t) & , 1/2 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = X_{100}.$

$$10. \mathbf{M10} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{1,1}(t) & , 1/3 \\ B(t) - 3t & , 1/3 \\ BB(t) & , 1/3 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_1, X_{48}, X_{100}\}.$

$$11. \mathbf{M11} : \begin{cases} P_0 : \begin{cases} B(t) + 3\Phi_{1,1}(t) & , 1/4 \\ B(t) - 3t & , 1/4 \\ B(t) + hillside_{0.5,5}(t) & , 1/4 \\ BB(t) & , 1/4 \end{cases} \\ P_1 : B(t) \end{cases}$$

$variables = \{X_1, X_{48}, X_{100}\}.$

Finally, the full list of models involved is, in summary, as follows:

1. L1 OU	26. L5 OU	51. L9 sB	76. L15 OU
2. L1 OUt	27. L5b OU	52. L9 ssB	77. L15 OUt
3. L1 B	28. L5 OUt	53. L10 OU	78. L15 B
4. L1 sB	29. L5 B	54. L10 B	79. L15 sB
5. L1 ssB	30. L5 sB	55. L10 sB	80. G1
6. L2 OU	31. L5 ssB	56. L10 ssB	81. G1b
7. L2 OUt	32. L6 OU	57. L11 OU	82. G2
8. L2 B	33. L6b OU	58. L11 OUt	83. G2b
9. L2 sB	34. L6 OUt	59. L11 B	84. G3
10. L2 ssB	35. L6b OUt	60. L11 sB	85. G4
11. L3 OU	36. L6 B	61. L11 ssB	86. G5
12. L3b OU	37. L6 sB	62. L12 OU	87. G6
13. L3 OUt	38. L6 ssB	63. L12 OUt	88. G7
14. L3b OUt	39. L7 OU	64. L12 B	89. G8
15. L3 B	40. L7b OU	65. L12 sB	90. M1
16. L3b B	41. L7 OUt	66. L12 ssB	91. M2
17. L3 sB	42. L7b OUt	67. L13 OU	92. M3
18. L3 ssB	43. L7 B	68. L13 OUt	93. M4
19. L4 OU	44. L7 sB	69. L13 B	94. M5
20. L4b OU	45. L7 ssB	70. L13 sB	95. M6
21. L4 OUt	46. L8 B	71. L13 ssB	96. M7
22. L4b OUt	47. L8 sB	72. L14 OU	97. M8
23. L4 B	48. L8 ssB	73. L14 OUt	98. M9
24. L4 sB	49. L8b OU	74. L14 B	99. M10
25. L4 ssB	50. L9 B	75. L14 sB	100. M11

# Bibliography

- Abraham, C., G. Biau, and B. Cadre (2006). On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics* 58(3), 619–633.
- Amaldi, E. and V. Kann (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science* 209(1), 237–260.
- Aneiros, G. and P. Vieu (2014). Variable selection in infinite-dimensional problems. *Statistics & Probability Letters* 94, 12–20.
- Aneiros-Pérez, G. and P. Vieu (2006). Semi-functional partial linear regression. *Statistics & Probability Letters* 76(11), 1102–1110.
- Antoniadis, A., X. Brossat, J. Cugliari, and J.-M. Poggi (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing* 11(01), 1350003.
- Arauzo-Azofra, A., J. L. Aznarte, and J. M. Benítez (2011). Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems with Applications* 38(7), 8170–8177.
- Ash, R. B. and M. F. Gardner (2014). *Topics in Stochastic Processes: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Academic Press.
- Athreya, K. B. and S. N. Lahiri (2006). *Measure Theory and Probability Theory*. Springer.
- Audibert, J.-Y. and A. B. Tsybakov (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics* 35(2), 608–633.
- Baíllo, A., A. Cuevas, and J. A. Cuesta-Albertos (2011). Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics* 38(3), 480–498.
- Baíllo, A., A. Cuevas, and R. Fraiman (2011). *Classification methods for functional data*, pp. 259–297. In [Ferraty and Romain \(2011\)](#).

- Barba, I., E. Miró-Casas, E. Pladevall, R. Sebastián, J. R. Berrendero, J. Torrecilla, A. Cuevas, and D. García-Dorado (2015). High fat diet induces metabolic changes associated to increased oxidative stress in male hearts. *Draft*.
- Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions on* 5(4), 537–550.
- Beniwal, S. and J. Arora (2012). Classification and feature selection techniques in data mining. *International Journal of Engineering Research & Technology (IJERT)* 1(6).
- Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.
- Berrendero, J. R., A. Cuevas, and J. L. Torrecilla (2015a). On near perfect classification and functional Fisher rules via reproducing kernels. *arXiv:1507.04398*, submitted.
- Berrendero, J. R., A. Cuevas, and J. L. Torrecilla (2015b). The mRMR variable selection method: a comparative study for functional data. *Journal of Statistical Computation and Simulation* (to appear).
- Berrendero, J. R., A. Cuevas, and J. L. Torrecilla (2015c). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica* (to appear).
- Biau, G., B. Cadre, and Q. Paris (2015). Cox process functional learning. *Stat. Inference Stoch. Process.* 18(3), 257–277.
- Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 1705–1732.
- Billingsley, P. (2013). *Convergence of Probability Measures*. John Wiley & Sons.
- Bosq, D. and D. Blanke (2008). *Inference and Prediction in Large Dimensions*. John Wiley & Sons.
- Breiman, L. (2001). Random forests. *Machine Learning* 45(1), 5–32.
- Brown, G., A. Pocock, M.-J. Zhao, and M. Luján (2012). Conditional likelihood maximization: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research* 13(1), 27–66.
- Brusco, M. J. (2014). A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis* 77, 38–53.
- Burfield, R., C. Neumann, and C. P. Saunders (2015). Review and application of functional data analysis to chemical data—the example of the comparison, classification and database search of forensic ink chromatograms. *Chemometrics and Intelligent Laboratory Systems*, to appear.

- Cadre, B. (2013). Supervised classification of diffusion paths. *Math. Methods Statist.* 22(3), 213–225.
- Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics* 35(6), 2313–2351.
- Cao, R., A. Cuevas, and W. G. Manteiga (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis* 17(2), 153–176.
- Carrizosa, E., B. Martín-Barragán, and D. R. Morales (2011). Detecting relevant variables and interactions in supervised classification. *European Journal of Operational Research* 213(1), 260–269.
- Carroll, R. J., A. Delaigle, and P. Hall (2013). Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier. *The Annals of Applied Statistics* 41(6), 2739–2767.
- Cérou, F. and A. Guyader (2006). Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics* 10, 340–355.
- Comminges, L. and A. S. Dalalyan (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics* 40(5), 2667–2696.
- Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine Learning* 20(3), 273–297.
- Cucker, F. and D. X. Zhou (2007). *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press.
- Cuesta-Albertos, J. A., M. Febrero-Bande, and M. O. de la Fuente (2015). The DD G-classifier in the functional setting. *arXiv:1501.00372*.
- Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference* 147, 1–23.
- Cuevas, A., M. Febrero, and R. Fraiman (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* 22(3), 481–496.
- De Boor, C. (1978). *A Practical Guide to Splines*. Springer.
- Delaigle, A. and P. Hall (2012a). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society B* 74(2), 267–286.
- Delaigle, A. and P. Hall (2012b). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics* 40(1), 322–352.

- Delaigle, A., P. Hall, and N. Bathia (2012). Componentwise classification and clustering of functional data. *Biometrika* 99(2), 299–313.
- Delsol, L., F. Ferraty, and A. Martinez Calvo (2011). *Functional Data Analysis: An Interdisciplinary Statistical Topic*. In Goldfarb et al. (2011).
- Demler, O. V., M. J. Pencina, and R. B. D’Agostino (2013). Impact of correlation on predictive ability of biomarkers. *Statistics in medicine* 32(24), 4196–4210.
- Devroye, L., L. Györfi, and G. Lugosi (2013). *A Probabilistic Theory of Pattern Recognition*, Volume 31. Springer Science & Business Media.
- Díaz-Uriarte, R. and S. Alvarez de Andrés (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3.
- Ding, C. and H. Peng (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* 3(2), 185–205.
- Doob, J. L. (1953). *Stochastic Processes*. Wiley.
- DsGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.
- Duda, R. O., P. E. Hart, and D. G. Stork (2012). *Pattern Classification*. John Wiley & Sons.
- Dudoit, S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association* 97(457), 77–87.
- Dueck, J., D. Edelman, T. Gneiting, and D. Richards (2014). The affinity invariant distance correlation. *Bernoulli* 20(4), 2305–2330.
- El Akadi, A., A. Amine, A. El Ouardighi, and D. Aboutajdine (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge and Information Systems* 26(3), 487–500.
- Estévez, P., M. Tesmer, C. Perez, and J. M. Zurada (2009). Normalized mutual information feature selection. *Neural Networks, IEEE Transactions on* 20(2), 189–201.
- Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* 70(5), 849–911.
- Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20(1), 101.
- Febrero-Bande, M. and M. Oviedo de la Fuente (2012). Statistical computing in functional data analysis: the R package *fda.usc*. *Journal of Statistical Software* 51(4), 1–28.

- Feldman, J. (1958). Equivalence and perpendicularity of Gaussian processes. *Pacific J. Math* 8(4), 699–708.
- Ferraty, F., P. Hall, and P. Vieu (2010). Most-predictive design points for functional data predictors. *Biometrika* 97(4), 807–824.
- Ferraty, F. and Y. Romain (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press.
- Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2), 179–188.
- Fraiman, R., Y. Giménez, and M. Svarc (2015). Feature selection for functional data. *arXiv:1502.02123*.
- Frénay, B., G. Doquire, and M. Verleysen (2013). Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. *Neurocomputing* 112, 64–78.
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* 84(405), 165–175.
- Galeano, P., E. Joseph, and R. E. Lillo (2014). The Mahalanobis distance for functional data with applications to classification. *Technometrics*, to appear.
- Gao, Y.-F., B.-Q. Li, Y.-D. Cai, K.-Y. Feng, Z.-D. Li, and Y. Jiang (2013). Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection. *Molecular BioSystems* 9(1), 61–69.
- Gertheiss, J., A. Maity, and A.-M. Staicu (2013). Variable selection in generalized functional linear models. *Stat* 2(1), 86–101.
- Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorical explanatory variables. *The Annals of Applied Statistics*, 2150–2180.
- Goldfarb, B., C. Pardoux, M. Summa, and M. Touati (2011). *Statistical Learning and Data Science*. Chapman & Hall.
- Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439), 531–537.
- Gómez-Verdejo, V., M. Verleysen, and J. Fleury (2009). Information-theoretic feature selection for functional data classification. *Neurocomputing* 72(16), 3580–3589.

- Gönen, M. A. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.* 12, 2211–2268.
- González-Manteiga, W. and P. Vieu (2011). *Methodological richness of functional data analysis*, pp. 197–203. In Goldfarb et al. (2011).
- Graves, S., G. Hooker, and J. Ramsay (2009). *Functional Data Analysis with R and MATLAB*. Springer.
- Grosenick, L., S. Greer, and B. Knutson (2008). Interpretable classifiers for fMRI improve prediction of purchases. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 16(6), 539–548.
- Guyon, I., S. Gunn, M. Nikravesh, and L. A. Zadeh (2006). *Feature Extraction: Foundations and Applications*. Springer.
- Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3), 389–422.
- Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Ph. D. thesis, The University of Waikato.
- Hall, P. and H. Miller (2011). Determining and depicting relationships among components in high-dimensional variable selection. *Journal of Computational and Graphical Statistics* 20(4), 988–1006.
- Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science* 21(1), 1–14.
- Hastie, T., A. Buja, and R. Tibshirani (1995). Penalized discriminant analysis. *The Annals of Statistics* 23(1), 73–102.
- Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin (2009). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27(2), 83–85.
- Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer.
- Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley.
- Hsing, T. and H. Ren (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *The Annals of Statistics* 37(2), 726–755.
- Hsu, H.-H., C.-W. Hsieh, and M.-D. Lu (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications* 38(7), 8144–8150.

- Hua, J., W. D. Tembe, and E. R. Dougherty (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition* 42(3), 409–424.
- James, G. M. and T. J. Hastie (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B*, 533–550.
- James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that’s interpretable. *The Annals of Statistics* 37(5A), 2083–2108.
- Janson, S. (1997). *Gaussian Hilbert Spaces*. Cambridge University Press.
- Jiang, B. and J. S. Liu (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics* 42(5), 1751–1786.
- Kailath, T. (1971). RKHS approach to detection and estimation problems I: Deterministic signals in Gaussian noise. *IEEE Transactions on Information Theory* 17(5), 530–549.
- Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems* 37(2), 255–259.
- Karabulut, E. M., S. A. Özel, and T. Ibrikçi (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology* 1, 323–327.
- Kittler, J. (1978). *Feature set search algorithms*. Sijhoff and Noordhof.
- Kneip, A. and P. Sarda (2011). Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics* 39(5), 2410–2447.
- Kohavi, R. and G. H. John (1997). Wrappers for feature subset selection. *Artificial intelligence* 97(1), 273–324.
- Kong, J., S. Wang, and G. Wahba (2015). Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine* 34(10), 1708–1720.
- Kwak, N. and C.-H. Choi (2002). Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(12), 1667–1671.
- Laha, R. G. and V. K. Rohatgi (1979). *Probability Theory*. Wiley.
- Lazar, C., J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. De Schaetzen, R. Duque, H. Bersini, and A. Nowe (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9(4), 1106–1119.

- Leardi, R., R. Boggia, and M. Terrile (1992). Genetic algorithms as a strategy for feature selection. *Journal of chemometrics* 6(5), 267–281.
- Lee, E. R. and B. U. Park (2012). Sparse estimation in functional linear regression. *Journal of Multivariate Analysis* 105(1), 1–17.
- Li, B. and Q. Yu (2008). Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis* 52(10), 4790–4800.
- Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* 107(499), 1129–1139.
- Li, Y., N. Wang, and R. J. Carroll (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association* 108(504), 1284–1294.
- Lindquist, M. A. and I. W. McKeague (2009). Logistic regression with Brownian-like predictors. *Journal of the American Statistical Association* 104(488), 1575–1585.
- Liptser, R. and A. N. Shiryaev (2013). *Statistics of Random Processes: I. General Theory*. Springer.
- Liu, H. and H. Motoda (2012). *Feature Selection for Knowledge Discovery and Data Mining*. Springer.
- López-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* 104(486), 718–734.
- Lukić, M. and J. Beder (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society* 353(10), 3945–3969.
- Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability* 41(5), 3284–3305.
- Maldonado, S. and R. Weber (2009). A wrapper method for feature selection using support vector machines. *Information Sciences* 179(13), 2208–2217.
- Mandal, M. and A. Mukhopadhyay (2015). A novel PSO-based graph-theoretic approach for identifying most relevant and non-redundant gene markers from gene expression data. *International Journal of Parallel, Emergent and Distributed Systems* 30(3), 175–192.
- Martin-Barragan, B., R. Lillo, and J. Romo (2014). Interpretable support vector machines for functional data. *European Journal of Operational Research* 232(1), 146–155.
- Matsui, H. (2014). Variable and boundary selection for functional data via multiclass logistic regression modeling. *Computational Statistics & Data Analysis* 78, 176–185.

- McKeague, I. W. and B. Sen (2010). Fractals with point impact in functional linear regression. *Annals of Statistics* 38(4), 2559.
- Mörters, P. and Y. Peres (2010). *Brownian Motion*. Cambridge University Press.
- Mosler, K. and P. Mozharovskyi (2014). Fast DD-classification of functional data. *arXiv:1403.1158*.
- Mundra, P. and J. C. Rajapakse (2010). SVM-RFE with MRMR filter for gene selection. *NanoBioscience, IEEE Transactions on* 9(1), 31–37.
- Nguyen, D. V. and D. M. Rocke (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* 18(1), 39–50.
- Nguyen, X. V., J. Chan, S. Romano, and J. Bailey (2014). Effective global approaches for mutual information based feature selection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 512–521. ACM.
- Parzen, E. (1961). An Approach to Time Series Analysis. *The Annals of Mathematical Statistics* 32(4), 951–989.
- Parzen, E. (1962). Extraction and detection problems and reproducing kernel Hilbert spaces. *Journal of the Society for Industrial & Applied Mathematics, Series A: Control* 1(1), 35–62.
- Peng, H., F. Long, and C. Ding (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27(8), 1226–1238.
- Pigoli, D. and L. M. Sangalli (2012). Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives. *Computational Statistics & Data Analysis* 56(6), 1482–1498.
- Ponsa, D. and A. López (2007). Feature selection based on a new formulation of the minimal-redundancy-maximal-relevance criterion. In *Pattern Recognition and Image Analysis*, Marti, J. et al. eds., pp. 47–54.
- Preda, C., G. Saporta, and C. Lévéder (2007). PLS classification of functional data. *Computational Statistics* 22(2), 223–235.
- Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of artificial intelligence research* 4, 77–90.
- Ramsay, J. (1982). When the data are functions. *Psychometrika* 47(4), 379–396.
- Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer.

- Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer.
- Rényi, A. (1959). On measures of dependence. *Acta Mathematica Hungarica* 10(3-4), 441–451.
- Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti (2011). Detecting novel associations in large data sets. *Science* 334(6062), 1518–1524.
- Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *The Journal of Machine Learning Research* 3, 1371–1382.
- Robnik-Šikonja, M. and I. Kononenko (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning* 53(1-2), 23–69.
- Rosasco, L., S. Villa, S. Mosci, M. Santoro, and A. Verri (2013). Nonparametric sparsity and regularization. *The Journal of Machine Learning Research* 14(1), 1665–1714.
- Ryali, S., K. Supekar, D. A. Abrams, and V. Menon (2010). Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage* 51(2), 752–764.
- Saeyns, Y., I. Inza, and P. Larrañaga (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507–2517.
- Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. *The Annals of Statistics* 43(4), 1716–1741.
- Segall, A. and T. Kailath (1975). Radon-Nikodym derivatives with respect to measures induced by discontinuous independent-increment processes. *The Annals of Probability* 3(3), 449–464.
- Sen, P. K. (1977). Almost sure convergence of generalized U-statistics. *The Annals of Probability* 5(2), 287–290.
- Seth, S. and J. C. Principe (2010). Variable selection: A statistical dependence perspective. In *Ninth International Conference on Machine Learning and Applications (ICMLA), 2010*, pp. 931–936. IEEE.
- Shepp, L. (1966). Radon-Nikodym Derivatives of Gaussian Measures. 37(2), 321–354.
- Speed, T. (2011). A correlation for the 21st century. *Science* 334(6062), 1502–1503.
- Székely, G. J. and M. L. Rizzo (2009). Brownian distance covariance. *The Annals of Applied Statistics* 3(4), 1236–1265.
- Székely, G. J. and M. L. Rizzo (2012). On the uniqueness of distance covariance. *Statistics & Probability Letters* 82(12), 2278–2282.

- Székely, G. J. and M. L. Rizzo (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* 117, 193–213.
- Székely, G. J. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* 143(8), 1249–1272.
- Szekely, G. J. and M. L. Rizzo (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* 42(6), 2382–2412.
- Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6), 2769–2794.
- Tian, T. S. and G. M. James (2013). Interpretable dimension reduction for classifying functional data. *Computational Statistics & Data Analysis* 57(1), 282–296.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* 58(1), 267–288.
- Tuddenham, R. and M. Snyder (1954). Physical growth of California boys and girls from birth to eighteen years. *Publ. Child. Dev. Univ. Calif.* 1(2), 183–364.
- Ullah, S. and C. F. Finch (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology* 13, 43.
- Varberg, D. (1961). On equivalence of Gaussian measures. *Pacific Journal of Mathematics* 11(2), 751–762.
- Varberg, D. (1964). On Gaussian measures equivalent to Wiener measure. *Transactions of the American Mathematical Society* 113, 262–273.
- Vergara, J. R. and P. A. Estévez (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications* 24(1), 175–186.
- Walters-Williams, J. and Y. Li (2009). Estimation of mutual information: A survey. In *Rough Sets and Knowledge Technology*, pp. 389–396. Springer.
- Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing*. CRC Press.
- Wang, J.-L., J.-M. Chiou, and H.-G. Mueller (2015). Review of Functional Data Analysis. *arXiv:1507.05135*.
- Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association* (to appear).
- Xiaobo, Z., Z. Jiewen, M. J. Povey, M. Holmes, and M. Hanpin (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta* 667(1), 14–32.

- Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100(470), 577–590.
- Yenigün, C. D. and M. L. Rizzo (2015). Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation* 85(8), 1692–1705.
- Yu, L. and H. Liu (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research* 5, 1205–1224.
- Zhang, X., B. U. Park, and J.-I. Wang (2013). Time-varying additive models for longitudinal data. *Journal of the American Statistical Association* 108(503), 983–998.
- Zhang, Y., C. Ding, and T. Li (2008). Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics* 9(Suppl 2), S27.
- Zhao, Y., H. Chen, and R. T. Ogden (2014). Wavelet-based weighted LASSO and screening approaches in functional linear regression. *Journal of Computational and Graphical Statistics* (to appear).
- Zhong, W. and L. Zhu (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation* 85(11), 2331–2345.
- Zhou, J., N.-Y. Wang, and N. Wang (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica* 23(1), 25–50.