Universidad Autónoma de Madrid Departamento de Matemáticas

Tesis Doctoral

On reproducing kernel methods in functional statistics

Beatriz Bueno Larraz

Madrid, 2018

Directores: José Ramón Berrendero Díaz
Antonio Cuevas González

A mis abuelos

Agradecimientos

Aunque parezca mentira, hace ya diez años que "volé" de Huesca con destino Madrid. Estudiar en la UAM fue la primera decisión importante que tomé en mi vida, y de la que no podría estar más satisfecha. Quiero mostrar mi agradecimiento a muchas de las personas con las que me he cruzado durante este tiempo, por su ayuda, apoyo y compañía.

En primer lugar quiero agradecer a mis directores, José Ramón y Antonio, porque esta tesis es tan vuestra como mía. Muchas gracias por todo lo que me habéis enseñado (no solamente en lo académico), por vuestro apoyo y dedicación, por vuestra confianza y por vuestra paciencia con mis "pescaos". En definitiva, por darme la oportunidad de trabajar con vosotros y mostrarme lo bonita que es la estadística, habéis sido unos "padres" estupendos.

Gracias a José Luis, por ejercer de hermano mayor, por los cafés interminables y por escuchar siempre mis quejas sin rechistar. Y al resto del grupo de estadística de la UAM, especialmente a Javi y a Amparo. Muchas gracias también al resto de profesores que me han ayudado a llegar hasta aquí. Quiero mencionar especialmente a dos de ellos. A Alberto Suárez, por ser el primero en acercarme a la investigación y animarme a seguir estudiando, primero el máster y después el doctorado. Y a Alfredo Vila, mi profesor del instituto, por convencerme de estudiar Matemáticas y ser el primero en hacerme ver su elegancia y magnitud.

Estos años hubiesen sido mucho más difíciles sin los doctorandos del departamento, que han hecho que todos los días estuviese encantada de ir hasta la universidad. Quisiera empezar por los que han pasado por el despacho en algún momento, pero la primera tiene que ser obligatoriamente Raquel: por tu sonrisa, tus maldades, por dejarme sobrepasar siempre el límite de abrazos, por las confesiones de sofá, las conversaciones estúpidas, las bolitas de papel... Sin duda vas a ser lo que más voy a echar de menos de estos cuatro años y te deseo todo lo mejor. A Dani (además de por toda tu ayuda en lo académico) por ser el primero en acogerme en el sótano y alegrarnos los días con tus locuras. El despacho no volvió a ser lo mismo desde que te fuiste. A Felipe, por tu entusiasmo, por todos los viajes y por inculcarnos a las nuevas generaciones el "amor" por el 103. A Ale, por tu alegría, tu cariño, por poner algo de orden de vez en cuando, ¡y por aguantarnos! A Julio, la incorporación más reciente, por no salir corriendo, por las ganas que le pones siempre a todo y por toda tu ayuda en las tesis. Y a Bea, por

ayudarme tanto con la docencia y por muchísimas más cosas, pero sobre todo por los abrazos.

Muchas gracias también a los que estaban "por arriba". A los que ya se han marchado: Marcos, Irina, Adri, David, Carlos, Javi, María, Jose, Iason, Leyter, Martí... Y a los que aún siguen por aquí: Marta, Jaime, Nikita, Álex, Fran, Diego, Adri, Flo, Manuel, José Luis, Sonja... También, aunque sean forasteros, quiero mencionar a los "jóvenes funcionales", especialmente a Javi y a Paula, por hacer que esperase siempre con impaciencia el próximo congreso.

I would like to thank Professor Claudia Klüppelberg for hosting me in her Chair at TUM. Thank you to Johannes, who made it possible for me to go to Munich in the first place and for making it feel like home. And thank you to the other PhD students there, specially to Tyson, Nadine and Thiago. Vielen Dank für eure Unterstützung!

Tampoco me puedo olvidar de la gente que me ha acompañado estos años fuera de la universidad. En primer lugar, muchísimas gracias a mis padres y a mi hermana, por animarme siempre a volar alto y confiar en mí, por hacer esto posible desde el principio y por todo vuestro apoyo, aunque no siempre entendieseis mis quejas. Estoy muy agradecida también a mis abuelos, aunque no hayan podido ver esta tesis terminada, por inculcarme desde pequeñita que "el saber es más importante que el tener". Seguro que estarían encantados de ver hasta qué punto he seguido su consejo. Y muchas gracias a Chema por estar a mi lado durante tantos años, porque has "sufrido" esta tesis casi tanto como yo, por todo tu apoyo, por alegrarme todos los días al volver a casa (incluso los no tan buenos) y por entusiasmarte por mi trabajo a veces incluso más que yo misma.

Muchas gracias a Marta por acompañarme prácticamente desde el primer día que pisé Madrid, y por todo lo que has hecho por mí desde entonces (que sería imposible de resumir en un par de frases). También gracias a los de Huesca, los de siempre, por hacerme sentir que esa sigue siendo mi casa. Gracias especialmente a Paula y Sheila por estar siempre ahí, independientemente de los años que pasen. Y a Raúl, por aventurarse conmigo a dejar el nido hace ya una década.

Por último, quiero dar las gracias al Departamento de Matemáticas de la UAM, al programa FPI-MINECO y a los proyectos MTM2013-44045-P y MTM2016-78751- P, que me han permitido llevar a cabo este trabajo.

Beginnings are usually scary, and endings are usually sad,
but it is everything in between that makes it all worth living.
Bob Marley

Contents

Abstract

Throughout this thesis we delve into some functional data problems from a novel mathematical point of view, which both clarifies the existing techniques and helps us to develop new ideas based on a purely functional approach. These new tools are often simpler and more efficient than the mere extensions of multivariate techniques, since they are designed focusing on the particular nature of the data. The guiding thread of this document is the use of Reproducing Kernel Hilbert Spaces (RKHS's). These spaces turn out to be specially useful to establish well-founded connections between functional problems and their multivariate counterparts. In particular, we focus on the following statistical problems:

- the definition of a suitable functional extension of the classical Mahalanobis distance,
- functional linear regression with both scalar and functional response, with an application to functional time series forecasting,
- and functional logistic regression.

Several simulations and experiments have been developed for all the methods proposed throughout the thesis. We use the statistical programming language R and all the code can be provided on demand.

RKHS's are the common thread of this work, so Chapter 1 is mainly devoted to them. These spaces have proven to be very useful in different areas of mathematics, not only statistics. Therefore, they have been defined from rather different perspectives. After a brief introduction to functional data, we compile diverse definitions of RKHS's and analyze the connections between them. Further on we present some of the numerous applications of these spaces, that we find specially appealing. At the end of this same chapter we summarize the main mathematical contributions of this work.

In Chapter 2 we suggest a suitable functional extension of the Mahalanobis distance, a classical tool in multivariate analysis. As it is well-known, the main advantage of the multivariate Mahalanobis distance when compared to the Euclidean metric is the fact that it takes into account the covariance structure of the data. Mahalanobis distance is basically a weighted version of the Euclidean one. It is defined in terms of the inverse of the covariance matrix. The functional counterpart of the covariance matrix would be

the covariance operator. Thus, the obvious difficulty for a direct functional extension of the notion of Mahalanobis distance is the non-invertibility of the covariance operator in infinite-dimensional cases.

Our definition of the functional Mahalanobis distance is suggested and motivated in terms of the RKHS associated with the stochastic process that generates the data. We first notice that the original (finite-dimensional) Mahalanobis distance between two realizations x_i and x_j of a random vector coincides with the norm in the RKHS (associated with the covariance matrix of the data) of $x_i - x_j$. However, this definition can not be directly extended to the functional case since the realizations of a stochastic process do not belong to the RKHS generated by its covariance operator. To circumvent this problem we suggest to replace each trajectory $x_i \equiv x_i(s)$ of the process with a "smoothed version" belonging to the RKHS. Such smoothed version is obtained as the solution of a classical minimization problem in statistics, penalizing the norm in the RKHS of the function. As mentioned in that chapter, this minimization has a simple explicit solution given by the Representer Theorem. Thus, we finally propose to define the functional Mahalanobis distance between two trajectories as the distance, in the RKHS norm, between the corresponding smoothed versions of these trajectories.

We show that the proposed distance is a true metric. Also, it depends only on a unique real smoothing parameter (the amount of penalization in the minimization problem) which is fully motivated in RKHS terms. Moreover, it shares some properties of its finite dimensional counterpart: it is invariant under isometries, it can be consistently estimated from the data and its sampling distribution is known under Gaussian models. The paper Berrendero et al. (2018b) corresponds essentially to the contents of Chapter 2.

The other main topic of this thesis is functional regression. In Chapter 3 we focus on variable selection for linear regression problems with functional predictors. By "variable selection" we mean a procedure to replace the whole trajectories of the functional explanatory variables with their values at a finite number of carefully selected instants (or "impact points"). The major advantage of variable selection in comparison with other dimension reduction techniques (like principal components analysis or partial least squares) is that it keeps more in touch with the original curves, in the sense that the results are directly interpretable in terms of the original data, which is usually desired in real data problems.

In the first part of the chapter we consider a functional regression model with scalar response. The basic idea of our approach is to use the RKHS associated with the underlying process, instead of the most common $L^2[0,1]$ space, in the definition of the model. This turns out to be especially suitable for variable selection purposes, since the finite-dimensional linear model based on the selected "impact points" can be seen as a particular case of the RKHS-based linear functional model. This is not feasible with the standard L^2 approach, since this space lacks of continuous evaluation functionals of type $x \mapsto x(t_0)$. In this RKHS framework, we address the consistent estimation

of the optimal design of impact points. We consider also the practical problem of deciding how many variables one should retain and we give a consistent estimator for this quantity. This first part of the chapter corresponds to the contents of the paper Berrendero et al. (2018a).

The second part of Chapter 3, which can be found in the paper Bueno-Larraz and Klepsch (2018), contains an extension of the previous methodology to functional regression with functional response. After that, the model is applied to forecasting in stationary functional time series. We propose to model the dependence of the data with a non-standard autoregressive (AR) structure, motivated in terms of the RKHS generated by the covariance kernel of the data. The stationarity of the time series entails that the RKHS does not change with time. The solution of the standard autoregressive model in $L^2[0,1]$ would involve the inverse of the covariance operator, which is not defined. However, with this new RKHS-based model the problem of the non-invertibility of the covariance operator can be circumvented.

The proposed model is fully functional and it is proved to be part of the more general family of Banach space valued autoregressive (ARB) processes. This allows us to use some results previously derived for these processes. In particular, we immediately get conditions for the existence of a unique stationary solution. Most theoretical results proved in the first part of Chapter 3 are adapted in the second part to the setting of functional response. Besides, we ensure the uniform convergence of the estimated response curves.

Finally, in Chapter 4 we address the problem of functional logistic regression. The most common approach in the literature is to directly extend the multiple logistic model, replacing the inner product in $\mathbb{R}^d$ with the inner product in $L^2[0,1]$. In contrast, we propose to use the inner product of the RKHS associated with the process. It is a well-known fact that the Gaussian homoscedastic model for binary classification in $\mathbb{R}^d$ entails the logistic model. In the functional setting we prove that, whenever the mean functions of both classes belong to the RKHS, one obtains our RKHS-based logistic model. If the mean functions belong instead to the image of the covariance operator (which is a subspace of the corresponding RKHS), one gets the classical L^2 model. In this regard, our RKHS model can be seen as a generalization of the L^2 one. In addition, as in Chapter 3, this RKHS approach is specially suitable to perform variable selection on the curves (in the sense that the finite-dimensional logistic model based on a set of evaluations of the process is a particular case).

The natural idea is to use the maximum likelihood (ML) procedure to estimate the slope function, but this method presents some difficulties, which are an important part of our discussion. Although we prove that the expected maximum likelihood function has a single maximum in the RKHS, the ML estimator might not be defined under rather general conditions. For multiple logistic regression this problem only arises for linearly separable samples, but it is drastically worsened for functional data. We analyze two different scenarios of non-existence:

- For a family of stochastic processes, including the Brownian motion, the ML estimator does not exist for any given sample with probability one. The classical L^2 model suffers from this problem too.
- For Gaussian processes the ML estimator does not exist with increasing probability.

We easily circumvent the second problem in practice in view of our variable selection aim, since we keep only a small number of variables. In order to deal with the first problem, we propose to use Firth's estimator, which is based on a re-sampling of the responses that avoids linear separability of the classes. We propose two different implementations of our methodology: one based on a greedy iterative maximization and a sequential one that goes over all the points of the grid. Our proposals have been tested for functional binary classification problems.

Resumen

A lo largo de esta tesis profundizamos en algunos problemas con datos funcionales desde un nuevo punto de vista matemático, que clarifica las técnicas existentes y al mismo tiempo nos ayuda a desarrollar nuevas ideas basadas en un enfoque puramente funcional. Estas nuevas herramientas suelen ser más simples y eficientes que las meras extensiones de técnicas multivariantes, ya que están específicamente diseñadas teniendo en cuenta la naturaleza de los datos. El hilo conductor de este trabajo es el uso de los espacios de Hilbert con núcleo reproductor (RKHS's según sus siglas en inglés). Estos espacios resultan de gran utilidad para establecer conexiones plenamente fundadas entre problemas funcionales y sus respectivos problemas multivariantes. En particular, nos centramos en los siguientes problemas estadísticos:

- la definición de una extensión adecuada de la distancia de Mahalanobis clásica,
- regresión lineal funcional, con respuesta tanto escalar como funcional, aplicada a la predicción de series temporales funcionales,
- y regresión logística funcional.

Se han llevado a cabo simulaciones y experimentos para todos los métodos propuestos en esta tesis. El lenguaje de programación utilizado es R y todo el código desarrollado se puede proporcionar bajo demanda. Los RKHS's son el denominador común de este trabajo, de forma que el capítulo 1 está básicamente dedicado a ellos. Estos espacios han probado ser de gran utilidad en diferentes áreas de las matemáticas, no sólo en estadística. Por lo tanto, han sido definidos desde varios puntos de vista diferentes. Tras una breve introducción a los datos funcionales, recogemos diversas definiciones de los RKHS's y analizamos las relaciones entre ellas. A continuación presentamos algunas de las numerosas aplicaciones de estos espacios, las cuales nos parecen especialmente interesantes. Al final de dicho capítulo resumimos las principales contribuciones de esta tesis.

En el capítulo 2 sugerimos una posible extensión funcional de la distancia de Mahalanobis, una herramienta clásica del análisis multivariante. Como es bien sabido, la principal ventaja de la distancia de Mahalanobis multivariante en comparación con la distancia Euclídea es que la primera tiene en cuenta la estructura de covarianza de los datos. La distancia de Mahalanobis es básicamente una versión ponderada de la distancia Euclídea, definida en función de la inversa de la matriz de covarianza. El

homólogo funcional de la matriz de covarianza es el operador de covarianza. Por tanto, la dificultad obvia a la hora de extender de forma directa la distancia de Mahalanobis es la no-invertibilidad del operador de covarianza en espacios funcionales.

Nuestra definición de la distancia de Mahalanobis está sugerida y motivada en términos del RKHS asociado con el proceso estocástico que genera los datos. En primer lugar observamos que la distancia de Mahalanobis clásica (en dimensión finita) entre dos realizaciones x_i y x_j de un vector aleatorio coincide con la norma en el RKHS (asociado a la matriz de covarianza de los datos) de $x_i - x_j$. Sin embargo, esta definición no puede ser extendida directamente al caso funcional, ya que las realizaciones de un proceso estocástico no pertenecen al RKHS generado por su operador de covarianza. Para solventar este problema sugerimos reemplazar cada trayectoria $x_i \equiv x_i(s)$ del proceso por una "versión suavizada" que sí pertenezca al RKHS. Dicha versión suavizada se obtiene como solución de un problema de minimización clásico en estadística, penalizando la norma en el RKHS de la función. Como se menciona en el capítulo, esta minimización tiene una solución explícita relativamente sencilla, que viene dada por el "Representer Theorem". Por lo tanto, proponemos definir la distancia de Mahalanobis funcional entre dos trayectorias como la distancia, en la norma del RKHS, entre sus correspondientes versiones suavizadas.

Demostramos que la distancia propuesta es realmente una métrica. Además, depende de un único parámetro real (el nivel de penalización elegido para la minimización) que está plenamente motivado en términos de los RKHS's. Asimismo, la distancia propuesta comparte algunas propiedades con su homóloga en dimensión finita: es invariante bajo isometrías, puede ser estimada consistentemente a partir de los datos y su distribución muestral es conocida para modelos Gaussianos. El artículo Berrendero et al. (2018b) se corresponde esencialmente con los contenidos del capítulo 2.

El otro tema principal de esta tesis es la regresión funcional. En el capítulo 3 nos centramos en la selección de variables para problemas de regresión lineal con predictores funcionales. Con "selección de variables" nos referimos a un procedimiento mediante el cual las trayectorias completas de las variables explicativas funcionales son remplazadas por sus valores en un número finito de instantes cuidadosamente elegidos (o "puntos de impacto"). La principal ventaja de seleccionar variables frente a otras técnicas de reducción de dimensión (como el análisis de componentes principales o la regresión de mínimos cuadrados parciales) es que está más en contacto con los datos originales, lo que suele ser deseable en problemas reales.

En la primera parte del capítulo consideramos el problema de regresión funcional con respuesta escalar. La idea básica de nuestro enfoque es el uso del RKHS asociado al proceso subyacente, en lugar del espacio $L^2[0,1]$ comúnmente usado, en la definición del modelo. Este nuevo enfoque resulta especialmente útil para la selección de variables, ya que el modelo lineal finito-dimensional basado en los "puntos de impacto" seleccionados puede verse como un caso particular del modelo funcional basado en RKHS's. Esto no puede darse con la definición L^2 estándar, ya que ese espacio carece de funcionales de

evaluación continuos del tipo $x \mapsto x(t_0)$. En el contexto RKHS, abordamos el problema de la estimación consistente del diseño óptimo basado en los puntos de impacto. Consideramos a su vez el problema práctico de decidir cuántas variables se deben seleccionar y damos un estimador consistente para dicha cantidad. Esta primera parte del capítulo se corresponde con los contenidos del artículo Berrendero et al. (2018a).

La segunda parte del capítulo 3, que puede encontrarse en el artículo Bueno-Larraz and Klepsch (2018), contiene la extensión de la metodología anterior al problema de regresión funcional con respuesta funcional. A continuación el modelo se aplica a la predicción de series temporales funcionales. Nosotros proponemos modelizar la dependencia de las curvas mediante una estructura autorregresiva (AR) no estándar, motivada en términos del RKHS generado por el núcleo de covarianza de los datos. La estacionaridad de la serie temporal implica que el RKHS utilizado no varíe con el tiempo. Obtener la solución del modelo autorregresivo estándar en $L^2[0,1]$ requeriría invertir el operador de covarianza, cuya inversa no está definida. Sin embargo, este nuevo modelo basado en RKHS nos permite solucionar el problema de la no-invertibilidad del operador de covarianza.

El modelo propuesto es completamente funcional y se prueba que forma parte de la familia de procesos autorregresivos en espacios de Banach (ARB). Esto nos permite usar resultados previos derivados para dichos procesos. En particular, se obtienen condiciones para asegurar la existencia de una única solución estacionaria. La mayoría de los resultados teóricos obtenidos en la primera mitad del capítulo 3 son adaptados en la segunda al caso de respuesta funcional. En este caso probamos además la convergencia uniforme de las curvas estimadas.

Finalmente, en el capítulo 4 abordamos el problema de regresión logística funcional. El enfoque más común en la literatura es extender directamente el modelo logístico múltiple, reemplazando el producto escalar en $\mathbb{R}^d$ por el producto interno en $L^2[0,1]$. En cambio, nosotros proponemos usar el producto escalar del RKHS asociado al proceso. Es bien conocido que el modelo de clasificación binaria homocedástica en $\mathbb{R}^d$ implica el modelo logístico. En el contexto funcional probamos que, siempre que las funciones de medias de ambas clases pertenezcan al RKHS, se obtiene nuestro modelo logístico basado en el RKHS. Si en cambio las funciones de medias están en la imagen del operador de covarianza (que es un subespacio del correspondiente RKHS), se obtiene el modelo L^2 clásico. En este sentido, nuestro modelo RKHS puede verse como una generalización del modelo L^2. Además, como en el capítulo 3, este enfoque RKHS es especialmente útil para seleccionar variables en las curvas (en el sentido de que el modelo logístico finito-dimensional basado en un conjunto de evaluaciones del proceso es un caso particular).

La idea natural es usar el procedimiento de máxima verosimilitud (MV) para estimar la función de pendientes de regresión ("slope function"), pero este método presenta algunas dificultades, que son una parte importante de nuestra discusión. Aunque probamos que la función de verosimilitud esperada tiene un único máximo en el RKHS, el

estimador de máxima verosimilitud (EMV) podría no estar definido bajo condiciones bastante generales. En el modelo de regresión logística múltiple este problema surge únicamente para muestras linealmente separables, pero empeora drásticamente para datos funcionales. En concreto, analizamos dos escenarios distintos de no-existencia:

- Para una familia de procesos estocásticos, incluyendo el movimiento Browniano, el EMV no existe con probabilidad uno para cualquier muestra dada. El modelo L^2 clásico también sufre este problema.
- Para procesos Gaussianos el EMV no existe con probabilidad creciente.

El segundo problema se puede evitar fácilmente en vista de nuestro objetivo de seleccionar variables, ya que nos quedamos únicamente con un pequeño número de ellas. Con el fin de resolver el primer problema, proponemos usar el estimador de Firth, basado en un remuestreo de las respuestas para evitar la separación lineal de las clases. Proponemos dos implementaciones distintas de esta metodología: una basada en un algoritmo de optimización parcial (o "greedy") y otra secuencial que revisa todos los puntos de la malla. Ambas propuestas han sido probadas en problemas de clasificación binaria funcional.

Some notation

We introduce here the common notation used in this document. Some specific notation for each topic is introduced in the corresponding chapter.

Given a measurable space (S, Σ, μ), the $L^2(S)$ space is defined as the set of measurable functions f from S to $\mathbb{R}$ such that $\int_S |f|^2 \mathrm{d}\mu$ is finite. Throughout this work we mainly use $S = [0, 1]$ with the Lebesgue measure, or the probability space $(\Omega, \mathcal{A}, P)$. Sometimes we restrict ourselves to $C(S)$, the space of continuous functions over S endowed with the uniform norm. Given a Hilbert (or Banach) space H, the space of bounded linear operators from H to H is denoted as $\mathcal{L}$. As usual, the norm of $F \in \mathcal{L}$ is defined as the supremum value of $\|F(f)\|_H$ for $f \in H$ such that $\|f\|_H \leq 1$. Given a symmetric and positive definite function K (usually a covariance function), its associated RKHS is denoted as $\mathcal{H}(K)$. The notation used for the norms in all these spaces is summarized in the next page.

Through this work X denotes a stochastic process and $X(s)$ is the s-th real marginal variable, for $s \in [0, 1]$, defined in $(\Omega, \mathcal{A}, P)$. The realizations (trajectories) of this process are denoted by $x = X(\omega)$ for $\omega \in \Omega$. As usual in statistics, we use an upper hat sign to denote the data-driven estimators. Sometimes a subscript with the sample size is also added.

The superscript $'$ denotes either the derivative of a function or the transpose of a matrix (or vector). For very few exceptions, it could denote a new version of an already defined quantity. The meaning is clear for each occurrence. Given a set of points $T = \{t_1, \ldots, t_n\}$ and a function f, $f(T)$ denotes the column vector of evaluations $(f(t_1), \ldots, f(t_n))'$. Equivalently for functions in several variables. Usually an asterisk will denote the optimal value under some optimality criterion.

$\mathbb{I}_A$ stands for the indicator function of the set A. When used without any sub-index, $\mathbb{I}$ represents the identity operator in a function space.

For the sake of readability, we include here a brief list of the main symbols and abbreviations.

Symbols

$\|\cdot\|$	Norm of $L^2(\Omega)$
$\|\cdot\|_2$	Norm of $L^2[0,1]$ or the Euclidean norm of a vector in $\mathbb{R}^d$
$\|\cdot\|_K$	Norm of $\mathcal{H}(K)$
$\|\cdot\|_{c_0}$	Norm of $\mathcal{H}(c_0)$ (time series context)
$\|\cdot\|_\Sigma$	Norm of $\mathcal{H}(\Sigma)$
$\|\cdot\|_\infty$	Uniform norm of $C(S)$
$\|\cdot\|_{\mathcal{L}}$	Norm of $\mathcal{L}$
$\|\cdot\|_{HS}$	Hilbert-Schmidt norm
$\langle x, y\rangle$	Inner product for $x, y \in \mathcal{L}^2(\Omega)$
$\langle x, y\rangle_2$	Inner product for $x, y \in \mathcal{L}^2[0,1]$ or $x, y \in \mathbb{R}^d$
$\langle f, g\rangle_{\mathcal{H}}$	Inner product for $f, g \in \mathcal{H}$
$\langle f, g\rangle_K$	Inner product for $f, g \in \mathcal{H}(K)$
$\langle f, g\rangle_{c_0}$	Inner product for $f, g \in \mathcal{H}(c_0)$ (time series context)
$\langle X, f\rangle_K$	Inverse of Loève's isometry $\Psi_X^{-1}(f)$, for a process X and $f \in \mathcal{H}(K)$
$\langle X, f\rangle_{c_0}$	Inverse of Loève's isometry $\Psi_X^{-1}(f)$, for a process X and $f \in \mathcal{H}(K)$ (time series context)
$\xrightarrow{d}$	Convergence in distribution
$\xrightarrow{P}$	Convergence in probability
$\xrightarrow{\text{a.s.}}$	Almost sure convergence
$\xrightarrow{L^2}$	Convergence in $L^2(\Omega)$
$P_0 << P_1$	P_0 is absolutely continuous with respect to P_1
$\mathrm{d}P_1/\mathrm{d}P_0$	Radon-Nikodym derivative of P_1 with respect to P_0
$T_p \prec T_{p+1}$	T_p is subvector of T_{p+1} (the components of T_p are included within those of T_{p+1})
$\otimes$	Tensor product of Hilbert spaces or tensor product operator
α	Smoothing parameter
adj	Adjoint matrix
argmin	Argument of the minima
argmax	Argument of the maxima
B	Standard Brownian motion
c_j	$\mathrm{cov}\big(X(t_j), Y\big)$ for $t_j \in T_p$
$c_n(\cdot,\cdot)$	Lagged covariance function of a time series
c_{T_p}	Vector $\big(\mathrm{cov}\big(X(t_1), Y\big), \ldots, \mathrm{cov}\big(X(t_p), Y\big)\big)'$ for $t_1, \ldots, t_p \in T_p$
$c_{T_p,j}$	Vector $\big(\mathrm{cov}\big(X(t_1), X(t_j)\big), \ldots, \mathrm{cov}\big(X(t_p), X(t_j)\big)\big)'$ for $t_1, \ldots, t_p \in T_p$
cov	Covariance
$C(S)$	Space of continuous functions over S

d_{FM}^k	Mahalanobis-based semidistance proposed in Galeano et al. (2015)
det	Determinant of a matrix
ε	Random error
ϵ	Small positive value
e_j	Eigenvectors of a covariance matrix or eigenfunctions of a cov. operator
$\mathbb{E}$	Mathematical expectation
$\mathcal{F}$	General function space
g^*	Bayes classifier
$\mathcal{H}$	Generic Hilbert space
$\mathcal{H}(K)$	Reproducing Kernel Hilbert Space associated with K
$\mathcal{H}(c_0)$	Reproducing Kernel Hilbert Space associated with c_0 (time series context)
$\mathcal{H}(\Sigma)$	Finite-dim. Rep. Kernel Hilbert Space associated with matrix Σ
$\mathbb{I}$	Identity operator
$\mathbb{I}_A$	Indicator function of the set A
inf	Infimum value
$\mathcal{K}$	Covariance operator
K	Covariance function of a stochastic process
λ_j	Eigenvalues of a covariance matrix or a covariance operator
$\bar{\lambda}$	Supremum of λ_j
L	Log likelihood function
L^*	Bayes error
$\mathcal{L}$	Space of bounded linear operators
ln	Natural logarithm
$L^2(\Omega)$	Space of random variables with finite variance
$L^2[0,1]$	Space of square integrable functions over $[0,1]$
m	Mean function of a process (or sample size of a time series)
M	Classical Mahalanobis distance
M_α	Functional Mahalanobis-type distance
$\mathcal{N}$	Standard Gaussian random variable
$\mathbb{N}$	Set of natural numbers
p	Number of selected points or probability of class 1 ($\mathbb{P}(Y=1)$)
P	Probability measure
P_S	Orthogonal projection onto the closed subspace S
$\mathbb{P}$	Probability of a random event
π_j	Prior probability of population j
Ψ_X	Loève's isometry (Eq. (1.3))
$\mathbb{R}$	Set of real numbers

σ_j^2	$\mathrm{var}\big(X(t_j)\big)$
Σ	Covariance matrix
Σ_{T_p}	Covariance matrix of the variables $X(t_j)$ indexed by the points in T_p
$\mathrm{span}(S)$	Closure of the set of all finite linear combinations of elements in S
sup	Supremum value
Θ_p	Compact subspace of $[0,1]^p$ defined in (3.5)
T_p	Vector of points $(t_1,\ldots,t_p)' \in [0,1]^p$
var	Variance
x_α	Smoothed trajectory given in Equation (2.13)
$\bar{x}$	Sample mean of X
$\widetilde{X}$	Centered stochastic process $X - m$
$\mathbb{Z}$	Set of integer numbers
Z_{T_p}	Orthogonal projection of the r.v. Z on $\mathrm{span}\{X(t_j) - m(t_j), t_j \in T_p\}$

Abbreviations

a.s.	Almost surely
AIC	Akaike information criterion
AR	Autoregressive (process)
ARB	Autoregressive (process) in a Banach space
BIC	Bayesian information criterion
Bm	Brownian motion
CV	Cross validation
EM	Expectation Maximization
FAR	Functional Autoregressive
fBm	Fractional Brownian motion
FCAR-sparse	Sparse functional autoregressive process of Definition 3.20
FDA	Functional Data Analysis
fFPE	Functional final prediction error-type
FLR	Functional logistic regression
fPCA	Functional Principal Component Analysis
gBm	Geometric Brownian motion
HS	Hilbert-Schmidt
HSIC	Hilbert-Schmidt Independence Criterion
iBm	Integrated Brownian motion
i.e.	"id est" (that is)

knn	k-nearest neighbors
KPS	Variable selection procedure proposed in Kneip et al. (2016)
KR	Prediction method proposed in Kokoszka and Reimherr (2013)
LASSO	Least absolute shrinkage and selection operator
MCD	Minimum covariance determinant
MH	Maxima Hunting
ML	Maximum likelihood
MLE	Maximum likelihood estimator
MMD	Maximum Mean Discrepancy
nonP	Non-parametric regression method proposed in Ferraty and Vieu (2006)
OB	Optimal Bayes classifier
ONB	Orthonormal basis
OU	Ornstein-Uhlenbeck
PCA	Principal Component Analysis
PLS	Partial Least Squares
PM10	Particulate matter concentrations data set
PVS	Partitioning Variable Selection
RK	Method for variable selection proposed in Berrendero et al. (2017)
RKHS	Reproducing Kernel Hilbert Space
RMSE	Relative mean square error
r.v.	Random variable
SC	Sign choice (property)
SLLN	Strong law of large numbers
SMM	Support Measure Machine
s.t.	Such that
SVM	Support Vector Machine
wav	1^{st} functional logistic regression method in Mousavi and Sørensen (2017)

Chapter 1

Introduction

1.1 Functional Data Analysis

The classical theory of mathematical statistics (as started at the beginning of 20th century) was essentially limited to deal with data observations on the real line. By the middle of the century, the theory was developed to cope with multivariate data (i.e., data in the Euclidean space $\mathbb{R}^d$). In the last three decades, the technological progress has posed new statistical problems which require the treatment of new types of data, including high-dimensional and functional data.

Throughout this thesis we will delve into some functional data problems, mainly from a theoretical point of view. The collection of methods and techniques developed in this setup is commonly known as Functional Data Analysis (FDA), term probably coined by Ramsay (1982). This field is currently very active and has many specialized directions, so let us establish what we mean by functional data. In what follows our data will consist of real valued functions $x_1(t), \ldots, x_n(t)$ defined on the interval $[0, 1]$ (or any other compact set of the real line). Of course this definition can be extended to include more sophisticated models, but these are out of the scope of this work.

In contrast to classical multivariate data, the existence of such a "continuous" kind of data has been sometimes questioned, since in real problems the curves $x(t)$ must be discretized on a grid. This has been a concern specially among the machine learning community, where FDA techniques are not yet very popular. However, modern measurement devices allow us to record data over increasingly fine grids. Such discretized data can be interpolated to reconstruct true functions on a compact interval. This has some important methodological advantages, as the fact of passing to "continuous" (infinite-dimensional) data often leads to simpler, easier to interpret models. Besides, working with proper functions has other important advantages against high-dimensional vectors. For instance, we can extract additional information from the curves, like rates of change or derivatives. Another important factor is computational performance. Purely functional-derived methods can often deal with a huge amount of data in a more efficient way, taking advantage of the properties of infinite-dimensional spaces. For this reason FDA is sometimes considered as a part of the wide area of big data.

From a historical point of view, one of the earliest contributions to FDA was the paper of Dauxois et al. (1982), a ground-breaking study of the principal components for functional data. However, it was not until the release (in 1997) of the first edition of Ramsay and Silverman (2005) that this discipline got off the ground. This monograph was the first attempt to collect all the progresses made thus far on the topic and its informative tone helped to reach a broad audience. In a more recent work, Ramsay and Silverman (2002), the same authors emphasize the applications of the statistical techniques exposed in their previous book, exemplifying them with real data problems. All the examples and computational details of that book are compiled in Graves et al. (2009), from where the popular R-package `fda` arose. Later on, Ferraty and Vieu (2006) expose the characteristics and difficulties of functional data from a more mathematical point of view. These authors take a non-parametric approach to FDA problems. Also from a mathematical and non-parametric perspective the book by Bosq and Blanke (2007) is mainly focused on statistical techniques for prediction problems. A more recent book combining theory and practice is Horváth and Kokoszka (2012). Unlike the previous works mentioned here, this one has a significant part devoted to dependent data, in particular to time series analysis. More recently, Hsing and Eubank (2015) provides an account of the main concepts involved in functional data theory, including Reproducing Kernel Hilbert Spaces (RKHS), the main focus of this thesis.

A recent and quite popular review of theory and methods on the topic can be found in Cuevas (2014). In addition, the book of Ferraty and Romain (2011) is a collection of surveys from different authors on several FDA problems. Other works about different statistical problems with functional data are, for instance, Zhao et al. (2004) and Shi and Choi (2011).

In this thesis we address three different problems of FDA, mainly from a mathematical point of view. A common feature in our original contributions here is the use of Reproducing Kernel Hilbert Spaces (which we will just denote as RKHS's). Thus, let us establish the mathematical definition of functional data we will adhere to. From our perspective, the data are realizations (trajectories) drawn from a continuous time stochastic process, whose marginal variables $X(s)$, $s \in [0,1]$, are defined in a common probability space $(\Omega, \mathcal{F}, P)$. That is, we understand stochastic processes as random variables taking values on an infinite-dimensional space of functions. The most common choices for such space are $L^2[0,1]$ and $C[0,1]$, the spaces of square integrable functions and continuous functions over $[0,1]$, respectively. In general, we will assume that the stochastic processes we work with have a continuous strictly positively definite covariance function $K(s,t)$ and a continuous mean function $m(s)$. This model-based approach to FDA, relying on stochastic processes, has been proven very useful. However, in the words of Biau et al. (2015), *"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 leap from finite-dimensional problems towards functional ones poses new chal-

lenges. We list hereunder some of the main difficulties one has to cope with when dealing with functional data.

- Most times functional data can not be recorded in a continuous way, so that they are discretized on a grid. Although the step size of this grid may be very small, there are still many different representations potentially suitable for the same measurements. A popular choice is to represent the data on a basis and to keep only a finite number of elements (properly chosen). This problem can be circumvented with the use of non-parametric techniques. In addition, the choice of the functional space into which the realizations of the process fall can drastically affect the theoretical results. For instance, the space might not be a Hilbert space, thus lacking an inner product (this is the case of $C[0,1]$). If we choose $L^2[0,1]$ as our framework, the point-wise evaluations of the functions are not properly defined as linear continuous transformations.

- There is no obvious complete order among curves in functional spaces. Thus, some core concepts like the median for multivariate data, which requires a notion of centrality, can not be directly extended. Besides, the norms of the possible ambient spaces are not equivalent, which could significantly change the obtained results for a fixed set of curves.

- The well-known phenomenon named "curse of dimensionality" for multivariate data is drastically magnified in the usual function (infinite-dimensional) spaces. As a consequence, one usually gets slow convergence rates for nonparametric methods, since such rates depend on the so-called "small ball probabilities". These are the probabilities of the closed balls of radius ϵ tending to zero. Typically for random vectors of dimension d, such probabilities are in the order of ϵ^d but this order is much smaller in function spaces; see Bongiorno and Goia (2017) and Ferraty and Vieu (2006). One possible solution to this problem, which we explore in the third chapter of this thesis, is to reduce the dimension of the data.

- Even though the curves are discretized as a (likely high-dimensional) vector, the continuous character of the data causes many variables to be highly correlated. Specially the variables corresponding to close time points. Thus, there are also a lot of redundant entries in these random vectors. This preclude the use of many classical tools designed for multivariate data. In particular, this could lead to invertibility problems with the covariance matrices. In fact, many statistical techniques which require to invert the covariance matrix can not be directly extended. For stochastic processes the role of the covariance matrix is played by the integral covariance operator, which is typically non-invertible due to its compactness.

- From a deeper theoretical perspective, one of the sharpest problems when working in functional spaces is the lack of a natural translation-invariant measure like the Lebesgue measure in $\mathbb{R}^d$. A major consequence of this fact is the corresponding

lack of natural density functions for continuous stochastic processes. In some problems where two probability measures are involved, this lack of densities can be mitigated through the use of Radon-Nikodym derivatives (fixing one of the measures as a reference). We will see later on an explicit example of this technique for functional binary classification. See Section 3.3 of Cuevas (2014) for more details.

Throughout this thesis we connect some finite dimensional problems with their functional counterparts through their formulations in terms of RKHS's. This will help us to partially overcome some of the above mentioned difficulties.

1.2 RKHS: theory and applications

Reproducing Kernel Hilbert Spaces (RKHS's), first introduced by Aronszajn (1950), have been used in many different fields of mathematics. Emanuel Parzen pioneered the applications of RKHS's to statistics; Parzen (1959). We totally agree with Parzen's statement: *"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"*; Parzen (1961a). We should point out that, in Parzen's terminology, the meaning of "time series" was not the same as it is nowadays, since it was applied to continuous-time stochastic processes $\{X(s),\, s \in [0,1]\}$.

Over the past few years RKHS theory has become very popular in the statistics and machine learning communities. The book of Berlinet and Thomas-Agnan (2004) brings together the RKHS methodology and statistical applications till the date. Appendix F of Janson (1997) provides a good short summary of RKHS theory. Apart from that, one of the most well-known applications of RKHS's in machine learning are the Support Vector Machines (SVM's). The books of Schölkopf and Smola (2002) and Steinwart and Christmann (2008) give a quite complete overview of the connection between RKHS theory and SVM's.

We present next a recap of the different definitions and motivations of RKHS's commonly found in the literature. We also give a quick overview of the, in our opinion, most relevant applications of RKHS's in statistics.

1.2.1 Different ways of seeing RKHS's

Reproducing kernel Hilbert spaces are a special type of Hilbert spaces with some nice features. The definition of these spaces, which we will denote usually by $\mathcal{H}(K)$, is associated with a positive-semidefinite kernel function $K = K(s,t)$. Among their many interesting properties, let us start recalling one (that motivates the name of such spaces) which will be particularly useful in what follows.

Reproducing property. $f(s) = \langle f, K(\cdot, s)\rangle_K$, for all $f \in \mathcal{H}(K)$, $s \in [0,1]$,

where $\langle \cdot, \cdot \rangle_K$ stands for the inner product of $\mathcal{H}(K)$. Then, RKHS's are typically defined as Hilbert spaces such that

- the kernel function K fulfills this reproducing property and
- all the functions $K(s, \cdot)$, $s \in [0,1]$, belong to the space.

As it often happens with many important mathematical ideas, RKHS's came along in diverse areas from different perspectives. Hereunder we introduce some alternative definitions. All of them are equivalent to the previous one.

RKHS's via finite linear combinations of a kernel function

We start with the definition to which we mainly refer to throughout this thesis. Given a symmetric positive-semidefinite function $K = K(s,t)$ (also called a function of positive type) we can construct a unique RKHS associated with K. This function K is known as the kernel of the space. In fact, by Moore-Aronszajn theorem (for instance Steinwart and Christmann (2008, p.118)), we know that every symmetric positive-semidefinite function K is the kernel of a unique RKHS.

We first introduce an auxiliary space, associated with K, which we will denote as $\mathcal{H}_0(K)$. It is defined as the set of all finite linear combinations of type $\sum_i^n a_i K(s, t_i)$, that is,

$$\mathcal{H}_0(K) := \Big\{ f : \ f(s) = \sum_{i=1}^{n} a_i K(s, t_i), \ a_i \in \mathbb{R}, \ t_i \in [0,1], \ n \in \mathbb{N} \Big\}. \tag{1.1}$$

In such space we define an inner product $\langle \cdot, \cdot \rangle_K$ by

$$\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_j, t_i), \tag{1.2}$$

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

The RKHS associated with K, denoted by $\mathcal{H}(K)$, is defined as the completion of $\mathcal{H}_0(K)$. More precisely, $\mathcal{H}(K)$ is the set of functions $f : [0,1] \to \mathbb{R}$ obtained as t-pointwise limits of Cauchy sequences $\{f_n\}$ in $\mathcal{H}_0(K)$; see Berlinet and Thomas-Agnan (2004, p. 16). Thus, in heuristic terms, one could say that $\mathcal{H}(K)$ is made of all linear combinations of type $f(s) = \sum_{i=1}^n a_i K(s, t_i)$ plus all the functions which can be obtained as limits of them. Then, it is clear that this space fulfills the two properties of the first definition of RKHS's. A natural question is when we can ensure that we have identifiability of the functions. It is easy to see that the elements of $\mathcal{H}_0(K)$ have a unique representation in terms of K whenever K is strictly positive definite.

In our context, K will usually be the covariance function of the L^2-process X. From the reproducing property we see that $K(s,\cdot)$ behaves, in some sense, like Dirac's delta. Thus, we could think about using this fact to perform variable selection (as we do in Chapter 3). However, in this case there is a not-so-nice feature in the RKHS associated with the process $X(t)$: under very general conditions, this space does not contain, with probability one, the trajectories of the process X; see, e.g., (Lukić and Beder, 2001, Cor. 7.1), (Pillai et al., 2007, Th. 11). This will have some consequences later on, since in the problems that we address it usually appears $\langle x_i, f\rangle_K$, where f is any function in $\mathcal{H}(K)$ and x_i a realization of the process. To address this problem, we follow the approach proposed by Parzen (1961a), which we expose next.

RKHS defined as the image of Loève's isometry

As pointed out at the beginning of this chapter, one crucial difference between the finite and the functional (infinite-dimensional) settings is that in function spaces we do not have the Lebesgue measure as a reference. As a consequence, in the finite-dimensional setting many probability measures can be characterized in terms of their densities with respect to Lebesgue measure, but this is not the case for functional data. This problem is partially circumvented through the use of Randon-Nikodym derivatives. As we will see later on, sometimes these derivatives depend on the so-called Loève's isometry associated with the underlying process. In those cases the RKHS methodology turns out to be very useful. In addition, as a further application of Loève's isometry, we will be able to address the problem just mentioned of the trajectories not belonging to $\mathcal{H}(K)$.

Our starting point from this perspective is the L^2 stochastic process X with covariance function K and mean function m. We define another Hilbert space which is also closely related to the process. We can derive a similar definition of a pre-Hilbert space as we did for $\mathcal{H}_0(K)$, but using the marginal variables of the process $X(s)$ instead of the kernel evaluations $K(s,\cdot)$. Denote by $\mathcal{L}_0(X)$, the linear (centered) span of X (i.e., the family of finite linear combinations of type $\sum_i a_i\big[X(t_i) - m(t_i)\big]$) and let $\mathcal{L}(X)$ be the $L^2(\Omega)$-completion of $\mathcal{L}_0(X)$. It is clear that $\mathcal{L}(X)$ is a closed subspace of the usual Hilbert space $L^2(\Omega)$ of random variables with finite second moments; this can be seen as the minimal Hilbert space including the (centered) variables $X(s)$. We denote the inner product in $L^2(\Omega)$ as $\langle\cdot,\cdot\rangle$. Then *Loève's isometry* (see Lukić and Beder (2001, Lemma 1.1)) is defined as

$$\Psi_X(U)(s) = \mathbb{E}[U(X(s) - m(s))] = \langle U, X(s) - m(s)\rangle, \quad U \in \mathcal{L}(X), \tag{1.3}$$

which up to now is just an injective linear mapping from $\mathcal{L}(X)$ to $L^2[0,1]$. We define the RKHS simply as the image of this mapping,

$$\mathcal{H}(K) = \{\Psi_X(U),\ U \in \mathcal{L}(X)\},$$

which is a Hilbert space endowed with the inner product inherited from $\mathcal{L}(X)$,

$$\langle f, g\rangle_K = \langle \Psi_X^{-1}(f), \Psi_X^{-1}(g)\rangle,\ f, g \in \mathcal{H}(K).$$

With this construction it is clear that Ψ_X is an isometry, since it is a linear bijection and preserves the inner product (Berlinet and Thomas-Agnan (2004, Th. 35)). We still should check that the image of this isometry is truly an RKHS. On the one hand, we recover the kernel functions $K(s,\cdot)$ by applying the isometry to the centered marginal variables $X(s)-m(s)$. On the other hand, these kernel evaluations fulfill the reproducing property,

$$\langle f, K(s,\cdot)\rangle_K = \langle \Psi_X^{-1}(f), X(s)-m(s)\rangle = \mathbb{E}[\Psi_X^{-1}(f)(X(s)-m(s))] = \Psi_X\big(\Psi_X^{-1}(f)\big)(s) = f(s),$$

where $f \in \mathcal{H}(K)$. Thus, as mentioned before, $\mathcal{H}(K)$ meets the two conditions to be an RKHS. Of course with this construction we also obtain that the finite linear combinations of $K(s,\cdot)$, $s \in [0,1]$, are dense in $\mathcal{H}(K)$, but this is not the main focus of this definition.

Alternatively, Loève's isometry can be defined as the continuous extension of the images of elements in the pre-Hilbert space $\mathcal{L}_0(X)$,

$$\Psi_X\Big(\sum_i a_i\big[X(t_i)-m(t_i)\big]\Big) = \sum_i a_i K(\cdot, t_i).$$

Let $x = X(\omega)$, for some $\omega \in \Omega$, be a trajectory of the process. Using this definition, as suggested by Parzen (1961a), we can identify $\langle x, f\rangle_K$ with $(\Psi_X^{-1}(f))(\omega) \equiv \Psi_x^{-1}(f)$. In particular the random variables $X(t_i)-m(t_i)$ are the inverse images of the functions $K(\cdot,t_i) \in \mathcal{H}(K)$ in such isometry. Thus, through the inverse of Loève's isometry we recover (for realizations of the process) the Dirac's delta behavior we had with the reproducing property.

RKHS defined as the image of the covariance operator

We have seen that the RKHS associated with the process X is closely related with its covariance structure. Sometimes, specially in time-series problems, the covariance structure of the process is rather expressed in terms of the covariance operator associated with the covariance function K. If the trajectories of the process $X = X(t)$ are in $L^2[0,1]$, the covariance operator $\mathcal{K} : L^2[0,1] \to L^2[0,1]$ is defined as

$$\mathcal{K}(f)(\cdot) = \int_0^1 K(s,\cdot)f(s)\mathrm{d}s = \mathbb{E}\big[\langle X-m, f\rangle_2\big(X(\cdot)-m(\cdot)\big)\big], \tag{1.4}$$

where $\langle\cdot,\cdot\rangle_2$ denotes the inner product in $L^2[0,1]$. Some classical, interesting properties of this operator can be found in Chapter III of Cucker and Smale (2001). For instance, in our case we will assume that the covariance function K is continuous, which implies that

$\mathcal{K}$ is a compact positive self-adjoint operator. However, unlike the finite-dimensional cases, $\mathcal{K}$ is not invertible in $L^2[0,1]$.

In general we will denote the square root of an operator, $\mathcal{K}^{1/2}$, as the (unique) operator such that $\mathcal{K}^{1/2}\mathcal{K}^{1/2}$ equals $\mathcal{K}$. This square root operator inherits some nice properties from $\mathcal{K}$. Namely, it is also positive and self-adjoint (see, for example, Theorem 3.35 and Problem 3.32 of Kato (2013, Chapter 5)). In a similar way as we did before with Loève's isometry, we can define the RKHS as the image of the square root operator (e.g. Definition 7.2 of Peszat and Zabczyk (2007)),

$$\mathcal{H}(K) = \{\mathcal{K}^{1/2}(f),\ f \in L^2[0,1]\}, \tag{1.5}$$

with the inner product, for $f, g \in \mathcal{H}(K)$,

$$\langle f, g\rangle_K = \langle \mathcal{K}^{-1/2}(f),\ \mathcal{K}^{-1/2}(g)\rangle_2.$$

This expression is specially useful to rewrite the norm of the RKHS in terms of the $L^2[0,1]$-norm, which sometimes is easier to compute. In order to check the equivalence of this definition with the previous ones we will use the spectral decomposition of the functions in $\mathcal{H}(K)$.

Theorem 2 in Cucker and Smale (2001, Chapter III) states that, in our setting, $\mathcal{H}(K)$ consists of continuous functions. Moreover by the Spectral Theorem for compact and self-adjoint operators (for instance Theorem 2 of Chapter II of that book), there exist a complete orthonormal basis $\{e_j, j = 1, \ldots\}$ in $L^2[0,1]$ of eigenfunctions of $\mathcal{K}$. The sequence of positive real eigenvalues $\{\lambda_j, j = 1, \ldots\}$ is either finite or tends to zero as $j \to \infty$. Besides, the maximum of λ_j (which are all non-negative) equals $\|\mathcal{K}\|_{\mathcal{L}}$, the norm of $\mathcal{K}$ in the space of bounded linear operators. By Corollaries 2 and 3 of that book, we know that the sum of all the eigenvalues equals $\int_0^1 K(s,s)\mathrm{d}s$ and that, whenever λ_j is greater than zero, its associated eigenfunction e_j is continuous. Then, as a consequence of Mercer's theorem (for instance Riesz and Szökefalvi-Nagy (1990, p. 245)), we can decompose the kernel function as

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t), \tag{1.6}$$

where the convergence is uniform in both variables.

In fact, since $\{e_j\}$ form an ONB in $L^2[0,1]$, by Equation (1.5) $\{\sqrt{\lambda_j} e_j\}$ is an ONB in $\mathcal{H}(K)$. Therefore we can rewrite the RKHS using the spectral decomposition of the functions on this base (this definition is presented, for instance, in Amini and Wainwright (2012) and Cucker and Smale (2001, Chapter III)):

$$\mathcal{H}(K) = \Big\{ f = \sum_{j=1}^{\infty} \alpha_j \sqrt{\lambda_j} e_j,\ \text{where } \{\alpha_j\}_{j=1}^{\infty} \in \ell_2(\mathbb{N}) \Big\},$$

$\ell_2(\mathbb{N})$ being the space of square-summable sequences. Then the inner product is rewritten, for $f = \sum_j \alpha_j \sqrt{\lambda_j} e_j$ and $g = \sum_j \beta_j \sqrt{\lambda_j} e_j$, as $\langle f, g \rangle_K = \sum_j \alpha_j \beta_j$.

From Equation (1.6) it is direct to see that the kernel functions $K(s, \cdot)$ belong to $\mathcal{H}(K)$ so defined, since $\{\sqrt{\lambda_j} e_j(s)\}_{j=1}^{\infty} \in \ell_2(\mathbb{N})$ (because $\sum_j \lambda_j e_j(s)^2 = K(s,s)$, which is finite). We can also rewrite the reproducing property with this spectral representation,

$$\langle f, K(s, \cdot) \rangle_K = \sum_{j=1}^{\infty} \alpha_j \sqrt{\lambda_j} e_j(s) = f(s),$$

where $f = \sum_j \alpha_j \sqrt{\lambda_j} e_j$. Note that this latest construction (via the eigenbasis) can be done for any positive definite function K, regardless of whether it is a covariance function or not.

RKHS's in terms of evaluation operators

RKHS's appear also in purely analytical problems. Unlike L^p spaces, which formally consist of equivalence classes of functions, RKHS's are Hilbert spaces whose elements are true functions. This is why these spaces are often known as "function spaces" among the analysis community. In this non-statistical context these spaces are defined from a different perspective with no especial emphasis on concepts such as covariance.

Given a Hilbert space consisting of real valued functions over $[0, 1]$ and a function f in this space, the evaluation operators δ_s, $s \in [0, 1]$, are defined as $\delta_s(f) = f(s)$. Such Hilbert space is an RKHS if the evaluation operators are bounded (i.e. continuous) for all $s \in [0, 1]$ (e.g. Section 4.2 of Steinwart and Christmann (2008)). That is, if there exists $C_s > 0$ such that

$$|f(s)| = |\delta_s(f)| \leq C_s \|f\|_K, \quad \text{for all } f \in \mathcal{H}(K),$$

where we have denoted the RKHS by $\mathcal{H}(K)$ as before. It is clear from this last equation that convergence in the RKHS norm implies pointwise convergence. In fact, whenever the kernel function is continuous, it also implies uniform convergence. Given a sequence f_n in $\mathcal{H}(K)$ that converges in the RKHS-norm to another function $f \in \mathcal{H}(K)$,

$$\begin{aligned} |f_n(s) - f(s)| &= |\langle K(s, \cdot), f_n - f \rangle_K| \leq \|K(s, \cdot)\|_K \|f_n - f\|_K = K(s,s)^{1/2} \|f_n - f\|_K \\ &\leq \sup_{t \in [0,1]} K(t,t)^{1/2} \|f_n - f\|_K = M \|f_n - f\|_K, \end{aligned}$$

where we use Cauchy-Schwartz inequality and the fact that the continuous function K attains its maximum in the compact interval $[0, 1]$.

The equivalence of this definition and the previous ones can be easily stated using Riesz Representation Theorem (for instance, Theorem 3.4 of Conway (1990)).

Theorem 1.1. (Riesz Representation Theorem) *If L is a bounded linear operator on a Hilbert space $\mathcal{H}$, then there exists a unique $\varphi \in \mathcal{H}$ such that, for all $f \in \mathcal{H}$, $L(f) = \langle f, \varphi \rangle_{\mathcal{H}}$.*

From this result there exists $\varphi_s \in \mathcal{H}(K)$ such that $\delta_s(f) = \langle f, \varphi_s \rangle_K$, for all $f \in \mathcal{H}(K)$. Thus, by means of the definition of δ_s we recover the reproducing property, where $K(s, \cdot) = \varphi_s(\cdot)$ is the reproducing kernel.

1.2.2 Applications in statistics

The use of RKHS's in statistics has bounced back over the past few years. We introduce here some applications that we found particularly interesting. Apart from the problems considered here, additional works which involve RKHS's are, among others, Berrendero et al. (2017); Berlinet and Thomas-Agnan (2004); Hsing and Eubank (2015) and Yuan and Cai (2010). Although we focus only on RKHS's consisting of real-valued functions, other more general RKHS's are useful for certain problems (see Kadri et al. (2016) and the references therein for more details).

The kernel trick

In the machine learning community, one of the most common applications of RKHS's is the so called kernel method, or kernel trick (on which the well-known Support Vector Machines are founded). These kernel methods have been traditionally applied to regularization theory and supervised learning problems with scalar output. The general idea of this procedure is presented here, without any specific application (for instance, see the book of Shawe-Taylor and Cristianini (2004) for details). In addition, for further information about kernels with non-scalar outputs we refer to the review of Álvarez et al. (2012).

The kernel trick can be applied as a previous step for any statistical rule that only involves inner products of the data. The data are thus embedded in a suitable feature space in such a way that the non-linear relations are transformed into linear ones. This represents the equivalent of working with the transformed data into an RKHS, as we will see.

We consider an embedding map ϕ from $\mathbb{R}^d$ (the space where the realizations are contained) to a Hilbert space $\mathcal{H}$ of higher dimension. Then the kernel function is defined as

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}, \tag{1.7}$$

which allows us to compute directly the inner products of the embedded data, without knowing explicitly the map ϕ. Then, the learning algorithms for multivariate data can

be extended by simply replacing the inner products with a kernel function. The evaluations of this kernel function on the pairs of data points are summarized in the so-called kernel matrix or Gram matrix. This matrix contains all the information about the data that is transferred to the learning algorithm. Since the embedding into a potentially infinite dimensional space is carried out just via this finite-dimensional matrix, it does not affect the computational efficiency. Besides, once that the relationships between data samples are merely linear, there are numerous statistical techniques available to solve the problem under consideration.

So far the function K is not necessarily a reproducing kernel. However, Theorem 3.11 of Shawe-Taylor and Cristianini (2004) states that a function K can be decomposed as in Equation (1.7) if and only if it is symmetric and positive-semidefinite. Thus, as already pointed out, K is the kernel of a unique RKHS, which would be the feature space. The idea behind the proof is simply using as embedding map $\phi(x) = K(x, \cdot)$, for $x \in \mathbb{R}^d$. The relationship between the feature space (that is, the RKHS) and the original space in which the data lie, as well as the transit from one space to the other, is studied in Schölkopf et al. (1999). In addition, even knowing that we are looking for a reproducing kernel, it is not always evident which is the optimal choice for a given problem. This question has been analyzed in depth, for instance in Chapter 4 of Steinwart and Christmann (2008) or in Section 13.1 of Schölkopf and Smola (2002).

Embedding of probability measures

The embedding map ϕ defined in the previous paragraphs was tailored to transform multivariate data into a suitable RKHS. This notion can be extended in order to perform embeddings of probability measures. This has a variety of interesting applications as we will see.

The most common way to perform this kind of embedding is to associate a given probability measure ν over $[0, 1]$ with the function $\mu_\nu \in \mathcal{H}(K)$ (the RKHS with kernel K) such that

$$\langle f, \mu_\nu \rangle_K = \int f(x) \mathrm{d}\nu(x), \quad \text{for all } f \in \mathcal{H}(K).$$

Whenever this integral defines a compact operator, the existence of such a unique function μ_ν is ensured by Riesz Representation Theorem stated above. That is, in this case the RKHS is wide enough to ensure the identifiability of the embedded measure. Taking $f = K(s, \cdot)$ in the previous equation and resorting to the reproducing property of K we obtain that

$$\mu_\nu(s) = \int K(s, x) \mathrm{d}\nu(x). \tag{1.8}$$

This kind of embedding has been widely studied, for instance by Sriperumbudur et al. (2010), where the authors analyze, among other aspects, when this embedding is injective. For instance, one sufficient condition for this to hold is that the kernel of the

RKHS should be integrally strictly positive definite. That is,

$$\int_0^1 K(s,t)\mathrm{d}\lambda(s)\mathrm{d}\lambda(t) > 0,$$

for all non-zero signed Borel measures λ on $[0,1]$. It is easy to see that such a kernel is also strictly positive definite, but the converse is not true. In addition, Smola et al. (2007) apply this kind of embedding to a variety of problems. We introduce hereunder some of these applications.

In Gretton et al. (2007) the authors propose a discrepancy measure between probability distributions named Maximum Mean Discrepancy (MMD), which was mainly applied to homogeneity testing problems. Given two probability measures ν and ρ on $[0,1]$, the MMD between them is defined as $\|\mu_\nu - \mu_\rho\|_K$. In the same work the authors define two statistical homogeneity tests to contrast the null hypothesis $H_0 : \nu = \rho$ against the alternative hypothesis $H_1 : \nu \neq \rho$. The first test, which was previously proposed in Borgwardt et al. (2006), is based on a sample-derived threshold and the null hypothesis is rejected whenever this value is exceeded. This threshold is computed using a bootstrap approximation of the distribution of $MMD(\nu,\rho)$ calculated through a biased estimator. This test is improved by the same authors in Gretton et al. (2009), where they give a consistent estimate of the distribution of this discrepancy measure under the null hypothesis, which does not rely on any bootstrap approximation. The second test Gretton et al. (2007) is based on the asymptotic distribution of an unbiased estimator of $MMD(\nu,\rho)^2$. These three proposals are compiled in Gretton et al. (2012). In addition, a brief version of these ideas can be found in Section 2.1 of Smola et al. (2007).

The above-defined Maximum Mean Discrepancy is also used to define an independence measure, presented for instance in Section 2.3 of Smola et al. (2007). Given two real random variables X and Y with probability measures ν_X and ν_Y respectively, they are independent if and only if its joint distribution ν_{XY} coincides with the product $\nu_X\nu_Y$. Then, in the same spirit as before, one could check the independence of the variables by measuring the distance between the embeddings of ν_{XY} and $\nu_X\nu_Y$. In order to define the embeddings of these two-dimensional measures we need to introduce the tensor product of two RKHS's, $\mathcal{H}(K_1)$ and $\mathcal{H}(K_2)$, which is denoted by $\mathcal{H}(K_1)\otimes\mathcal{H}(K_2)$. It is easy to see that this is also an RKHS, $\mathcal{H}(R)$, with reproducing kernel $R((s_1,s_2),(t_1,t_2))$ equal to $K_1(s_1,t_1)K_2(s_2,t_2)$. Thus, the embeddings of ν_{XY} and $\nu_X\nu_Y$ into this product space are defined as,

$$\begin{aligned}
\mu_{\nu_{XY}}(\cdot,\star) &= \int_0^1 R((s,t),(\cdot,\star))\mathrm{d}\nu_{xy}(s,t) \quad \text{and} \\
\mu_{\nu_X\nu_Y}(\cdot,\star) &= \int_0^1\int_0^1 R((s,t),(\cdot,\star))\mathrm{d}\nu_x(s)\mathrm{d}\nu_y(t),
\end{aligned}$$

respectively. Therefore, the MMD for this problem is rewritten as $\|\mu_{\nu_{XY}} - \mu_{\nu_X\nu_Y}\|_R$. In Sejdinovic et al. (2013) the authors prove that this independence measure is equivalent

to the Hilbert-Schmidt Independence Criterion (HSIC) defined in Gretton et al. (2005) and Gretton et al. (2008), which is also based on this type of embeddings. Namely, this criterion is defined as

$$HSIC(x, y) = \|\mu_{\nu_{XY}} - \mu_{\nu_X} \otimes \mu_{\nu_Y}\|_{HS},$$

where now $\otimes$ denotes the tensor product operator and $\|\cdot\|_{HS}$ stands for the Hilbert-Schmidt norm of operators from $\mathcal{H}(K_2)$ to $\mathcal{H}(K_1)$. We recall that the squared Hilbert-Schmidt norm of such an operator T is defined by $\sum_{i,j}\langle Te_j, u_i\rangle^2_{K_1}$ for $\{u_i\}$, $\{e_j\}$ orthonormal bases of $\mathcal{H}(K_1)$ and $\mathcal{H}(K_2)$, respectively. A variation of this criterion is used in Jitkrittum et al. (2017) also to define an independence test, but measuring the distance of a properly chosen finite set of evaluations of the probability measures.

Furthermore, the expression of the embedding given in Equation (1.8) is reminiscent of that of a kernel density estimator if ν is a empirical probability distribution. In Section 2.5 of Smola et al. (2007) the authors propose a density estimator based also on these embeddings, but from a rather different perspective. In particular, if we denote as ν_n a empirical distribution measure, they propose a restricted maximum entropy distribution estimator as,

$$\widehat{\nu} = \underset{\nu}{\operatorname{argmax}}\, H(\nu) \;\text{ such that }\; \|\mu_{\nu_n} - \mu_\nu\|_K < \epsilon,$$

for $\epsilon > 0$ small and H some entropy-like function. Since this optimization problem is posed for all possible distribution measures ν, it can not be computationally addressed. Thus, the authors propose to estimate instead the coefficients of a mixture of distributions in a fixed family.

There are many other interesting applications of these embeddings, or modifications of them. For instance, Muandet et al. (2012) propose an extension of the SVM for classification purposes to deal with the embeddings μ_{ν_X} and μ_{ν_Y}, X, Y being two real random variables. The authors call this extension Support Measure Machines (SMM). Thus, in this case there are two kernels involved, the first one from which we construct the embeddings and the second one, $R(\mu_{\nu_X}, \mu_{\nu_Y})$, used to perform the "kernel trick".

Functional data classification

The classification problem with multivariate data is widely known and its formulation can be easily extended to the functional setting. For clarity we start summarizing the basic concepts of the extension we consider. We focus just on the binary supervised classification problem, where the explanatory variable X, taking values in a space $\mathcal{F}$, belongs either to population zero or one, with respective probability measures P_0 and P_1. The membership of an observation to these classes is indicated by a random variable Y whose range is $\{0, 1\}$. In our context, typical choices for $\mathcal{F}$ are $L^2[0, 1]$ or $C[0, 1]$ (that is, X would be a stochastic process), while for multivariate data $\mathcal{F}$ would

be $\mathbb{R}^d$. Thus, the supervised classification problem consists of predicting the label y of a new functional observation x using a training sample of well classified observations.

Different classifiers are constructed for this purpose, which are merely measurable functions $g : \mathcal{F} \to \{0, 1\}$. The performance of a classifier is assessed in terms of the classification error $L = \mathbb{P}(g(X) \neq Y)$. The minimum error that can be attained for a given problem is the well-known Bayes error L^*. This error is obtained with the Bayes classifier $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$, where $\mathbb{I}$ is the indicator function and $\eta(x)$ is the conditional probability $\mathbb{P}(Y = 1|X = x)$. This optimal classifier is usually unknown in real problems, so the general aim is to design discrimination rules that give reasonable approximations of it.

When the samples are drawn from variables in $\mathbb{R}^d$ whose density functions are defined, the function $\eta(x)$, and then also the optimal classifier, can be rewritten in terms of these density functions. Namely, this classifier is given by

$$g^*(x) = \mathbb{I}_{\left\{\frac{f_1(x)}{f_0(x)} > \frac{1-p}{p}\right\}},$$

where f_0, f_1 are the density functions of the two populations and p is the prior probability of class one (i.e. $p = \mathbb{P}(Y = 1)$). In Dai et al. (2017) the authors propose a functional extension of this classifier for processes whose covariance operators have a common eigenbasis. The curves are transformed into finite-dimensional vectors by retaining a small number of elements in its Karhunen-Loève expansions. More precisely, the curves are projected onto the common sequence of eigenfunctions, and the previous quotient is taken using the densities of these projections.

As we already mentioned, we lack (Lebesgue) density functions in functional spaces. However, a more general expression of the optimal classifier can be obtained by using Radon-Nikodym derivatives. This can be done whenever the probability measures P_0 and P_1 are mutually absolutely continuous (that is, $P_0(A) = 0$ if and only if $P_1(A) = 0$), which is commonly denoted as $P_1 << P_0$ and $P_0 << P_1$. Then, if $\mathrm{d}P_1/\mathrm{d}P_0$ stands for the Radon-Nikodym derivative of P_1 with respect to P_0, the Bayes rule is given by

$$g^*(x) = \mathbb{I}_{\left\{\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(x) > \frac{1-p}{p}\right\}}.$$

See Theorem 1 of Baíllo et al. (2011) for additional details about this result. Note that this holds for both multivariate and functional data. Recall that the Hájek-Feldman dichotomy for Gaussian measures states that two measures are either mutually absolutely continuous or mutually singular (Feldman, 1958). When we are working with finite-dimensional Gaussian random vectors only degenerate distributions are mutually singular. However, this is not the case with stochastic processes, where many interesting problems have mutually singular distributions. A very simple example can be obtained taking P_0 as the distribution of a standard Brownian Motion $\{B(t),\ t \in [0, 1]\}$ and P_1 the distribution of $\{\sigma B(t),\ t \in [0, 1]\}$ for some positive $\sigma \neq 1$.

Whenever the expressions of the Radon-Nikodym derivatives are known, we can explicitly write the Bayes classifier. As a consequence of Theorem 5A of Parzen (1961a) we have that, given two Gaussian processes with continuous trajectories and continuous covariance function $K(s,t)$, $s,t \in [0,1]$, one of them (P_1) with mean function $m(s) = \mathbb{E}X(s)$ and the other one (P_0) with zero mean, then $P_1 << P_0$ if and only if $m \in \mathcal{H}(K)$. Thus, if $m \notin \mathcal{H}(K)$, the probability measures are mutually singular. Moreover, if $P_1 << P_0$, (in the same Theorem 5A)

$$\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(x) = \exp\left\{\Psi_x^{-1}(m) - \frac{1}{2}\|m\|_K^2\right\}. \tag{1.9}$$

Theorem 2.1 of Beder (1987) is a more general version of this result, in the sense that it can be applied for s in a general set. One can also obtain an explicit expression for the Bayes error of this optimal classifier. Let p be the probability of the first class, then the Bayes error is given by (Theorem 2 of Berrendero et al. (2017))

$$L^* \;=\; (1-p)\Phi\left(-\frac{\|m\|_K}{2} - \frac{\log[(1-p)p^{-1}]}{\|m\|_K}\right) + p\;\Phi\left(-\frac{\|m\|_K}{2} + \frac{\log[(1-p)p^{-1}]}{\|m\|_K}\right), \tag{1.10}$$

being Φ the cumulative distribution function of a standard Gaussian random variable. Whenever both classes are equiprobable ($p = 1/2$), this expression becomes much simpler, just $\Phi(-\|m\|_K/2)$.

There are other interesting processes whose Radon-Nikodym derivatives are explicitly known. In most cases they depend on the covariance and mean functions of the processes. Some examples can be found in Segall and Kailath (1975); Shepp (1966); Varberg (1961, 1964).

An interesting fact is that the absolute continuity or mutual singularity of the measures determines the accuracy of the prediction. In Berrendero et al. (2017) the authors give an interpretation on these terms of the "near perfect classification" phenomenon described in Delaigle and Hall (2012). Such result is stated for the same classification problem mentioned before Equation (1.9). Specifically, denote by λ_j, e_j the eigenvalues and eigenfunctions of the common covariance operator $\mathcal{K}$ (Equation (1.4)) and by μ_j the coefficients of the mean function m of population one in the basis of eigenfunctions $\{e_j\}$. Therewith, Theorem 1 of Delaigle and Hall (2012) establish some conditions on the ℓ^2-norm (the space of square-summable sequences) of the sequence $\{\lambda_j^{-1/2}\mu_j\}_{j=1}^{\infty}$ under which the classification is "near perfect". With this we mean that it is possible to construct a classifier whose error may be as small as desired. As pointed out in Theorem 4 of Berrendero et al. (2017), the condition to obtain this "near perfect classification" is equivalent to the probability measures being mutually singular, which in turn is equivalent to $m \notin \mathcal{H}(K)$. As already mentioned, this mutually singular case in functional spaces covers a large number of non-degenerate problems.

There is still other application of RKHS's to this discrimination problem, also analyzed in Berrendero et al. (2017). The authors use the properties of these spaces to perform

variable selection on the curves. In fact it is ensured that, in some cases, the Bayes rule coincides with the Fisher linear classifier on the selected marginals of the process. This is connected with Chapters 3 and 4 of this thesis, where similar variable selection techniques are studied for functional regression problems.

Regularization techniques: the Representer Theorem

Another important use of RKHS's when dealing with functional data is regularization. These spaces are useful to impose smoothness conditions and help to reduce noise and irrelevant information. As we have repeatedly mentioned in the previous paragraphs, one of the most common spaces to work with is $L^2[0,1]$. However, this space is too large from several points of view. Hence, in many statistical problems, one typically uses penalization or projection methods to exclude extremely rough functions.

The functions that compose Reproducing Kernel Hilbert Spaces contained in $L^2[0,1]$ are often smoother than those of the entire space. Roughly speaking, the functions in $\mathcal{H}(K)$ are "as smooth as the kernel". The general idea of these regularization techniques is to replace a function $f \in L^2[0,1]$ with a close function, in the sense of the L^2-norm, that belongs to $\mathcal{H}(K)$. However, the RKHS is not a closed subspace of $L^2[0,1]$, so one can not directly project f onto it. In addition, whenever zero is not an eigenvalue of the covariance operator $\mathcal{K}$, the RKHS is dense in $L^2[0,1]$ (Remark 4.9 of Cucker and Zhou (2007)). Thus, a classical regularization approach is to solve the minimization problem

$$\underset{h\in\mathcal{H}(K)}{\operatorname{argmin}} \left(\|f-h\|_2^2 + \alpha \|h\|_K^2 \right), \tag{1.11}$$

where α is a real positive parameter that modulates the penalization. Therewith we obtain a trade-off between fitting and roughness of the curve. For a complete review on this kind of regularization techniques we refer to the book Cucker and Zhou (2007), or its summarized version Cucker and Smale (2001). As stated in Proposition 8.6 of that book, the unique solution of this minimization problem is given by $h = (\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}(f)$, where $\mathbb{I}$ denotes the identity operator. Thus, this result gives us a simple solution to the previous optimization problem in an infinite-dimensional space.

This result has a sample version: for $(x_1,y_1),\ldots,(x_n,y_n) \in \mathbb{R}^2$ and a positive real number α, the solution of the optimization problem

$$\underset{h\in\mathcal{H}(K)}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \big(y_i - h(x_i)\big)^2 + \alpha\|h\|_K \right\},$$

can always be written as $h(s) = \sum_{j=1}^{n} a_j K(t_j, s)$. This result is known as the Representer Theorem and it was originally proved by Kimeldorf and Wahba (1970). A general version of this result, beyond scalar variables, is presented in Theorem 4.2 of Schölkopf and Smola (2002). Given a sample $(x_1,y_1),\ldots,(x_n,y_n) \in \mathcal{X}\times\mathbb{R}$, where $\mathcal{X}$ may be a

functional set, a loss function $\ell : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R}$ and a strictly monotonic increasing penalization function $\omega : [0, \infty) \to \mathbb{R}$, the minimization problem now is

$$\underset{h \in \mathcal{H}(K)}{\operatorname{argmin}} \left\{ \ell((x_1, y_1, h(x_1)), \ldots, (x_n, y_n, h(x_1))) + \omega(\|h\|_K) \right\}.$$

This result is applied to functional regression with scalar response and to binary functional classification in Preda (2007), where the author also proves some rates of convergence of the solutions of both problems under less restrictive conditions than Schölkopf and Smola (2002). The subsequent natural extension of the Representer Theorem can be found in Micchelli and Pontil (2005, Th. 4.2), where it is extended to vector-valued RKHS's, i.e. the responses y_i belong to a general Hilbert space. Appendix B of Kadri et al. (2016) also presents this result. This latter paper is analyzed in more detail in the next paragraphs.

Operator-valued kernels

Although throughout this document we mainly focus on RKHS's consisting of real valued functions, more general spaces can be defined for different problems. For instance, RKHS's of vector valued functions are theoretically studied in Micchelli and Pontil (2005); Carmeli et al. (2006) and Carmeli et al. (2010). Even more general kernels, particularly the ones whose values are functions themselves, or operators between function spaces, have been studied in Caponnetto et al. (2008) and Kadri et al. (2016). We briefly introduce here the learning problem analyzed in this latter work.

As we have seen when introducing the kernel trick, RKHS's allow us to easily apply nonlinear methods to multiple learning problems. In Kadri et al. (2016) the authors develop kernel methods for regression and classification problems where both the attributes and the values (or labels) are functions. First they generalize the notion of RKHS to deal with these types of data. The first step is to define the functional spaces where the data live. We denote by $\mathcal{X}$ and $\mathcal{Y}$ the spaces of real valued functions with domains D_x and D_y respectively. In addition, $\mathcal{L}(\mathcal{Y})$ is the space of bounded linear operators from $\mathcal{Y}$ to $\mathcal{Y}$.

Then, the basic idea is to apply the kernel trick for functional data, to easily define non-linear methods. Thus, operator-valued kernels could supersede the inner products between data in functional spaces. In particular, these kernels are functions K from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{L}(\mathcal{Y})$. The kernel matrix (or Gram matrix) in this case is called "block operator kernel matrix". Given a set of functions $f_1, \ldots, f_n \in \mathcal{X}$, it is defined as a matrix with entries $K(f_i, f_j) \in \mathcal{L}(\mathcal{Y})$. That is, this matrix is an operator in $\mathcal{L}(\mathcal{Y}^n)$.

The first definition of RKHS presented at the beginning of Section 1.2.1 is extended as follows to this setting in Definition 5 of Kadri et al. (2016), as follows. The functional Hilbert space $\mathcal{F}$ is an RKHS with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ if

- the map $f \to K(g, f)(h)$ belongs to $\mathcal{F}$ for every $f, g \in \mathcal{X}$ and $h \in \mathcal{Y}$,
- (reproducing property) for every $F \in \mathcal{F}$, $g \in \mathcal{X}$ and $h \in \mathcal{Y}$, it holds

$$\langle F, K(g, \cdot)(h) \rangle_{\mathcal{F}} = \langle F(g), h \rangle_{\mathcal{Y}}. \tag{1.12}$$

That is, the RKHS consists of operators from $\mathcal{X}$ to $\mathcal{Y}$. Note that the naïve approach based on a kernel of type $K : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ would fail since, given $F \in \mathcal{F}$ and $g \in \mathcal{X}$, the reproducing property for such a kernel would provide $\langle F, K(g, \cdot) \rangle_{\mathcal{F}}$ for the left hand side of Equation (1.12) and $F(g)$ in the right hand side. But $F(g)$ is a function in $\mathcal{Y}$ instead of a real number.

Theorems 2 and 3 of Kadri et al. (2016) give an extension of Moore-Aronszajn theorem for Mercer kernels. In addition, the authors apply these kernels to problems like standard functional regression with functional response. They also give an extension of the Representer Theorem to this setting.

1.3 Original contributions

In each chapter of this thesis we address a different statistical problem with functional data from an RKHS perspective. We summarize here the main theoretical contributions derived for each of them. In addition, we have coded algorithms in the programming language R (see R Core Team (2013)) for each of the proposed techniques.

In **Chapter 2** we introduce a functional extension of the classical Mahalanobis distance. The expression of Mahalanobis distance for multivariate data involves the inverse of the covariance matrix. The functional counterpart of the covariance matrix is the covariance operator $\mathcal{K}$ defined in Equation (1.4). As mentioned, this operator is not invertible in $L^2[0, 1]$, so a direct extension of the distance is not possible. There are already a couple of interesting proposals trying to circumvent this problem. We propose a rather different perspective, motivated in RKHS terms, which is fully mathematically founded.

Our proposed definition relies on the fact that the original Mahalanobis distance from x to m in the finite-dimensional case coincides with the norm in the RKHS of $x - m$. The classical distance can be thus expressed as a simple sum involving the inverse eigenvalues of the covariance matrix of the data. As pointed out in Lemma 2.1, this expression can be extended to the infinite-dimensional case, becoming an infinite sum. One could naively define the functional Mahalanobis distance in terms of this series. However, as it is apparent from the proof of the mentioned lemma, the series is convergent only when applied to a function in the RKHS. Usually one wants to compute distances between the trajectories of the process, which do not belong to the RKHS. Then our approach is based on replacing each trajectory with a "smoother" version that does belong to the RKHS. We propose to define the functional Mahalanobis-type distance as the distance

in the RKHS norm between these smoothed trajectories, which are obtained as the solution of the minimization problem defined in Equation (2.11). Both this solution and the Mahalanobis distance thus defined have simple explicit expressions, derived in Proposition 2.2. The minimization introduces a real smoothing parameter α. With Proposition 2.8 we prove that the choice of this parameter is not critical, in the sense that our distance is continuous in α. Another benefit of this result is that it implies the point convergence of the quantile functions of the distance, which is useful if one uses it to define a depth measure.

We prove that our proposal shares some interesting properties with the classical definition. For instance:

- Mahalanobis distance is invariant when an invertible transformation is applied to the point cloud. The proposed extension is invariant when an isometry in $L^2[0,1]$ is applied to the curves (Theorem 2.4).
- It is a well-known fact that Mahalanobis distance for Gaussian data in $\mathbb{R}^d$ distributes as a sum of d independent chi-squared variables. We prove (Theorem 2.5) that, for Gaussian processes, our distance follows an infinite sum of independent chi-squared variables. We also derive explicit expressions for the mean and variance of this distribution.

The obvious way to estimate the proposed distance in practice is to replace the eigenvalues and eigenfunctions of $\mathcal{K}$ in the expression of Proposition 2.2 with their sample counterparts. Under very general conditions this estimator is proved to be almost surely consistent (Theorem 2.10). Besides, we have also obtained the asymptotic behavior of the Mahalanobis distance between the mean function and its sample counterpart (Theorem 2.13).

In order to check the practical relevance of the proposal, we apply it to three problems in which Mahalanobis distance is classically used: exploratory analysis, functional binary classification and mean inference.

Chapter 3 has two parts. In the first part we address the problem of functional regression with scalar response. We propose to replace the the inner product in $L^2[0,1]$ of the standard functional regression model with the inverse of Loève's isometry (Eq. (1.3)) of a slope function in $\mathcal{H}(K)$. This model, defined in Equation (3.3), is especially suitable to perform variable selection. In fact, it reduces to the classical finite-dimensional linear model when the slope function belongs to $\mathcal{H}_0(K)$ (Eq. (1.1)). The points $t_1, \ldots, t_j \in [0,1]$ that define this sparse slope function are sometimes called "impact points". Note that a finite model like in (3.7) can not be obtained with the standard L^2 model. This would require the evaluation functionals $x \mapsto x(t_0)$ to be continuous, which is not the case.

An important advantage of this model is that it gives a theoretical ground to variable selection. We derive three suitable optimization criteria to obtain meaningful points

$t_1, \ldots, t_p$ for prediction. These criteria are proved to be equivalent in Proposition 3.1. One of them is easily implementable in practice and it can be rewritten in an iterative way that directly suggests a greedy implementation of the proposal (Proposition 3.3). Besides, this criterion only depends on the covariances between the variables $\{X(t_1), \ldots, X(t_p)\}$ and the response, which are simple to estimate, making the algorithm really fast even for large data sets.

We propose to select the points that optimize the sample version of this criterion. Whenever the true slope function belongs to $\mathcal{H}_0(K)$, depending on p^* points, we are able to consistently estimate them. If we assume that we know the true number of points p^* to select, our estimator of the points converges almost surely and in quadratic mean to the true ones (Theorem 3.8). Once that the points are selected, one can easily predict the scalar response. This prediction is also proved to be almost sure consistent and it also converges in quadratic mean (Theorem 3.9). However, for most applications one does not know the optimal number of variables to select. In Equation (3.27) we propose a suitable estimator for this quantity, which converges almost surely to the true value p^* (Theorem 3.11). Besides, we analyze how our estimator of the points behaves when one prefers to select a large conservative number of variables p (i.e. $p > p^*$). In Theorem 3.12 we prove that the selected points are "close" to the true points eventually as the sample size increases.

In the **second part of Chapter 3** we extend the previous methodology to functional regression with functional response. The definition of the model should be adapted to select a set of points that is meaningful to predict the whole response curves. In this case the cross-covariance function between the regressors and response processes plays an important role. This function is essential part of the adapted optimality criteria, so it should be uniformly almost surely estimated to perform variable selection. This introduces some restrictions on the slope function $\beta(s,t)$. Specifically, it is required that $\beta(s,\cdot) \in \mathcal{H}(K)$ for every $s \in [0,1]$ and that the stochastic process $U(s) = \Psi_X^{-1}(\beta(s,\cdot))$ satisfies $\mathbb{E}[\sup_s U(s)^2] < \infty$. In this setting, we are able to extend the consistency results developed in the first part of the chapter: Theorem 3.16 for the optimal points, Theorem 3.17 for the estimated response curves and Theorem 3.18 for the number of selected points.

This extended model is adapted to perform variable selection in functional time series. The model with finite-dimensional kernel (i.e. $\beta(s,\cdot) \in \mathcal{H}_0(K)$) is proved to be an autoregressive (AR) process in $C[0,1]$ with a unique strictly stationary solution (Proposition 3.21). The estimated lagged-covariance functions for AR processes are strongly consistent (Lemma 3.23), so all the previous results hold in this case. The proposal is proved to be competitive both in prediction accuracy and computational efficiency.

Chapter 4 is devoted to functional logistic regression. The most common approach to functional logistic regression is defined in terms of the inner product in $L^2[0,1]$. However we prove in Theorem 4.1 that, when the distributions of X given $Y = 0, 1$ are Gaussian and homoscedastic, the logistic model induced by this model involves the inner product

in the RKHS. The derivation of this result requires Radon-Nikodym derivatives as in Equation (1.9). As in Chapter 3, this model is specially suited for variable selection. Whenever the slope function belongs to $\mathcal{H}_0(K)$ and depends on the points $t_1, \ldots, t_p$, it reduces to the finite-dimensional logistic model with regressors $\{X(t_1), \ldots, X(t_p)\}$.

We propose to select the points and the coefficients of the slope function using the maximum likelihood (ML) criterion. We carefully analyze whether the ML estimator exists (Theorems 4.4 and 4.6). It turns out that non-existence problems of the ML estimator in the finite dimensional case are drastically worsened in the functional setting. For instance, the probability with which ML does not exists tends to one when the sample size goes to infinity. In order to circumvent these problems in practice we use Firth's estimator, which exists for every sample, and select a small number of points (always less than 10).

1.4 Publications and preprints associated with this thesis

The content of Chapter 2, in which we provide an extension of the classical Mahalanobis distance for functional data, is summarized in Berrendero et al. (2018b). In Chapter 3 we analyze a variable selection technique for linear regression with functional predictors. The first part of the chapter is focused on scalar response problems and it can be found in Berrendero et al. (2018a). The second part includes an extension to stationary functional time series and it is available in Bueno-Larraz and Klepsch (2018). The results of Chapter 4 about functional logistic regression are included in a preliminary draft currently in progress.

Chapter 2

Functional Mahalanobis distance

2.1 Introduction

The classical (finite-dimensional) Mahalanobis distance and its applications

Let X be a random variable taking values in $\mathbb{R}^d$ with non-singular covariance matrix Σ. In many practical situations it is required to measure the distance between two points $x_1, x_2 \in \mathbb{R}^d$ when considered as two possible observations drawn from X. Clearly, the usual (square) Euclidean distance $\|x_1 - x_2\|^2 = (x_1 - x_2)'(x_1 - x_2) := \langle x_1 - x_2, x_1 - x_2 \rangle$ is not a suitable choice since it disregards the standard deviations and the covariances of the components of x_i (given a column vector $x \in \mathbb{R}^d$ we denote by x' the transpose of x). Instead, the most popular alternative is perhaps the classical Mahalanobis distance, $M(x_1, x_2)$, defined as

$$M(x_1, x_2) \;=\; \big((x_1 - x_2)'\Sigma^{-1}(x_1 - x_2)\big)^{1/2}. \tag{2.1}$$

Very often the interest is focused on studying "how extreme" a point x is within the distribution of X; this is typically evaluated in terms of $M(x, m)$, where m stands for the vector of means of X.

This distance is named after the Indian statistician P. C. Mahalanobis (1893-1972) who first proposed and analyzed this concept (Mahalanobis, 1936) in the setting of Gaussian distributions. Nowadays, some popular applications of the Mahalanobis distance arise in the following fields (the list of references is far from exhaustive):

Supervised classification: this use of M is mostly linked to Gaussian models. It is easy to show that in a supervised classification problem of discriminating among k Gaussian homoscedastic populations with (mean, covariance) parameters $(m_1, \Sigma), \ldots, (m_k, \Sigma)$ and prior probabilities $\pi_1, \ldots, \pi_k$, the optimal (Bayes) rule to classify a coming observation x is just to assign x to population j defined by

$$M^2(x, m_j) - 2\log \pi_j = \max_{1 \le i \le k} \big(M^2(x, m_i) - 2\log \pi_i\big) \tag{2.2}$$

Outliers detection: Robust estimators of $M(x, m_j)$ are proposed in Rousseeuw and van Zomeren (1990) in order to identify outliers and leverage points. In Penny (1996) Mahalanobis distance, combined with jackknife methods, is also used as a tool for outliers detection.

Multivariate depth measures: the function $D(x) = 1/(1 + M^2(x, m))$ has been used as an assessment of "how deep" is the point x into the a population with mean and covariance parameters m and Σ, respectively. See, for example, Zuo and Serfling (2000) for a discussion on the relative merits of this proposal when compared with other popular depth measures.

Hypothesis testing: It is a well-known result in multivariate analysis that, whenever X has a Gaussian distribution in $\mathbb{R}^d$, that is $X \sim N_d(m_0, \Sigma)$, with a non-singular covariance matrix Σ, we have

$$nM^2(\bar{x}, m_0) \sim \chi^2_d, \tag{2.3}$$

where $\bar{x}$ denotes the vector of sample means of a random sample $x_1, \ldots, x_n$ from X, and χ^2_d stands for the distribution chi-square with d degrees of freedom. Of course this result can be used to get a significance test for $H_0 : m = m_0$ versus $H_1 : m \neq m_0$.

When Σ is unknown we can estimate it with the usual empirical estimator from the sample $x_1, \ldots, x_n$. In that case we have

$$n(\bar{x} - m_0)' S^{-1} (\bar{x} - m_0) \sim T^2_{d,n-1}, \tag{2.4}$$

where $T^2_{p,n-1}$ denotes the Hotelling distribution with d and $n - 1$ degrees of freedom. A two-sample version of this statistic (and the corresponding test) is also well-known; see, e.g., Rencher (2012, Ch. 5) for additional details on Hotelling's statistic.

Goodness of fit: while this application is perhaps less relevant than those mentioned in the previous paragraphs, the paper by Mardia (1975) shows some interesting connections between the moments of Hotelling's statistic and some statistics commonly used in testing multivariate normality.

On the difficulties of defining a Mahalanobis-type distance for functional data

The framework of this thesis is Functional Data Analysis (FDA). In other words, we deal with statistical problems involving functional data. Thus our sample is made of trajectories $x_1(t), \ldots, x_n(t)$ in $L^2[0, 1]$ drawn from a second order stochastic process $X(t),\ t \in [0, 1]$ with $m(t) = \mathbb{E}[X(t)]$. The inner product and the norm in $L^2[0, 1]$ are denoted by $\langle \cdot, \cdot \rangle_2$ and $\| \cdot \|_2$, respectively. We will henceforth assume that the covariance function $K(s, t) = \mathrm{cov}(X(s), X(t))$ is continuous and positive definite. The function

K defines a linear operator $\mathcal{K} : L^2[0,1] \to L^2[0,1]$, called covariance operator (already defined in Equation (1.4)), given by

$$\mathcal{K}f(t) \;=\; \int_0^1 K(t,s)f(s)\mathrm{d}s.$$

The aim of this chapter is to extend the notion of the multivariate (finite-dimensional) Mahalanobis distance (2.1) to the functional case when $x_1, x_2 \in L^2[0,1]$. Clearly, in view of (2.1), the inverse $\mathcal{K}^{-1}$ of the functional operator $\mathcal{K}$ should play some role in this extension if we want to keep a close analogy with the multivariate case. Unfortunately, such a direct approach utterly fails since, typically, $\mathcal{K}$ is not invertible in general as an operator, in the sense that there is no linear continuous operator $\mathcal{K}^{-1}$ such that $\mathcal{K}^{-1}\mathcal{K} = \mathcal{K}\mathcal{K}^{-1} = \mathbb{I}$, the identity operator.

To see the reason for this crucial difference between the finite and the infinite-dimensional cases, let us recall that some elementary linear algebra yields the following representations for Σx and $\Sigma^{-1}y$,

$$\Sigma x = \sum_{i=1}^{d} \lambda_i (e_i{}'x) e_i, \qquad \Sigma^{-1} y = \sum_{i=1}^{d} \frac{1}{\lambda_i} (e_i{}'y) e_i \tag{2.5}$$

where $\lambda_1, \ldots, \lambda_d$ are the, strictly positive, eigenvalues of Σ and $\{e_i, \ldots, e_d\}$ the corresponding orthonormal basis of eigenvectors.

In the functional case, the classical Karhunen-Loève Theorem (see, e.g., Ash and Gardner (2014)) provides $X(t) = \sum_j Z_j e_j(t)$ (in L^2 uniformly on t) where $\{e_j\}$ is the basis of orthonormal eigenfunctions of $\mathcal{K}$ and the $Z_j = \langle X, e_j \rangle_2$ are uncorrelated random variables with $\mathrm{var}(Z_j) = \lambda_j$, the eigenvalue of $\mathcal{K}$ corresponding to e_j. Then, we have

$$\mathcal{K}x = \mathcal{K}\Big(\sum_{i=1}^{\infty} \langle x, e_i \rangle_2 e_i \Big)$$

Note that the continuity of $K(s,t)$ implies $\int_0^1 \int_0^1 K(s,t)^2 \mathrm{d}s\mathrm{d}t < \infty$, thus $\mathcal{K}$ is in fact a compact, Hilbert-Schmidt operator. In addition, it is easy to check that $\int_0^1 \int_0^1 K(s,t)^2 \mathrm{d}s\mathrm{d}t = \sum_{i=1}^{\infty} \lambda_i^2$ so that, in particular, the sequence $\{\lambda_i\}$ converges to zero very quickly. As a consequence, there is no hope of keeping a direct analogy with (2.5) since

$$\mathcal{K}^{-1}x = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} \langle x, e_i \rangle_2 e_i \tag{2.6}$$

will not define in general a continuous operator with a finite norm. Still, for some particular functions $x = x(t)$ the series in (2.6) might be convergent. Hence we could use it formally to define the following template which, suitably modified, could lead

to a general, valid definition for a Mahalanobis-type distance between two functions x and m,

$$\widetilde{M}(x,m) = \left(\sum_{i=1}^{\infty} \frac{\langle x-m, e_i \rangle_2^2}{\lambda_i} \right)^{1/2}, \tag{2.7}$$

for all $x, m \in L^2[0,1]$ such that the series in (2.7) is finite. We are especially concerned with the case where x is a trajectory from a stochastic process $X(t)$ and m is the corresponding mean function. As we will see later on, this entails some especial difficulties.

The organization of this chapter

In the next section the connection of the Mahalanobis distance with RKHS theory is introduced, together with the proposed definition. In Section 2.3 some properties of the proposed distance are presented and compared with those of the original multivariate definition. Then, a consistent estimator is analyzed in Section 2.4. Finally, some numerical outputs corresponding to different statistical applications can be found in Section 2.5.

2.2 A new definition of Mahalanobis distance for functional data

Motivated by the previous considerations, Galeano et al. (2015) and Ghiglietti et al. (2017) have suggested two functional Mahalanobis-type distances, that we comment at the end of this section. These proposals are natural extensions to the functional case of the multivariate notion (2.1). Moreover, as suggested by the practical examples considered in both works, these options performed quite well in many cases. However, we believe that there is still some room to further explore the subject for the reasons we will explain below.

In this section we propose a further definition of a Mahalanobis-type distance, denoted M_α. Its most relevant features can be summarized as follows:

- M_α is also inspired in the natural template (2.7). The serious convergence issues appearing in (2.7) are solved by smoothing.
- M_α depends on a single, real, easy to interpret smoothing parameter α whose choice is not critical, in the sense that the distance has some stability with respect to α. Hence, it is possible to think of a cross-validation or bootstrap-based choice of α. In particular, no auxiliary weight function is involved in the definition.

- $M_\alpha(x, m)$ is a true metric which is defined for any given pair x, m of functions in $L^2[0, 1]$. Besides, it shares some invariance properties with the finite-dimensional counterpart (2.1).
- If $m(s) = \mathbb{E}[X(s)]$, the distribution of $M_\alpha(X, m)$ is explicitly known for Gaussian processes. In particular, $\mathbb{E}[M_\alpha^2(X, m)]$ and $\text{var}(M_\alpha^2(X, m))$ have explicit, relatively simple expressions.

The main contribution of this work is to show that the theory of Reproducing Kernel Hilbert Spaces (RKHS) provides a natural and useful framework to propose an extension of the Mahalanobis distance to the functional setting, satisfying the above mentioned properties.

2.2.1 RKHS's and the Mahalanobis distance

We rely on the second definition of RKHS presented in Section 1.2.1, where the starting element in the construction of the space is a positive semidefinite function $K(s, t)$, $s, t \in [0, 1]$. For our purposes, K will be the continuous positive definite covariance function of the process X that generates our functional data.

To see the connection with the Mahalanobis distance, instead of the whole stochastic process $X(s)$, $s \in [0, 1]$, let us consider a random vector $(X(t_1), \ldots, X(t_d))$. The covariance function $K(s, t)$ would be then replaced with the covariance matrix Σ whose (i, j)-entry is $K(t_i, t_j)$. From Moore-Aronszajn's Theorem we know that there exists a unique RKHS, $\mathcal{H}(\Sigma)$, in $\mathbb{R}^d$ whose reproducing kernel is Σ (see, Hsing and Eubank (2015, p.47–49) or Berlinet and Thomas-Agnan (2004, p. 19)).

From the definition of $\mathcal{H}_0(\Sigma)$ (Equation (1.1)) it is clear that, in this case, this space is just the image of the linear application defined by Σ, that is, it consists of the vectors that can be written as $x = \Sigma a$ for some $a \in \mathbb{R}^d$. Moreover, according to (1.2), the inner product between two elements $x = \Sigma a$ and $y = \Sigma b$ of this space is given by $\langle x, y\rangle_\Sigma = a'\Sigma b$. On the other hand, since $\mathcal{H}_0(\Sigma)$ is here a finite-dimensional space, it agrees with its completion $\mathcal{H}(\Sigma)$.

If we assume that Σ has full rank (if not, the generalized inverse should be used), this product can be rewritten as

$$\langle x, y\rangle_\Sigma \;=\; a'\Sigma b \;=\; a'\Sigma\Sigma^{-1}\Sigma b \;=\; x'\Sigma^{-1}y.$$

Then, the squared distance between two vectors $x, y \in \mathcal{H}(\Sigma)$ associated with this inner product can be expressed as

$$\|x - y\|_\Sigma^2 \;=\; \langle x - y, x - y\rangle_\Sigma \;=\; (x - y)'\Sigma^{-1}(x - y) = \sum_{i=1}^{d} \frac{((x - m)'e_i)^2}{\lambda_i}, \qquad (2.8)$$

where in the last equality we have used the second equation in (2.5).

We might summarize the above elementary discussion in the following statements:

(a) The RKHS distance $\|x-y\|_\Sigma$ in the RKHS associated with a finite-dimensional covariance operator, given by a positive definite matrix Σ, can be expressed as a simple sum involving the inverse eigenvalues of Σ, as shown in (2.8).

(b) Such RKHS distance coincides with the corresponding Mahalanobis distance between x and y.

At this point it is interesting to note that the above statement (a) can be extended to the infinite-dimensional case, as pointed out in the following lemma.

Lemma 2.1. *Let $\lambda_1 \geq \lambda_2 \geq \ldots$ be the positive eigenvalues of the integral operator associated with the kernel K. Let us denote by e_i the corresponding unit eigenfunctions. For $x \in \mathcal{H}(K)$,*

$$\|x\|_K^2 \;=\; \sum_{i=1}^{\infty} \frac{\langle x, e_i\rangle_2^2}{\lambda_i}, \tag{2.9}$$

and then the RKHS can be also rewritten as

$$\mathcal{H}(K) = \Big\{\; x \in L^2[0,1] \;:\; \sum_{i=1}^{\infty} \frac{\langle x, e_i\rangle_2^2}{\lambda_i} < \infty \;\Big\}.$$

In particular, the functions $\{\sqrt{\lambda_i} e_i\}$ are an orthonormal basis for $\mathcal{H}(K)$.

Proof. This result is just a rewording of the following theorem (already introduced in Section 1.2.1), whose proof can be found in Amini and Wainwright (2012):

Theorem.- Under the indicated conditions, the RKHS associated with K can be written

$$\mathcal{H}(K) = \Big\{ x \in L^2[0,1] \;:\; x = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i} e_i, \;\text{for}\; \sum_{i=1}^{\infty} a_i^2 < \infty \Big\}, \tag{2.10}$$

where the convergence of the series is in $L^2[0,1]$. This space is endowed with the inner product $\langle x, y\rangle_K \;=\; \sum_i a_i b_i$, where $x = \sum_i a_i \sqrt{\lambda_i} e_i$ and $y = \sum_i b_i \sqrt{\lambda_i} e_i$.

The result follows by noting that for any $x \in L^2[0,1]$ we can write

$$x \;=\; \sum_{i=1}^{\infty} \langle x, e_i\rangle_2 \, e_i \;=\; \sum_{i=1}^{\infty} \frac{\langle x, e_i\rangle_2}{\sqrt{\lambda_i}} \sqrt{\lambda_i} e_i.$$

Then, if the coefficients $\langle x, e_i\rangle_2$ tend to zero fast enough so that $\sum_i \langle x, e_i\rangle_2^2 \lambda_i^{-1} < \infty$, we have $x \in \mathcal{H}(K)$ and we get the expression (2.9) for $\|x\|_K^2$. □

This result sheds some light on the following crucial question: to what extent the formal expression (2.7) can be used to give a general definition of the functional Mahalanobis distance? In other words, for which functions $x \in L^2[0,1]$ does the series in (2.7) converge in L^2? The answer is clear in view of Lemma 2.1: expression (2.7) is well defined if and only if $x \in \mathcal{H}(K)$. This amounts to ask for a strong, very specific, regularity condition on x.

The bad news is that, as a consequence of a well-known result (see, e.g. Lukić and Beder (2001)), Cor. 7.1) if $X = X(s)$ is a Gaussian process with mean and covariance functions m and K, respectively, such that $m \in \mathcal{H}(K)$ and $\mathcal{H}(K)$ is infinite-dimensional, then $\mathbb{P}(X \in \mathcal{H}(K)) = 0$, whenever the probability $\mathbb{P}$ is assumed to be complete. Hence, with probability one, expression (2.7) is not convergent for the trajectories drawn from the stochastic process X.

2.2.2 The proposed definition

In view of the discussion above (see statement (b) before Lemma 2.1), it might seem natural to define the (square) Mahalanobis functional distance between a trajectory $x = x(s)$ of the process $X(s)$ and a function $m \in L^2[0,1]$ by $M^2(x,m) = \|x - m\|_K^2$. However, this idea does not work since, as indicated above, the trajectories $x = x(s)$ of $X = X(s)$ do not belong to $\mathcal{H}(K)$ with probability one.

This observation suggest us the simple strategy we will follow here: given two functions $x, m \in L^2[0,1]$, just approximate them by two other functions $x_\alpha, m_\alpha \in \mathcal{H}(K)$ and calculate the distance $\|x_\alpha - m_\alpha\|_K$. It only remains to decide how to obtain the RKHS approximations x_α and m_α. One could think of taking x_α as the "closest" function to x in $\mathcal{H}(K)$ but this approach also fails since $\mathcal{H}(K)$ is dense in $L^2[0,1]$ whenever all λ_i are strictly greater than zero (Remark 4.9 of Cucker and Zhou (2007)). Thus, every function $x \in L^2[0,1]$ can be arbitrarily well approximated by functions in $\mathcal{H}(K)$.

This leads us in a natural way to the following penalization approach. Let us fix a penalization parameter $\alpha > 0$. Given any $x \in L^2[0,1]$, define

$$x_\alpha = \underset{f \in \mathcal{H}(K)}{\operatorname{argmin}} \|x - f\|_2^2 + \alpha \|f\|_K^2. \tag{2.11}$$

As we will see below, the "penalized projection" x_α is well-defined. In fact it admits a relatively simple closed form. Finally, the definition we propose for the functional Mahalanobis distance is

$$M_\alpha(x,m) = \|x_\alpha - m_\alpha\|_K. \tag{2.12}$$

As mentioned, given a realization x of the stochastic process we have relatively simple expressions for both the smoothed trajectory x_α and the proposed distance. In the next result we summarize these expressions.

Proposition 2.2. *Given a second order process with covariance K, we denote as $\mathcal{K}$ the integral covariance operator of Equation* (1.4) *associated with K. Then the smoothed trajectories x_α defined in (2.11) satisfy the following basic properties:*

(a) Let $\mathbb{I}$ be the identity operator on $L^2[0,1]$. Then, $\mathcal{K}+\alpha\mathbb{I}$ is invertible and

$$x_\alpha = (\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}x = \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j+\alpha}\langle x, e_j\rangle_2\, e_j, \tag{2.13}$$

where λ_j, $j=1,2,\cdots$ are the eigenvalues of $\mathcal{K}$ (which are strictly positive under our assumptions) and e_j stands for the unit eigenfunction of $\mathcal{K}$ corresponding to λ_j.

(b) Denoting as $\mathcal{K}^{1/2}$ the square root operator defined by $(\mathcal{K}^{1/2})^2=\mathcal{K}$, the norm of x_α in $\mathcal{H}(K)$ satisfies

$$\|x_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\ \langle x, e_j\rangle_2^2 = \|\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}x\|_2^2, \tag{2.14}$$

and therefore,

$$M_\alpha(x,m)^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\ \langle x-m, e_j\rangle_2^2.$$

Proof. (a) The fact that $\mathcal{K}+\alpha\mathbb{I}$ is invertible is a consequence of Theorem 8.1 in Gohberg and Goldberg (2013, p. 183). The expression for x_α follows straightforwardly from Proposition 8.6 of Cucker and Zhou (2007, p.139). Moreover, expression (8.4) in Gohberg and Goldberg (2013, p. 184) yields

$$(\mathcal{K}+\alpha\mathbb{I})^{-1}\, y = \frac{1}{\alpha}(\mathbb{I}-\mathcal{K}_1)y, \tag{2.15}$$

where

$$\mathcal{K}_1 y = \sum_{j=1}^{\infty} \frac{\lambda_j}{\alpha+\lambda_j}\langle y, e_j\rangle_2\, e_j. \tag{2.16}$$

Then, using the Spectral theorem for compact and self-adjoint operators (for instance Theorem 2 of Chapter 2 of Cucker and Smale (2001)) we get:

$$x_\alpha = (\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}x = \frac{1}{\alpha}\sum_{j=1}^{\infty}\left(1-\frac{\lambda_j}{\alpha+\lambda_j}\right)\lambda_j\langle x, e_j\rangle_2\, e_j = \sum_{j=1}^{\infty}\frac{\lambda_j}{\alpha+\lambda_j}\langle x, e_j\rangle_2\, e_j.$$

(b) In Lemma 2.1 we have seen that $\sqrt{\lambda_j}e_j$ is an orthonormal basis of $\mathcal{H}(K)$. Then (2.13) together with Parseval's identity (in $\mathcal{H}(K)$) imply

$$\|x_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\ \langle x, e_j\rangle_2^2.$$

Moreover, from the Spectral Theorem $\mathcal{K}^{1/2}(x) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\langle x, e_i\rangle_2\, e_i$, then using (2.15) and (2.16), $\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1} = \alpha^{-1}\mathcal{K}^{1/2}(\mathbb{I}-\mathcal{K}_1)$, and also

$$\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}x = \sum_{j=1}^{\infty} \frac{\sqrt{\lambda_j}}{\lambda_j+\alpha}\langle x, e_j\rangle_2\, e_j.$$

Then, using again Parseval's identity (but now in $L^2[0,1]$) we get

$$\|\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}x\|_2^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\ \langle x, e_j\rangle_2^2 = \|x_\alpha\|_K^2.$$

□

Corollary 2.3. *The expression M_α given in* (2.12) *defines a metric in $L^2[0,1]$.*

Proof. This result is a direct consequence of Proposition 2.2. Indeed, from expression (2.13), the transformation $x \mapsto x_\alpha$ form $L^2[0,1]$ to $\mathcal{H}(K)$ is injective (since the coefficients $\langle x, e_i\rangle_K$ completely determine x). This, together with the fact that $\|\cdot\|_K$ is a norm, yields the result. □

Remark. The expression $x_\alpha = (\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}x$ obtained in the first part of Proposition 2.2 has an interesting intuitive meaning: the transformation $x \mapsto \mathcal{K}x$ takes first the function $x \in L^2[0,1]$ to the space $\mathcal{H}(K)$, made of much nicer functions, with Fourier coefficients $\langle x, e_i\rangle_2$ converging quickly to zero, since we must have $\sum_{i=1}^{\infty}\langle x, e_i\rangle_2^2/\lambda_i < \infty$; see (2.10). Then, after this "smoothing step", we perform an "approximation step" by applying the inverse operator $(\mathcal{K}+\alpha\mathbb{I})^{-1}$, in order the get, as a final output, a function x_α that is both, close to x and smoother than x. Note also that the operator $(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}$ is compact. Thus, if we assume that the original trajectories are uniformly bounded in $L^2[0,1]$, the final result of applying on these trajectories the transformation $x \mapsto x_\alpha$ would be to take them to a pre-compact set of $L^2[0,1]$. This is very convenient from different points of view (beyond our specific needs here), in particular when one needs to find a convergent subsequence inside a given bounded sequence of x_α's.

2.2.3 Some earlier proposals

Motivated by the heuristic spectral version (2.7) of the Mahalanobis distance, Galeano et al. (2015) proposed the following definition, that avoids the convergence problems of the series in (2.7) (provided that $\lambda_i > 0$) at the expense of introducing a sort of smoothing parameter $k \in \mathbb{N}$,

$$d_{FM}^k(x,m) \;=\; \left(\sum_{i=1}^{k} \frac{\langle x-m, e_i\rangle_2^2}{\lambda_i}\right)^{1/2}. \tag{2.17}$$

We keep the notation d^k_{FM} used in Galeano et al. (2015). Let us note that $d^k_{FM}(x,m)$ is a semi-distance, since it lacks the identifiability condition $d^k_{FH}(x,m) = 0 \Rightarrow x = m$. The applications of d^k_{FM} considered by these authors focus mainly on supervised classification. While this proposal is quite simple and natural, it suffers from some insufficiencies when considered from the theoretical point of view. The most important one is the fact that the series (2.17) is divergent, with probability one, whenever x is a trajectory of a Gaussian process with mean function m and covariance function K (as we have just seen). So, d^k_{FM} is defined in terms of the k-th partial sum of a divergent series. As a consequence, one may expect that the definition might be strongly influenced by the choice of k. From the discussion above it is clear that in practice this effect might not noticed if x is replaced with a smoothed trajectory but, in that case, the smoothing procedure should be incorporated to the definition.

Another recent proposal is due to Ghiglietti et al. (2017). The idea is also to modify the template (2.7) to deal with the convergence issues. In this case, the suggested definition is

$$d_p(x,m) \;=\; \left(\int_0^\infty \sum_{i=1}^{\infty} \frac{\langle x-m, e_i\rangle_2^2}{e^{\lambda_i c}}\, g(c;p)\mathrm{d}c\right)^{1/2}, \tag{2.18}$$

where $p > 0$ and $g(c;p)$ is a weight function such that $g(0;p) = 1$, g is non-increasing and non-negative and $\int_0^\infty g(c;p)\mathrm{d}c = p$. Moreover, for any $c > 0$, $g(c;p)$ is assumed to be non-decreasing in p with $\lim_{p\to\infty} g(c;p) = 1$. This definition does not suffer from any problem derived from degeneracy but, still, it depends on two smoothing functions: the exponential in the denominator of (2.18) and the weighting function $g(c;p)$. As pointed out also in Ghiglietti et al. (2017), a more convenient expression for (2.18) is given by the following weighted version of the template, formal definition (2.7),

$$d_p(x,m) \;=\; \left(\sum_{i=1}^{\infty} \frac{\langle x-m, e_i\rangle_2^2}{\lambda_i}\, h_i(p)\right)^{1/2}, \tag{2.19}$$

where $h_i(p) = \int_0^\infty \lambda_i e^{-\lambda_i c} g(c;p)\mathrm{d}c$.

The applications of (2.18) offered in Ghiglietti et al. (2017) and Ghiglietti and Paganoni (2017) deal with hypotheses testing for two-sample problems of type $H_0 : m_1 = m_2$.

2.3 Some properties of the functional Mahalanobis distance

In this section we analyze in detail and prove some of the features of M_α we have anticipated above. In what follows $X = X(s)$, with $s \in [0,1]$ will stand for a second-order stochastic process with continuous trajectories and continuous mean and covariance functions, denoted by $m = m(s)$ and $K = K(s,t)$, respectively.

2.3.1 Invariance

In the finite dimensional case, one appealing property of the Mahalanobis distance is the fact that it does not change if we apply a non-singular linear transformation to the data. Then, the invariance for a large class of linear operators appears also as a desirable property for any extension of the Mahalanobis distance to the functional case. Here, we prove invariance with respect to operators preserving the norm. We recall that an operator L is an isometry if it maps $L^2[0,1]$ to $L^2[0,1]$ and $\|f\|_2 = \|Lf\|_2$. In this case, it holds $L^*L = \mathbb{I}$, where L^* stands for the adjoint of L.

Theorem 2.4. *Let L be an isometry on $L^2[0,1]$. Then, $M_\alpha(x,m) = M_\alpha(Lx,Lm)$ for all $\alpha > 0$, where M_α was defined in* (2.12).

Proof. Let $\mathcal{K}_L$ be the covariance operator of the process LX. The first step of the proof is to show that $\mathcal{K}_L = L\mathcal{K}L^*$. It is enough to prove that for all $f, g \in L^2[0,1]$, it holds $\langle \mathcal{K}_L f, g\rangle_2 = \langle L\mathcal{K}L^* f, g\rangle_2$. Observe that

$$\begin{aligned}\langle \mathcal{K}_L f, g\rangle_2 &= \int_0^1 \mathcal{K}_L f(t) g(t) \mathrm{d}t \\ &= \int_0^1 \int_0^1 \mathbb{E}\big[\big(LX(s) - Lm(s)\big)\big(LX(t) - Lm(t)\big)\big] f(s) g(t) \mathrm{d}s\mathrm{d}t.\end{aligned}$$

Then, using Fubini's theorem and the definition of the adjoint operator:

$$\langle \mathcal{K}_L f, g\rangle_2 = \mathbb{E}\big[\langle L(X-m), f\rangle_2 \langle L(X-m), g\rangle_2\big] = \mathbb{E}\big[\langle X-m, L^*f\rangle_2 \langle X-m, L^*g\rangle_2\big].$$

Analogously, we also have

$$\langle L\mathcal{K}L^* f, g\rangle_2 = \langle \mathcal{K}L^* f, L^* g\rangle_2 = \mathbb{E}\big[\langle X-m, L^*f\rangle_2 \langle X-m, L^*g\rangle_2\big].$$

From the last two equations we conclude $\mathcal{K}_L = L\mathcal{K}L^*$.

The second step of the proof is to observe that the eigenvalues λ_j of $\mathcal{K}_L$ are the same as those of $\mathcal{K}$, and the unit eigenfunction v_j of $\mathcal{K}_L$ for the eigenvalue λ_j is given by $v_j = Le_j$, where e_j is the unit eigenfunction corresponding to λ_j of $\mathcal{K}$. Indeed, using $L^*L = \mathbb{I}$ we have

$$\mathcal{K}_L v_j = L\mathcal{K}L^* v_j = L\mathcal{K}L^* L e_j = \lambda_j L e_j = \lambda_j v_j, \quad j = 1, 2, \ldots$$

Then, by (2.14) and using that L is an isometry,

$$\begin{aligned}M_\alpha(Lx, Lm) &= \|(Lx - Lm)_\alpha\|_{K_L} = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2} \langle Lx - Lm, Le_j\rangle_2^2 \\ &= \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2} \langle x - m, e_j\rangle_2^2 = M_\alpha(x, m).\end{aligned}$$

□

The family of isometries on $L^2[0,1]$ contains some interesting examples. For instance, all the symmetries and translations are isometries, as well as changes between orthonormal bases. Thus, this distance does not depend on the basis on which the data are represented.

2.3.2 Distribution for Gaussian processes

We have mentioned in the introduction of the chapter that the squared Mahalanobis distance to the mean for Gaussian data has a χ^2 distribution with d degrees of freedom, where d is the dimension of the data. In the functional case, the distribution of $M_\alpha(X,m)^2$ for a Gaussian process X equals that of an infinite linear combination of independent χ^2_1 random variables. We prove this fact in the following result and its corollary, and also give explicit expressions for the expectation and the variance of $M_\alpha(X,m)^2$.

Theorem 2.5. *Let $\{X(s) : s \in [0,1]\}$ be an L^2 Gaussian process with mean m and continuous positive definite covariance function K. Let $\lambda_1, \lambda_2, \cdots$ be the eigenvalues of $\mathcal{K}$ and let $e_1, e_2, \ldots$ be the corresponding unit eigenfunctions.*

(a) The squared Mahalanobis distance to the origin satisfies

$$M_\alpha(X,0)^2 = \|X_\alpha\|_K^2 = \sum_{j=1}^{\infty} \beta_j Y_j, \tag{2.20}$$

where $\beta_j = \lambda_j^2(\lambda_j+\alpha)^{-2}$ and Y_j, $j = 1,2,\cdots$, are non-central $\chi^2_1(\gamma_j)$ random variables with non-centrality parameter $\gamma_j = \mu_j^2/\lambda_j$, where $\mu_j := \langle m, e_j\rangle_2$.

(b) We have

$$\mathbb{E}\big[M_\alpha(X,0)^2\big] = \sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\lambda_j+\alpha)^2}\left(1+\frac{\mu_j^2}{\lambda_j}\right),$$

and

$$\operatorname{var}\big(M_\alpha(X,0)^2\big) = 2\sum_{j=1}^{\infty} \frac{\lambda_j^4}{(\lambda_j+\alpha)^4}\left(1+\frac{2\mu_j^2}{\lambda_j}\right).$$

Proof. (a) Using (2.14), $\|X_\alpha\|_K^2 = \sum_{j=1}^{\infty} \beta_j Y_j$, where $\beta_j = \lambda_j^2(\lambda_j+\alpha)^{-2}$ and $Y_j = \lambda_j^{-1}\langle X, e_j\rangle_2^2$. Since the process is Gaussian the variables $\lambda_j^{-1/2}\langle X, e_j\rangle$ are independent with normal distribution, mean $\lambda_j^{-1/2}\mu_j$ and variance 1 (see Ash and Gardner (2014), p. 40). The result follows.

(b) It is easy to see that the partial sums in (2.20) form a sub-martingale with respect to the natural filtration $\sigma(Y_1, \ldots, Y_N)$,

$$\mathbb{E}\Big[\sum_{j=1}^{N+1} \beta_j Y_j \,\Big|\, Y_1, \ldots, Y_N\Big] = \beta_{N+1}\mathbb{E}[Y_{N+1}] + \sum_{j=1}^{N} \beta_j Y_j \geq \sum_{j=1}^{N} \beta_j Y_j.$$

Moreover, if $\bar{\lambda} := \sup_j \lambda_j$, which is always finite,

$$\sup_N \mathbb{E}\Big[\sum_{j=1}^{N+1} \beta_j Y_j\Big] = \sum_{j=1}^{\infty} \frac{\lambda_j(\lambda_j + \mu_j^2)}{(\lambda_j + \alpha)^2} \leq \frac{\bar{\lambda}}{\alpha^2}\Big(\sum_{j=1}^{\infty} \lambda_j + \sum_{j=1}^{\infty} \mu_j^2\Big) < \infty,$$

because $m \in L^2[0,1]$ and $\sum_{j=1}^{\infty} \lambda_j = \int_0^1 K(t,t)\mathrm{d}t < \infty$ (see e.g. Cucker and Smale (2001), Corollary 3, p. 34). Now, Doob's convergence theorem implies $\sum_{j=1}^{N} \beta_j Y_j \to \sum_{j=1}^{\infty} \beta_j Y_j$ a.s. as $N \to \infty$, and Monotone Convergence theorem yields the expression for the expectation of $M_\alpha(X,0)^2$.

The proof for the variance is fairly similar. Using Jensen's inequality, we deduce

$$\mathbb{E}\Big[\Big(\sum_{j=1}^{N+1} \beta_j \left(Y_j - \mathbb{E}Y_j\right)\Big)^2 \,\Big|\, Y_1, \ldots, Y_N\Big] \geq \Big(\sum_{j=1}^{N} \beta_j \left(Y_j - \mathbb{E}Y_j\right)\Big)^2.$$

Moreover, since the variables Y_j are independent:

$$\begin{aligned}\sup_N \mathbb{E}\Big[\sum_{j=1}^{N} \beta_j \left(Y_j - \mathbb{E}Y_j\right)\Big]^2 &= \sum_{j=1}^{\infty} \beta_j^2 \mathrm{var}(Y_j) = 2\sum_{j=1}^{\infty} \frac{\lambda_j^3(\lambda_j + 2\mu_j^2)}{(\lambda_j + \alpha)^4} \\ &\leq \frac{2\bar{\lambda}^3}{\alpha^4}\Big(\sum_{j=1}^{\infty} \lambda_j + 2\sum_{j=1}^{\infty} \mu_j^2\Big) < \infty.\end{aligned}$$

Then, $\big(\sum_{j=1}^{N+1} \beta_j(Y_j - \mathbb{E}Y_j)\big)^2 \to \big(\sum_{j=1}^{\infty} \beta_j(Y_j - \mathbb{E}Y_j)\big)^2$ a.s., as $N \to \infty$, and using Monotone Convergence theorem,

$$\mathrm{var}\Big(M_\alpha(X,0)^2\Big) = \lim_{N\to\infty} \mathrm{var}\Big(\sum_{j=1}^{N} \beta_j Y_j\Big) = 2\sum_{j=1}^{\infty} \frac{\lambda_j^4}{(\lambda_j + \alpha)^4}\left(1 + \frac{2\mu_j^2}{\lambda_j}\right).$$

□

When we compute the squared Mahalanobis distance to the mean the expressions above simplify because $\mu_j = 0$ for each j, and then we have the following corollary.

Corollary 2.6. *Under the same assumptions of Theorem 2.5,* $M_\alpha(X,m)^2 = \sum_{j=1}^{\infty} \beta_j Y_j$, *where* $\beta_j = \lambda_j^2(\lambda_j + \alpha)^2$ *and* $Y_1, Y_2, \ldots$ *are independent* χ_1^2 *random variables. Moreover, the expectation* $\mathbb{E}[M_\alpha(X,m)^2]$ *equals* $\sum_{j=1}^{\infty} \lambda_j^2(\lambda_j + \alpha)^{-2}$ *and* $\mathrm{var}\left(M_\alpha(X,m)^2\right) = 2\sum_{j=1}^{\infty} \lambda_j^4(\lambda_j + \alpha)^{-4}$.

2.3.3 Stability with respect to α

Our definition of distance depends on a regularization parameter $\alpha > 0$. In this subsection we prove the continuity of M_α with respect to the tuning parameter α. The proof of the main result requires the following auxiliary lemma, which has been adapted from Corollary 8.3 in Gohberg and Goldberg (2013), p. 71. Recall that given a bounded operator $A : \mathcal{H} \to \mathcal{H}$ on a Hilbert space $\mathcal{H}$ we can define the norm

$$\|A\|_{\mathcal{L}} := \sup\{\|Ax\|_{\mathcal{H}} : \|x\|_{\mathcal{H}} \leq 1\}.$$

Lemma 2.7. *Let $A_j : \mathcal{H} \to \mathcal{H}$, $j = 1, 2, \ldots$, be a sequence of bounded invertible operators on a Hilbert space $\mathcal{H}$ which converges in norm $\|\cdot\|_{\mathcal{L}}$ to another operator A, and such that $\sup_j \|A_j^{-1}\|_{\mathcal{L}} < \infty$. Then A is also invertible, and $\|A_j^{-1} - A^{-1}\|_{\mathcal{L}} \to 0$, as $j \to \infty$.*

We apply the preceding lemma in the proof of the following result.

Proposition 2.8. *Let α_j be a sequence of positive real numbers such that $\alpha_j \to \alpha > 0$, as $j \to \infty$. Then, $\|X_{\alpha_j}\|_K \to \|X_\alpha\|_K$ a.s. as $j \to \infty$.*

Proof. Note that by Proposition 2.2(b), Equation (2.14), we have

$$\begin{aligned}\left|\|X_{\alpha_j}\|_K - \|X_\alpha\|_K\right| &\leq \|\mathcal{K}^{1/2}(\mathcal{K} + \alpha_j\mathbb{I})^{-1}X - \mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}X\|_2 \\ &\leq \|\mathcal{K}^{1/2}\|_{\mathcal{L}} \, \|(\mathcal{K} + \alpha_j\mathbb{I})^{-1} - (\mathcal{K} + \alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \, \|X\|_2.\end{aligned}$$

But it holds

$$\|(\mathcal{K} + \alpha_j\mathbb{I}) - (\mathcal{K} + \alpha\mathbb{I})\|_{\mathcal{L}} = |\alpha_j - \alpha| \to 0, \quad \text{as } j \to \infty,$$

and $\sup_j \|(\mathcal{K} + \alpha_j\mathbb{I})^{-1}\|_{\mathcal{L}} \leq \inf_j \alpha_j < \infty$ (see Gohberg and Goldberg (2013), (1.14), p. 228). Therefore, $\|(\mathcal{K} + \alpha_j\mathbb{I})^{-1} - (\mathcal{K} + \alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \to 0$, as $j \to \infty$, by Lemma 2.7. □

Observe that Proposition 2.8 implies the point convergence of the sequence of distribution functions of $M_{\alpha_j}(X, m)$ to that of $M_\alpha(X, m)$. This fact in turn implies the point convergence of the corresponding quantile functions.

2.4 A consistent estimator of the functional Mahalanobis distance

Given a sample $x_1(s), \ldots, x_n(s)$ of realizations of the stochastic process $X(s)$, we want to estimate the Mahalanobis distance between any trajectory of the process X and the

mean function m in a consistent way. Let $\bar{x}(s) = n^{-1}\sum_{i=1}^{n} x_i(s)$ be the sample mean and let

$$\widehat{K}(s,t) = \frac{1}{n}\sum_{i=1}^{n}\big(x_i(s)-\bar{x}(s)\big)\big(x_i(t)-\bar{x}(t)\big)$$

be the sample covariance function. The function $\widehat{K}$ defines the sample covariance operator $\widehat{\mathcal{K}}f(\cdot) = \int_0^1 \widehat{K}(\cdot,t)f(t)\mathrm{d}t$.

Define the following estimator for $M_\alpha(x,m)$:

$$\widehat{M}_{\alpha,n}(x,\bar{x}) := \|\widehat{x}_\alpha - \bar{x}_\alpha\|_{\widehat{K}_n}, \tag{2.21}$$

where $\widehat{x}_\alpha = (\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\widehat{\mathcal{K}}x$ and $\bar{x}_\alpha = (\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\widehat{\mathcal{K}}\bar{x}$.

In the following lemma we establish the consistency (in the operator norm) of $\widehat{\mathcal{K}}$ as an estimator of $\mathcal{K}$, as a preliminary step to show the consistency of $\widehat{M}_{\alpha,n}$.

Lemma 2.9. *Suppose that* $\mathbb{E}\|X\|_2^2 < \infty$. *Then* $\|\bar{x}-m\|_2 \to 0$, $\|\widehat{K}-K\|_{L^2([0,1]\times[0,1])} \to 0$ *and* $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{\mathcal{L}} \to 0$, *a.s. as* $n\to\infty$.

Proof. Mourier's SLLN (see e.g. Theorem 4.5.2 in Laha and Rohatgi (1979), p. 452) implies directly $\|\bar{x}-m\|_2 \to 0$ since $(\mathbb{E}\|X\|_2)^2 \le \mathbb{E}\|X\|_2^2 < \infty$ and $L^2[0,1]$ is a separable Banach space.

Consider the process $Z(s,t) = X(s)X(t)$. Then, $Z \in L^2([0,1]\times[0,1])$ and this is also a separable Banach space. Therefore, if $z_i(s,t) = x_i(s)x_i(t)$, $\bar{z} = n^{-1}z_i(s,t)$, and $m_z(s,t) = \mathbb{E}[X(s)X(t)]$, using again Mourier's SLLN we have

$$\|\bar{z} - m_z\|_{L^2([0,1]\times[0,1])} \to 0, \quad \text{a.s.}, \quad n\to\infty,$$

and also, since $\widehat{K}(s,t) = \bar{z}(s,t) - \bar{x}(s)\bar{x}(t)$, $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{HS} \to 0$ a.s., where $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{HS} = \|\widehat{K}-K\|_{L^2([0,1]\times[0,1])}$ stands for the Hilbert-Schmidt norm of the operator $\widehat{\mathcal{K}}-\mathcal{K}$.

Finally, for any $x \in L^2[0,1]$,

$$\|(\widehat{\mathcal{K}}-\mathcal{K})x\|_2^2 \;=\; \int_0^1 \langle \widehat{K}(t,\cdot)-K(t,\cdot),x\rangle_2^2\mathrm{d}t \;\le\; \|x\|_2^2\,\|\widehat{K}-K\|^2_{L^2([0,1]\times[0,1])}.$$

Thus, in particular, the operator norm is smaller than the Hilbert-Schmidt norm and we have $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{\mathcal{L}} \le \|\widehat{K}-K\|_{L^2([0,1]\times[0,1])} \to 0$ a.s. as $n\to\infty$. □

As already mentioned, by the square root of an operator F we mean the operator G such that $G^2 = F$.

Theorem 2.10. *If* $\mathbb{E}\|X\|_2^2 < \infty$, *then* $\widehat{M}_{\alpha,n}(X,\bar{x}) \to M_\alpha(X,m)$ *a.s., as* $n\to\infty$.

Proof. From Proposition 2.2(b), Eq (2.14), we have $\widehat{M}_{\alpha,n}(X,\bar{x}) = \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}(X-\bar{x})\|_2$. Therefore,

$$\begin{aligned}\left|\widehat{M}_{\alpha,n}(X,\bar{x}) - M_\alpha(X,m)\right| &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}(X-\bar{x}) - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}(X-m)\|_2 \\ &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}}\ \|\bar{x}-m\|_2 + \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}}\ \|X-m\|_2.\end{aligned}$$

By Lemma 2.9, $\|\bar{x}-m\|_2$ goes to zero a.s. as $n\to\infty$. Besides, $\|X-m\|_2$ is almost sure bounded. As a consequence, it is enough to show that $\|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \to 0$ a.s. For that purpose, observe that

$$\begin{aligned}&\|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \\ &\leq \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - \widehat{\mathcal{K}}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} + \|\widehat{\mathcal{K}}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \\ &\leq \|\widehat{\mathcal{K}}^{1/2}\|_{\mathcal{L}}\ \|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - (\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} + \|\widehat{\mathcal{K}}^{1/2} - \mathcal{K}^{1/2}\|_{\mathcal{L}}\ \|(\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}}.\end{aligned}$$

Therefore, to end the proof we will show that $\|\widehat{\mathcal{K}}^{1/2} - \mathcal{K}^{1/2}\|_{\mathcal{L}} \to 0$ a.s. as $n\to\infty$ and $\|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - (\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \to 0$ a.s. as $n\to\infty$. Since the square root is a continuous function in $[0,\infty)$, the first result follows from part one of Theorem VIII.20 of Reed and Simon (1980). The requirement of the function vanishing at infinity is irrelevant here since, from Lemma 2.9 we know $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{\mathcal{L}} \to 0$ a.s. as $n\to\infty$, which in particular implies that there exist a bound on the norm of operators $\widehat{\mathcal{K}}$.

Finally, observe that $\|\widehat{\mathcal{K}}-\mathcal{K}\|_{\mathcal{L}} = \|(\widehat{\mathcal{K}}+\alpha\mathbb{I}) - (\mathcal{K}+\alpha\mathbb{I})\|_{\mathcal{L}} \to 0$ a.s., and we also have $\sup_n \|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \leq \alpha^{-1} < \infty$. Then, Lemma 2.7 implies $\|(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1} - (\mathcal{K}+\alpha\mathbb{I})^{-1}\|_{\mathcal{L}} \to 0$ a.s. as $n\to\infty$. □

Corollary 2.11. *In fact, the result is true when measuring the distance between the mean and any function f in $L^2[0,1]$, that is, $\widehat{M}_{\alpha,n}(f,\bar{x}) \to M_\alpha(f,m)$ a.s., as $n\to\infty$.*

Putting together Theorems 2.10 and 2.5 we obtain the asymptotic distribution of $\widehat{M}_{\alpha,n}$:

Corollary 2.12. *Under the assumptions of Theorem 2.10 and Corollary 2.6, and with the same notation, $\widehat{M}_{\alpha,n}(X,\bar{x})$ converges in distribution to $\sum_{j=1}^{\infty}\beta_j Y_j$, where $\beta_j = \lambda_j^2(\lambda_j+\alpha)^{-2}$ and $Y_1, Y_2, \ldots$ are independent χ_1^2 random variables.*

We can also prove another consistency result involving the distances between the sample and the population means, which could be useful for doing inference on the mean.

Theorem 2.13. *If $\mathbb{E}\|X\|_2^2 < \infty$, and with the same notation of Theorem 2.5, it holds*

(a) for $Y_1, Y_2, \ldots$ independent χ_1^2 random variables,

$$\sqrt{n}\ \widehat{M}_{\alpha,n}(\bar{x},m) \ \xrightarrow{d}\ \Big(\sum_{j=1}^{\infty}\frac{\lambda_j^2}{(\lambda_j+\alpha)^2}Y_j\Big)^{\frac{1}{2}}, \tag{2.22}$$

(b) if $m \neq m_0 \in L^2[0,1]$,

$$\sqrt{n}\left(\widehat{M}_\alpha(\bar{x}, m_0) - \widehat{M}_\alpha(m, m_0)\right) \xrightarrow{d} M_\alpha(m, m_0)^{-1}\Big(\sum_{j=1}^{\infty} \frac{\lambda_j^3}{(\lambda_j+\alpha)^4}\langle m - m_0, e_j\rangle_2^2\Big)^{1/2} W,$$

with W *a standard Gaussian variable.*

Proof. (a) We can rewrite the left-hand side of Equation (2.22) as,

$$\sqrt{n}\ \widehat{M}_{\alpha,n}(\bar{x}, m) \ = \ \sqrt{n}(\widehat{M}_{\alpha,n}(\bar{x}, m) - M_\alpha(\bar{x}, m)) + \sqrt{n} M_\alpha(\bar{x}, m). \tag{2.23}$$

Now, from Equation (2.12) and Proposition 2.2, we have

$$\begin{aligned}
&\sqrt{n}\,|\widehat{M}_{\alpha,n}(\bar{x}, m) - M_\alpha(\bar{x}, m)| \\
&\quad \leq \ \sqrt{n}\,\|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}} + \alpha\mathbb{I})^{-1}(\bar{x} - m) - \mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}(\bar{x} - m)\|_2 \\
&\quad \leq \ \|\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}} + \alpha\mathbb{I})^{-1} - \mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}\|_{\mathcal{L}}\ \|\sqrt{n}(\bar{x} - m)\|_2.
\end{aligned}$$

As a part of the proof of Theorem 2.10 we have seen that the first norm in the right-hand side goes to zero a.s. as $n \to \infty$. From the Functional Central Limit Theorem (e.g., Theorem 8.1.1 of Hsing and Eubank (2015)), $\sqrt{n}(\bar{x} - m)$ converges in distribution in $L^2[0,1]$ to a Gaussian stochastic process Z with zero mean and covariance operator $\mathcal{K}$. Since the norm is a continuous function in this space, by the continuous mapping theorem the second term converges in distribution to the random variable $\|Z\|_2$. Thus, by Slutsky's theorem, the distribution of the product goes to zero, and this convergence holds also in probability since the limit is a constant.

We can rewrite the remaining term of Equation (2.23) as,

$$\sqrt{n} M_\alpha(\bar{x}, m) \ = \ \sqrt{n}\ \|\mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}(\bar{x} - m)\|_2 \ = \ \Big\|\sqrt{n}\ \Big(\frac{1}{n}\sum_{i=1}^{n} \chi_{i,\alpha} - \mu_\alpha\Big)\Big\|_2,$$

where we denote $\chi_{i,\alpha} = \mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1} x_i$ and $\mu_\alpha = \mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1} m$. Since $\mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}$ is a bounded linear operator and the process X is Bochner-integrable ($\mathbb{E}\|X\|_2 < \infty$), the expectation and the operator commute, that is,

$$\mathbb{E}[\chi_\alpha] \ = \ \mathbb{E}[\mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1} X] \ = \ \mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}\mathbb{E}[X] \ = \ \mu_\alpha.$$

Therefore, we can use again the Functional Central Limit Theorem with $\chi_{\alpha,i}$ and μ_α, since

$$\mathbb{E}\|\chi_\alpha\|_2^2 \ \leq \ \|\mathcal{K}^{1/2}(\mathcal{K} + \alpha\mathbb{I})^{-1}\|_{\mathcal{L}}^2\ \mathbb{E}\|X\|_2^2 \ < \ \infty,$$

which gives us that $\sqrt{n} M_\alpha(\bar{x}, m)$ converges in distribution to $\|\xi\|_2$, ξ being a random element with zero mean and whose covariance operator is the same as that of χ_α.

Using the same reasoning as at beginning of the proof of Theorem 2.4 and denoting as A^* the adjoint of the operator A, the covariance operator of $\chi_{\alpha,1}$ is given by

$$\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}[\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}]^* \;=\; \mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}(\mathcal{K}+\alpha\mathbb{I})^{-1}\mathcal{K}^{1/2},$$

since both $\mathcal{K}^{1/2}$ and $(\mathcal{K}+\alpha\mathbb{I})^{-1}$ are self-adjoint operators (for instance, Theorem 3.35 and Problem 3.32 of (Kato, 2013, Chapter 5) and Proposition 2.4 of (Conway, 1990, Chapter X)). Now since ξ is a zero-mean Gaussian process with compact covariance operator, it has an associated orthonormal basis of eigenfunctions (Spectral theorem for compact and self-adjoint operators, for instance Theorem 2 of Chapter 2 of Cucker and Smale (2001)). This operator has the same eigenfunctions as $\mathcal{K}$ and its eigenvalues are $\lambda_j^2(\lambda_j+\alpha)^{-2}$. Thus, using its Karhunen-Loève representation we get

$$\|\xi\|_2 \;=\; \|\sum_{j=1}^{\infty} Z_j e_j\|_2 \;=\; \Big(\sum_{j=1}^{\infty} Z_j^2\Big)^{\frac{1}{2}},$$

where e_j are the eigenfunctions of $\mathcal{K}$ and Z_j are independent Gaussian random variables with zero mean and variances $\lambda_j^2(\lambda_j+\alpha)^{-2}$ (the eigenvalues of the covariance operator of ξ). Then the result follows from the standardization of these Z_j, applying Slutsky's theorem to the sum of Equation (2.23).

(b) In the same spirit as Theorem 7 of Ghiglietti et al. (2017), note that we can rewrite

$$\sqrt{n}\,\big(\widehat{M}_\alpha(\bar{x},m_0)-\widehat{M}_\alpha(m,m_0)\big) \;=\; \sqrt{n}\,\frac{\widehat{M}_\alpha^2(\bar{x},m_0)-\widehat{M}_\alpha^2(m,m_0)}{\widehat{M}_\alpha(\bar{x},m_0)+\widehat{M}_\alpha(m,m_0)},$$

where the denominator converges a.s. to $2M_\alpha(m,m_0)$, as pointed out in Corollary 2.11. Denoting for the numerator $\widehat{T}=\widehat{\mathcal{K}}^{1/2}(\widehat{\mathcal{K}}+\alpha\mathbb{I})^{-1}$, which is a linear and self-adjoint operator (as indicated in the previous part for $\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}$),

$$\begin{aligned}
\sqrt{n}\,\big(\widehat{M}_\alpha^2(\bar{x},m_0)-\widehat{M}_\alpha^2(m,m_0)\big) \;&=\; \sqrt{n}\,\big(\|\widehat{T}(\bar{x}-m_0)\|_2^2-\|\widehat{T}(m-m_0)\|_2^2\big)\\
&=\; \sqrt{n}\,\langle\widehat{T}(\bar{x}-m)\,,\,\widehat{T}(\bar{x}-m_0)+\widehat{T}(m-m_0)\rangle_2\\
&=\; \sqrt{n}\,\langle\widehat{T}(\bar{x}-m)\,,\,\widehat{T}(\bar{x}-m)+2\widehat{T}(m-m_0)\rangle_2\\
&=\; \sqrt{n}\,\|\widehat{T}(\bar{x}-m)\|_2^2+\,2\langle\sqrt{n}(\bar{x}-m)\,,\,\widehat{T}^2(m-m_0)\rangle_2.
\end{aligned}$$

The first term of this sum goes to zero in probability. Indeed, it is equal to

$$\sqrt{n}\,\|\widehat{T}(\bar{x}-m)\|_2^2=\big(\sqrt{n}\,\widehat{M}_\alpha(\bar{x},m)\big)\widehat{M}_\alpha(\bar{x},m),$$

where we can apply Slutsky's theorem, since the first element converges to the distribution of Equation (2.22) (part (a) of the theorem) and the second one converges a.s. to $M_\alpha(m,m)=0$ (by Corollary 2.11).

The remaining term can be rewritten as,

$$2\langle\sqrt{n}(\bar{x}-m)\,,\,\widehat{T}^2(m-m_0)\rangle_2 \;=\; 2\langle\sqrt{n}(\bar{x}-m)\,,\,(\widehat{T}^2-T^2)(m-m_0)\rangle_2+\ldots$$

$$+ \, 2\langle\sqrt{n}(\bar{x}-m)\,,\, T^2(m-m_0)\rangle_2, \tag{2.24}$$

where T stands for $\mathcal{K}^{1/2}(\mathcal{K}+\alpha\mathbb{I})^{-1}$.

Again using Slutsky's theorem, we can see that the first term of the sum converges to zero in probability, since we can bound it as,

$$\begin{aligned}|\langle\sqrt{n}(\bar{x}-m), (\widehat{T}^2-T^2)(m-m_0)\rangle_2| \;&\leq\; \|\sqrt{n}(\bar{x}-m)\|_2 \, \|\widehat{T}^2-T^2\|_{\mathcal{L}} \, \|m-m_0\|_2 \\ &\leq\; \|\sqrt{n}(\bar{x}-m)\|_2 \, \|\widehat{T}-T\|_{\mathcal{L}}(\|\widehat{T}\|_{\mathcal{L}}+\|T\|_{\mathcal{L}})\|m-m_0\|_2,\end{aligned}$$

where $\|\widehat{T}-T\|_{\mathcal{L}} \overset{\text{a.s.}}{\to} 0$ (as in the proof of Theorem 2.10) and $\|\sqrt{n}(\bar{x}-m)\|_2$ converges in distribution to a real random variable (as mentioned in the proof of part (a)).

Now for the last term of Equation (2.24), we can express $T^2(m-m_0)$ using its spectral decomposition as

$$T^2(m-m_0) \;=\; \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\langle m-m_0, e_j\rangle_2 \, e_j.$$

Moreover, $\sqrt{n}(\bar{x}-m)$ converges in distribution to a zero-mean Gaussian process ξ with covariance operator $\mathcal{K}$ (Theorem 8.1.1 of Hsing and Eubank (2015)), whose Karhunen-Loève representation is $\sum_{j=1}^{\infty} Z_j e_j$ with $Z_j \sim N(0,\lambda_j)$. Then, using the continuous mapping theorem for stochastic processes (for instance Theorem 18.11 of Van der Vaart (2000)), the remaining inner product of Equation (2.24) converges to

$$\begin{aligned}2\langle\sqrt{n}(\bar{x}-m)\;,\; T^2(m-m_0)\rangle_2 \;&\overset{d}{\to}\; 2\Big\langle \sum_{j=1}^{\infty} Z_j e_j \;,\; \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\langle m-m_0, e_j\rangle_2 \, e_j)\Big\rangle_2 \\ &=\; 2\sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j+\alpha)^2}\langle m-m_0, e_j\rangle_2 \; Z_j \;=\; 2\sum_{j=1}^{\infty} W_j,\end{aligned}$$

where W_j are Gaussian random variables with zero mean and variance $\lambda_j^3(\lambda+\alpha)^{-4}\langle m-m_0, e_j\rangle_2^2$. Therefore, their sum equals $\big(\sum_{j=1}^{\infty}\lambda_j^3(\lambda_j+\alpha)^{-4}\langle m-m_0,e_j\rangle_2^2\big)^{1/2}W$ with $W\sim N(0,1)$ and the result follows. □

Remark. If we are interested in testing the hypothesis $m=m_0$, we could compute the power of the test if we know the asymptotic distribution of $\sqrt{n}(\widehat{M}_\alpha(\bar{x},m_0) - M_\alpha(m,m_0))$. Although it resembles part (b) of the previous result, note that this result does not approximate the desired distribution with M_α, since the term $\sqrt{n}\widehat{M}_\alpha(m,m_0)$ goes to infinity.

2.5 Statistical applications

The purpose of this section is to give a general overview of possible applications of the proposed distance. The selected models have been mostly chosen among those previously proposed in the literature. However, as usual in empirical studies, many other meaningful scenarios could be considered. Thus we make no attempt to reach any definitive conclusion. Only the long term practitioners' experience will lead to a safer judgment.

2.5.1 Exploratory analysis

The Mahalanobis distance can be used to analyze and summarize some interesting features of the data which, for instance, can be done by generating boxplots. We follow here the experimental setting proposed in Arribas-Gil and Romo (2014), where some real and simulated data sets are used for outliers detection and functional boxplots.

Outliers detection

The simulation study proposed in Arribas-Gil and Romo (2014) checks the performance of ten different methods. The curves are generated using three different combinations of a main process (from which most trajectories are drawn) and a contamination one (from which the outliers come from). Given a contamination rate c, $n - \lceil c \cdot n \rceil$ curves are drawn from the main process and $\lceil c \cdot n \rceil$ from the contamination one (we denote as $\lceil x \rceil$ the smallest integer greater than x).

- The first model is defined by,
  $$\text{main process: } X(t) = 30t(1-t)^{3/2} + \varepsilon(t),$$
  $$\text{contamination process: } X(t) = 30t^{3/2}(1-t) + \varepsilon(t),$$
  for $t \in [0,1]$, where ε is a Gaussian process with zero mean and covariance function $K(s,t) = 0.3\ \exp(-|s-t|/0.3)$.

- The second model is given by,
  $$\text{main process: } X(t) = 4t + \varepsilon(t),$$
  $$\text{contamination process: } X(t) = 4t + (-1)^u 1.8 + (0.02\pi)^{-1/2}\ e^{\frac{-(t-\mu)^2}{0.02}} + \varepsilon(t),$$
  where ε is a Gaussian process with zero mean and covariance function $K(s,t) = \exp(-|s-t|)$, u follows a Bernoulli distribution with parameter 0.5 and μ is uniformly distributed over $[0.25, 0.75]$.

- Finally, using the same definitions for ε and μ, the third model is given by,
  $$\text{main process: } X(t) = 4t + \varepsilon(t),$$
  $$\text{contamination process: } X(t) = 4t + 2\sin(4(t+\mu)\pi) + \varepsilon(t).$$

We run 100 simulations of each model with different contamination rates $c = 0$, 0.05, 0.1, 0.15 and 0.2. The sample size for each simulation is 100 and the curves are simulated in a discretized fashion over a grid of 50 equidistant points in $[0, 1]$. We have checked nine out of the ten methods exposed in Arribas-Gil and Romo (2014), whose code is provided by the authors. The details about the implementations of each method can be found on that paper. We have adapted the code provided by the authors to include our method.

In order to formally define what we exactly mean by "an outlier" in our case, we approximate the distribution of the random variable $\|X_\alpha - m_\alpha\|_K$ given in Corollary 2.6 through a Monte Carlo sample of size 2000 where the Monte Carlo observations are generated using the covariance structure of the original data.

Then we mark as outliers the curves whose distance to the mean is greater that the 95% of the distances for the simulated data. The main drawback of this method is that the distribution of Corollary 2.6 is computed using the covariance structure of the data. Therefore, if the number of outliers is large compared with the sample size, this estimation is biased. In order to partially overcome this problem, we compute the covariance function using the robust minimum covariance determinant (MCD) estimator.

Regarding our proposal, we have noticed that the choice of α does not affect the number of selected outliers significantly. We have chosen $\alpha = 0.01$, but an automatic technique (as the one proposed in Arribas-Gil and Romo (2014) for the choice of the factor of the adjusted outliergram) could be used as well.

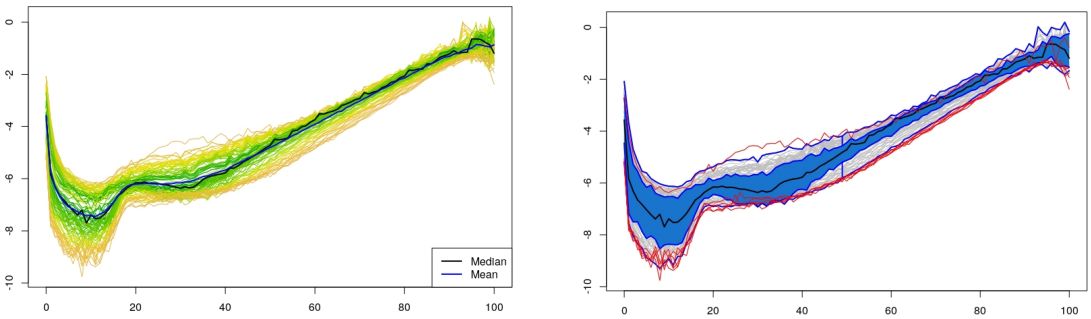
The rates of correct (p_c) and false (p_f) outliers detected for each method on the different settings can be found in Table 2.1. We can see that the Mahalanobis-based method proposed in this chapter (denoted *Mah. RKHS* in the table) is quite competitive.

Boxplots

As a part of the exploratory analysis of the data, we include the functional boxplots of two real data sets used also in Arribas-Gil and Romo (2014).

- Male mortality rates in Australia 1901-2003: this data set can be found in the R package "fds". It contains Australia male log mortality rates between 1901 and 2003, provided by The Australian Demographic Data Bank.
- Berkeley growth: this data set is available in the R package "fda". It contains height measures of 54 girls and 39 boys, under the age of 18, at 31 fixed points.

In Arribas-Gil and Romo (2014) the authors suggest to smooth the data, since the curves in the first set are very irregular. However, the distance proposed here has an intrinsic smoothing procedure, so we work directly with the original curves.



(a) Depths of the curves (b) Boxplot and outliers

Figure 2.1: Male mortality rates in Australia 1901-2003

We use the proposed Mahalanobis distance to define a depth measure, for a realization x of the process, by $(1 + M_\alpha^2(x, m))^{-1}$. Using this depth, we mark as the functional median the deepest curve of the set. The central band of the boxplot is built as the envelope of the 50% deepest curves, and the "whiskers" are constructed as the envelop of all the curves that are not marked as outliers. In order to detect the outliers we use the same procedure as before. However, the sample sizes now are too small to robustly estimate the covariance matrix over the grid, then we use the usual empirical covariance matrix.

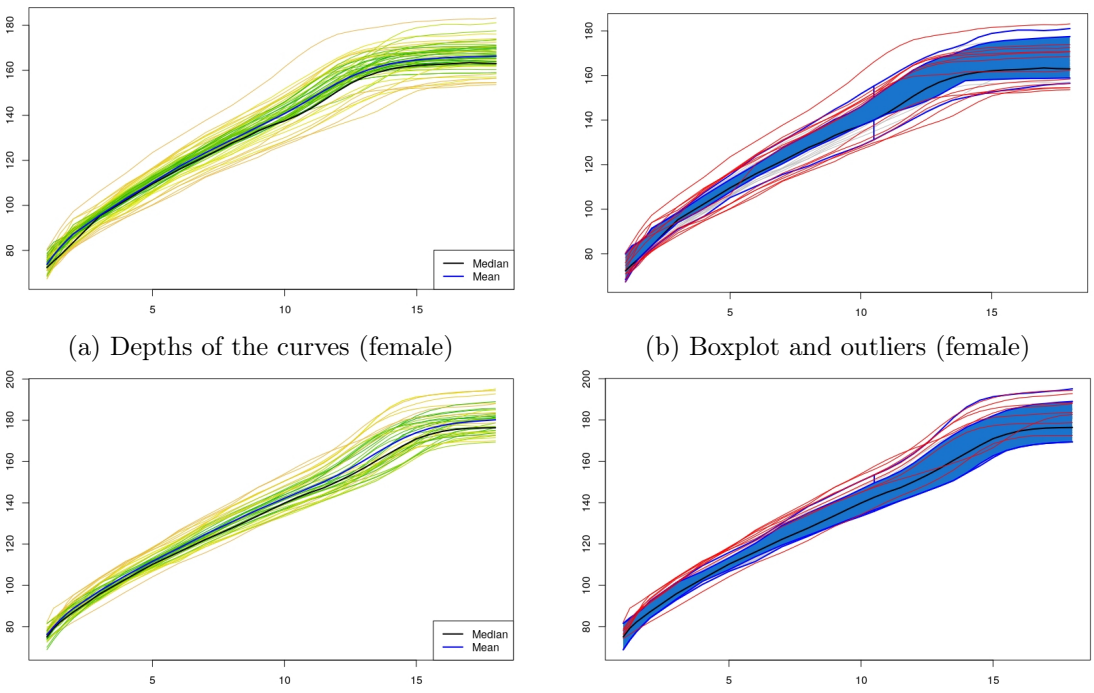
The curves marked as outliers for the male mortality set are years 1919 (influenza epidemic) and 1999-2003, which are among the curves detected using other different proposals in Arribas-Gil and Romo (2014). The resulting boxplot for this data set can be found in Figure 2.1b, where the outliers are plotted in red. Figure 2.1 includes also (on the left) a graphic representation of the depths: from green, the deepest curves, to ochre, the outer ones.

The boxplots for the Berkely growth sets, female and male, are shown in Figure 2.2. The distributions of the distances in this case are far from the theoretical distribution derived for Gaussian processes. In an attempt to overcome this problem, the parameter α is adjusted automatically in order to reduce the Kullback-Leibler divergence between both distributions. The selected values of α with this procedure are 0.089 for the female set and 0.1 for the male set. In any case, the number of outliers detected is quite large when compared to the sample size.

Female: 3, 8, 10, 13, 15, 18, 26, 29, 42, 43, 48 and 53.

Male: 5, 10, 15, 27, 29, 32, 35 and 37.

But if we look at the estimated density functions corresponding to the distribution of M_α^2 on each set (Figure 2.3), we can see that these distributions have two modes. In fact, all the curves marked as outliers are the ones that fall into the second mode (whose



(a) Depths of the curves (female)

(b) Boxplot and outliers (female)

(c) Depths of the curves (male)

(d) Boxplot and outliers (male)

Figure 2.2: Berkeley growth

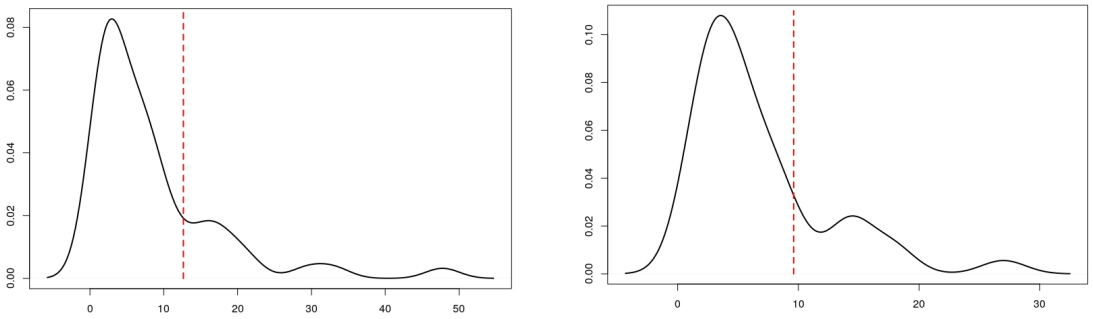
distance to the mean is greater that the red dotted line). This behavior is similar to the one of the Integrated Squared Error showed in Arribas-Gil and Romo (2014).

2.5.2 Functional binary classification

Mahalanobis distance can be used also for classification, classifying each curve through the distance to the nearest mean function, whenever the prior probabilities $\pi_1, \ldots, \pi_k$ of the classes are equal. When this is not the case, the rule to classify a coming observation x is just to assign it to the population j defined by

$$M_\alpha^2(x, m_j) - 2\log \pi_j = \min_{1 \le i \le k} \left(M_\alpha^2(x, m_i) - 2\log \pi_i \right),$$

where m_j stands for the mean functions (for instance, it is used in (Galeano et al., 2015, Section 3.3)). Here we present two different examples of binary classification with same prior probabilities. In order to check the performance of our proposal, we compare it with other classifiers presented below. The name used on the tables for each method is shown between brackets.



(a) Female (b) Male

Figure 2.3: Estimated density functions of the distributions of M_α^2 for Berkeley growth.

- Optimal Bayes classifier proposed in Dai et al. (2017) ("OB"). This is a functional extension of the classical multivariate Bayes classifier based on nonparametric estimators of the density functions corresponding to the main coefficients in Karhunen-Loève expansions. Here the curves are projected onto a common sequence of eigenfunctions and the densities of these projections are used. The authors propose three approaches to estimate these densities. We have chosen the implementation which assumes that the densities are Gaussian since, according to their results, it seems to slightly outperform the others. The number of eigenfunctions used for the projections is fixed by cross-validation.
- Mahalanobis-based semidistance of Equation (2.17) proposed in Galeano et al. (2015) ("d_{FM}^k").
- k-nearest neighbors with 3 and 5 neighbors ("knn3" and "knn5"). In spite of its simplicity, this method tends to show a good performance when dealing with functional data.

Our proposal is denoted as "M_α". Now the parameter α is fixed by cross-validation, for $\alpha \in [10^{-4}, 10^{-1}]$. For heteroscedastic problems, we have implemented our binary classifier mimicking an improvement that is usually made in the multivariate context. In that finite setting, given two equiprobable populations with covariance matrices Σ_0, Σ_1, a curve x is assigned to class 1, according to the Quadratic Discriminant classifier, whenever

$$M^2(x, m_0) - M^2(x, m_1) > \log \frac{|\Sigma_1|}{|\Sigma_0|},$$

where the finite dimensional Mahalanobis distance M is defined in (2.1) (see, for instance, Section 8.3.7 of Izenman (2008)). This rule coincides with the Bayes classifier for heteroscedastic Gaussian predictors. Then, in most cases with multivariate data, using this approach gives better results than merely classifying to class 1 when $M^2(x, m_0) > M^2(x, m_1)$. In the case of functional data this is just an heuristic improvement. If m_0, K_0 and m_1, K_1 are the mean and covariance functions of each class,

the standard classifier would assign the curve x to the class such that $M^2_{\alpha,K_i}(x, m_i)$, $i = 0, 1$, is minimum (M_{α,K_i} stands for the distance M_α when using the covariance function K_i). Instead, we will classify x to class 1 if $M^2_{\alpha,K_0}(x, m_0) - M^2_{\alpha,K_1}(x, m_1) > C$, and to 0 if not. This constant C is computed as $\log((\lambda^1_1 \cdot \ldots \cdot \lambda^1_{10})/(\lambda^0_1 \cdot \ldots \cdot \lambda^0_{10}))$, where λ^0_j, λ^1_j, $j = 1, \ldots, 10$, are the ten greater eigenvalues of $\mathcal{K}_{K_0}$ and $\mathcal{K}_{K_1}$ respectively.

Cut Brownian Motion and Brownian Bridge

The first problem under consideration is to distinguish between two "cut" versions of a standard Brownian Motion and a Brownian Bridge. By "cut" we mean to take the process $X(s)$ on the interval $s \in [0, T]$, $T < 1$. We know an explicit expression for the Bayes error of this problem, which depends on the cut point T. For the case of equal prior probabilities of the classes, which will be the case here, this Bayes error is given by,

$$L^* \;=\; \frac{1}{2} - \Phi\left(\frac{(-(1-T)\log(1-T))^{1/2}}{(T(1-T))^{1/2}}\right) + \Phi\left(\frac{(-(1-T)\log(1-T))^{1/2}}{T^{1/2}}\right),$$

where Φ stands for the distribution function of a standard Gaussian random variable. Since both processes are almost indistinguishable around zero, $L^* \to 0.5$ when $T \to 0$. Also $L^* \to 0$ when $T \to 1$, since then one can decide the class with no error just looking at the last point of the curve.

The trajectories of both processes are shown in Figure 2.4 and the cut points considered, $0.75, 0.8125, 0.875, 0.9375$ and 1, are marked with vertical dotted lines. For each class, 50 samples are drawn for training and 250 for test. The experiment is run 500 times for each cut point, and the trajectories are sampled over an equidistant grid in $[0, 1]$ of size 50. Table 2.2 shows the percentages of misclassified curves, as well as the Bayes errors. Our proposal and knn with 5 neighbors seem to outperform the other methods for this problem.

Models based on truncated Karhunen–Loève expansions

We have implemented also the experimental setting proposed in Dai et al. (2017). The authors consider three different scenarios. For the first two, the curves of classes X^0 and X^1, are drawn from processes

$$X^i(t) \;=\; \mu_i(t) + \sum_{j=1}^{50} A_{j,i}\phi_j(t) + \varepsilon, \quad i = 0, 1,$$

where ε is a Gaussian variable with zero mean and variance 0.01. The function ϕ_j is the jth element in the Fourier basis, starting with,

$$\phi_1(t) = 1, \quad \phi_2(t) = \sqrt{2}\cos(2\pi t), \quad \phi_3(t) = \sqrt{2}\sin(2\pi t).$$

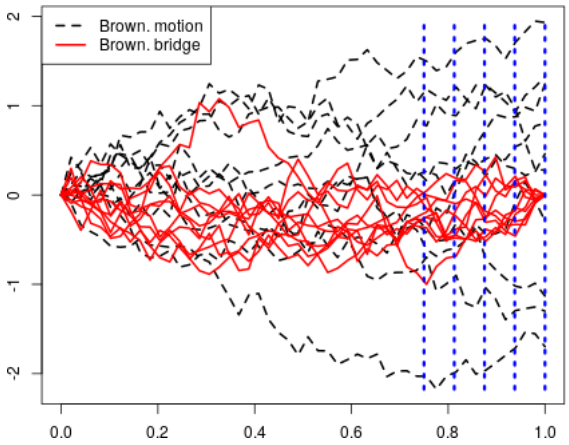


Figure 2.4: Trajectories of Brownian Motion and Bridge with cut points (vertical)

For Scenario A, the coefficients $A_{j,0}, A_{j,1}$ are independent Gaussian variables. For Scenario B they are independent centered exponential random variables. Finally, in Scenario C the processes are

$$X^i(t) \;=\; \mu_i(t) + \sum_{j=1}^{50} \frac{A_{j,i}}{B_i}\phi_j(t),\;\; i = 0, 1,$$

where $A_{j,0}, A_{j,1}$ are the same as in Scenario B and B_0, B_1 are independent variables with common distribution $\chi^2_{30}/30$. Thus, in this latter case the coefficients of the basis expansion are dependent but uncorrelated. The means and the variances of the coefficients $A_{j,i}$, $i = 0, 1$, are changed in order to check the "same" and "different" scenarios for mean and variances. Then $m_0(t) = 0$ always, and $m_1(t)$ is either 0 or t. In the same way, the variance of $A_{j,0}$ is always $\exp(-j/3)$ and the variance of $A_{j,1}$ is either $\exp(-j/3)$, or $\exp(-j/2)$. The curves are sampled on 50 equidistant points in $[0, 1]$.

The prior probabilities of both classes are set to 0.5 and two sample sizes, 50 and 100, are tested for training. For test we use 500 realizations of the processes. Each experiment is repeated 500 times. The misclassification percentages for all the different scenarios are shown in Table 2.3. Our proposal is mainly the winner, although in Scenario A it is overtaken by the Optimal Bayes classifier in the case of equal means and different variances. Also knn with 5 neighbors performs better sometimes in the case of different means and equal variances.

2.5.3 Testing for equality of means

Mahalanobis type distances can also be used to build inference tests for equality of means, as proposed in Ghiglietti et al. (2017). We try to replicate here the experimental

setting proposed by the authors. The problem now is to test the null hypothesis H_0 : $m_0 = m_1$ for $m_0, m_1 \in L^2[0,1]$ the mean functions of two processes X^0, X^1 which share the same covariance structure. Then we have two independent samples $x_1^0, \ldots, x_{n_0}^0$ and $x_1^1, \ldots, x_{n_1}^1$, with both $n_0, n_1 \to \infty$.

We can directly adapt the proof of Theorem 2.13(a) to this context. Note that $M_\alpha(\bar{x}^0, \bar{x}^1) = M_\alpha(\bar{x}^0 - \bar{x}^1, 0)$, which equals $M_\alpha(\bar{x}^0 - \bar{x}^1, m_0 - m_1)$ under H_0. Then one can use the same reasoning of that proof just reading $\bar{x}$ as $\bar{x}^0 - \bar{x}^1$, m as $m_0 - m_1$ and n as $N = n_0 n_1 (n_0 + n_1)^{-1}$. Thus, if n_0, n_1 go to infinity at the same rate (i.e. $n_0/n_1 \to 1$), then $\sqrt{N}\widehat{M}_{\alpha,n}(\bar{x}^0, \bar{x}^1)$ converges weakly to the distribution of Equation (2.22) under the null hypothesis, where λ_j, e_j are the eigenvalues and eigenfunctions of the common covariance operator. If the sizes of the samples increase at a different rate, the result would be similar but the limit distribution should be properly balanced.

With this we define the critical region,

$$R_\delta = \Big\{ \Big(\frac{n_0 n_1}{n_0 + n_1} \Big)^{\frac{1}{2}} \widehat{M}_\alpha(\bar{x}^0, \bar{x}^1) \ > \ \widehat{Q}_\delta \Big\},$$

being $\widehat{Q}_\delta$ the $1 - \delta$ quantile of the distribution in Equation (2.22) computed using the sample eigenvalues $\widehat{\lambda}_j$.

Whenever the processes fulfill $\mathbb{E}\|X^i\|^4 < \infty$, $i = 0, 1$, we can prove that the previous critical region is asymptotically of level δ, since we can adapt the proof of Theorem 4.3 of Ghiglietti and Paganoni (2014) to get the convergence of the quantiles $\widehat{Q}_\delta$. We should simply replace in that proof $h_k(p)$ with $\lambda_k^2(\lambda_k + \alpha)^{-2}$ and take k_ϵ such that $2\sum_{k=k_\epsilon+1}^{\infty} \lambda_k < \alpha\epsilon$ and $\lambda_{k_\epsilon} \leq 1$, so that (for terms B_N and C_N of the proof) we can use

$$\sum_{k=k_\epsilon+1}^{\infty} \lambda_k^2 (\lambda_k + \alpha)^{-2} \ \leq \ \alpha^{-2} \sum_{k=k_\epsilon+1}^{\infty} \lambda_k^2 \ \leq \ \alpha^{-2} \sum_{k=k_\epsilon+1}^{\infty} \lambda_k.$$

In order to follow the experimental setting of Ghiglietti et al. (2017) as much as possible, we use the processes,

$$X^0(s) = m_0(s) + \sum_{j=1}^{\infty} U_j^0 \sqrt{\theta_j} e_j(s), \quad X^1(s) = m_1(s) + \sum_{j=1}^{\infty} U_j^1 \sqrt{\theta_j} e_j(s),$$

for $s \in [0,1]$, where the coefficients U_j^0, U_j^1 are uniform random variables in $(-\sqrt{3}, \sqrt{3})$. The sequence of eigenvalues λ_j of the common covariance K are given by,

$$\lambda_j = \begin{cases} \frac{1}{j+1} & \text{if } j \in 1, 2, 3, \\ \frac{1}{(j+1)^4} & \text{if } j \geq 4, \end{cases}$$

and the eigenfunctions e_j,

$$e_j(t) = \begin{cases} 1 & \text{if } j = 1, \\ \sqrt{2}\sin(j\pi t) & \text{if } j \geq 2 \text{ even}, \\ \sqrt{2}\cos((j-1)\pi t) & \text{if } j \geq 2 \text{ odd}. \end{cases}$$

The mean functions are different for the four proposed scenarios, in order to change the dependence of $(m_0 - m_1)$ on the eigenfunctions e_j. However, the exact expressions for these mean functions are not explicitly stated in the original paper. The authors simply state the dependence of each mean function on the eigenfunctions e_j, $j = 1, \ldots$ (for instance, $m_1 - m_0$ belongs to $\text{span}\{e_1\}$ in the first scenario). Thus, we have chosen them in order to recover approximately the results exposed in Ghiglietti et al. (2017). The mean of the first class is always $m_0(s) = 4s(1-s)$. The different scenarios and the exact means chosen are,

- $\boldsymbol{j = 1}$: $(m_1 - m_0) \in \text{span}\{e_1\}$, $m_1(s) = m_0(s) + 0.25\lambda_1 e_1(s)$,
- $\boldsymbol{j \leq 3}$: $(m_1 - m_0) \in \text{span}\{e_1, e_2, e_3\}$, $m_1(s) = m_0(s) + 0.2\sum_{j=1}^{3} \lambda_j e_j(s)$,
- $\boldsymbol{j = 4}$: $(m_1 - m_0) \in \text{span}\{e_4\}$, $m_1(s) = m_0(s) + 2\sqrt{\lambda_4} e_4(s)$,
- $\boldsymbol{j \geq 5}$: $(m_1 - m_0) \in \text{span}\{e_j, j \geq 5\}$, $m_1(s) = m_0(s) + 2.1\sum_{j=5}^{\infty} \sqrt{\lambda_j} e_j(s)$.

It is stressed in bold the name used for each scenario on the tables. The curves are sampled on a grid of 50 equidistant points in $[0, 1]$. Since it is not possible to simulate infinitely many eigenfunctions for the last scenario, we cut the series for $j = 100$. For each class 300 samples are generated and each experiment is repeated 10^3 times. The test is carried out at significance level 0.05.

In Ghiglietti et al. (2017), the authors check their proposal using two different functions $g(c; p)$ for the integral of Equation (2.18). We have implemented here the version with $g(c; p) = \exp(-c/p)$, since it seems to slightly overtake the other one, according to their results. We use the three intermediate values of p out of the five tested in the original paper, $p = 0.1, 1$ and 10.

We compute the power of the test for different values of the smoothing parameter $\alpha = 5 \cdot 10^{-5}, 10^{-4}$ and 10^{-3}. An unique value for α could be fixed using an automatic procedure as the one proposed in Arribas-Gil and Romo (2014) for the adjusted outliergram: resample from the theoretical distribution to maintain the false positive rate close to the significance level of the test.

The results can be found in Table 2.4. The different methods are marked with the value of its parameter: α for our proposal and p for the one of Ghiglietti et al. (2017). As expected, when increasing the value of α, the power of the test decreases for the last scenarios, where the mean functions do not depend on the first eigenfunctions. On the current setting, the fifth eigenvalue is approximately $7.72 \cdot 10^{-4}$. Thus, if the value chosen for α is orders of magnitude greater than 10^{-4}, the test is not able to manage well the last scenarios.

c= 0	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	-	0.001 (0.003)	-	0.001 (0.002)	-	0.000 (0.002)
Adj. Fun. BP	-	0.007 (0.010)	-	0.006 (0.010)	-	0.007 (0.012)
Fun. HDR BP	-	0.010 (0.000)	-	0.010 (0.000)	-	0.010 (0.000)
Rob. Mah. Dist.	-	0.016 (0.014)	-	0.015 (0.013)	-	0.015 (0.015)
ISE	-	0.038 (0.020)	-	0.032 (0.021)	-	0.033 (0.021)
DB trimming	-	0.013 (0.007)	-	0.012 (0.006)	-	0.014 (0.007)
DB weighting	-	0.016 (0.012)	-	0.015 (0.011)	-	0.014 (0.011)
Outliergram	-	0.054 (0.025)	-	0.057 (0.027)	-	0.058 (0.022)
Adj. Ourliergram	-	0.012 (0.012)	-	0.011 (0.013)	-	0.011 (0.014)
Mah. RKHS	-	0.037 (0.015)	-	0.033 (0.018)	-	0.035 (0.016)
c= 0.05	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.186 (0.193)	0.001 (0.003)	0.208 (0.220)	0.000 (0.001)	0.184 (0.179)	0.000 (0.002)
Adj. Fun. BP	0.576 (0.282)	0.008 (0.012)	0.551 (0.330)	0.006 (0.010)	0.588 (0.344)	0.008 (0.012)
Fun. HDR BP	0.155 (0.084)	0.002 (0.004)	0.131 (0.096)	0.004 (0.005)	0.057 (0.091)	0.008 (0.005)
Rob. Mah. Dist.	0.976 (0.096)	0.008 (0.009)	0.361 (0.250)	0.008 (0.010)	0.104 (0.153)	0.015 (0.013)
ISE	0.865 (0.313)	0.033 (0.020)	1.000 (0.000)	0.038 (0.026)	1.000 (0.000)	0.033 (0.021)
DB trimming	0.947 (0.183)	0.008 (0.009)	0.957 (0.135)	0.008 (0.009)	0.994 (0.035)	0.006 (0.007)
DB weighting	0.894 (0.259)	0.008 (0.009)	0.941 (0.203)	0.012 (0.011)	0.957 (0.168)	0.011 (0.009)
Outliergram	0.998 (0.020)	0.038 (0.022)	0.998 (0.020)	0.033 (0.021)	1.000 (0.000)	0.036 (0.023)
Adj. Ourliergram	0.994 (0.035)	0.006 (0.008)	0.978 (0.070)	0.006 (0.009)	0.998 (0.020)	0.012 (0.014)
Mah. RKHS	0.998 (0.020)	0.022 (0.016)	1.000 (0.000)	0.027 (0.014)	1.000 (0.000)	0.031 (0.016)
c= 0.1	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.139 (0.123)	0.000 (0.001)	0.158 (0.151)	0.000 (0.002)	0.134 (0.128)	0.000 (0.002)
Adj. Fun. BP	0.549 (0.239)	0.005 (0.008)	0.593 (0.268)	0.008 (0.010)	0.632 (0.248)	0.008 (0.012)
Fun. HDR BP	0.073 (0.044)	0.003 (0.005)	0.083 (0.038)	0.002 (0.004)	0.047 (0.050)	0.006 (0.006)
Rob. Mah. Dist.	0.961 (0.105)	0.004 (0.007)	0.373 (0.170)	0.007 (0.009)	0.104 (0.108)	0.011 (0.014)
ISE	0.790 (0.335)	0.027 (0.017)	1.000 (0.000)	0.036 (0.021)	1.000 (0.000)	0.033 (0.022)
DB trimming	0.808 (0.340)	0.009 (0.009)	0.989 (0.045)	0.010 (0.010)	0.995 (0.030)	0.008 (0.011)
DB weighting	0.176 (0.247)	0.001 (0.004)	0.910 (0.232)	0.005 (0.008)	0.922 (0.258)	0.006 (0.008)
Outliergram	0.981 (0.040)	0.020 (0.014)	0.998 (0.014)	0.018 (0.012)	1.000 (0.000)	0.020 (0.016)
Adj. Ourliergram	0.897 (0.118)	0.006 (0.009)	0.971 (0.076)	0.006 (0.009)	1.000 (0.000)	0.007 (0.011)
Mah. RKHS	0.767 (0.148)	0.014 (0.012)	1.000 (0.000)	0.014 (0.011)	0.995 (0.030)	0.015 (0.013)
c= 0.15	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.098 (0.105)	0.000 (0.002)	0.114 (0.101)	0.000 (0.002)	0.134 (0.130)	0.000 (0.001)
Adj. Fun. BP	0.494 (0.215)	0.006 (0.010)	0.550 (0.242)	0.006 (0.009)	0.584 (0.247)	0.006 (0.009)
Fun. HDR BP	0.043 (0.032)	0.004 (0.006)	0.063 (0.016)	0.001 (0.003)	0.050 (0.029)	0.003 (0.005)
Rob. Mah. Dist.	0.927 (0.098)	0.001 (0.003)	0.324 (0.184)	0.004 (0.007)	0.152 (0.175)	0.005 (0.008)
ISE	0.778 (0.349)	0.027 (0.018)	0.999 (0.007)	0.040 (0.029)	1.000 (0.000)	0.034 (0.023)
DB trimming	0.444 (0.410)	0.009 (0.011)	0.981 (0.099)	0.016 (0.015)	0.993 (0.067)	0.009 (0.011)
DB weighting	0.020 (0.039)	0.001 (0.003)	0.659 (0.329)	0.002 (0.005)	0.634 (0.391)	0.002 (0.005)
Outliergram	0.879 (0.137)	0.011 (0.012)	0.984 (0.043)	0.008 (0.011)	0.999 (0.013)	0.008 (0.009)
Adj. Ourliergram	0.616 (0.220)	0.003 (0.007)	0.969 (0.099)	0.006 (0.008)	0.996 (0.019)	0.007 (0.010)
Mah. RKHS	0.295 (0.122)	0.013 (0.011)	0.988 (0.052)	0.008 (0.009)	0.941 (0.167)	0.007 (0.009)
c= 0.2	Model 1		Model 2		Model 3	
Method	p_c	p_f	p_c	p_f	p_c	p_f
Fun. BP	0.060 (0.090)	0.000 (0.001)	0.098 (0.104)	0.000 (0.001)	0.102 (0.094)	0.000 (0.000)
Adj. Fun. BP	0.376 (0.226)	0.003 (0.006)	0.509 (0.205)	0.005 (0.009)	0.540 (0.227)	0.003 (0.006)
Fun. HDR BP	0.034 (0.024)	0.004 (0.006)	0.047 (0.012)	0.001 (0.003)	0.036 (0.022)	0.003 (0.006)
Rob. Mah. Dist.	0.866 (0.167)	0.000 (0.002)	0.304 (0.171)	0.002 (0.005)	0.111 (0.118)	0.004 (0.007)
ISE	0.513 (0.396)	0.031 (0.023)	0.997 (0.018)	0.047 (0.031)	0.999 (0.010)	0.028 (0.023)
DB trimming	0.235 (0.314)	0.009 (0.013)	0.990 (0.037)	0.015 (0.014)	0.979 (0.121)	0.011 (0.011)
DB weighting	0.015 (0.025)	0.001 (0.004)	0.216 (0.228)	0.001 (0.003)	0.111 (0.179)	0.000 (0.002)
Outliergram	0.356 (0.202)	0.002 (0.005)	0.894 (0.158)	0.001 (0.004)	0.959 (0.146)	0.001 (0.004)
Adj. Ourliergram	0.248 (0.179)	0.001 (0.003)	0.959 (0.074)	0.004 (0.008)	0.999 (0.007)	0.008 (0.011)
Mah. RKHS	0.141 (0.089)	0.012 (0.011)	0.945 (0.127)	0.005 (0.007)	0.749 (0.232)	0.006 (0.009)

Table 2.1: Ratio of correct and false detected outliers.

t	Bayes	M_α	OB	d^k_{FM}	knn3	knn5
0.75	33.9	42.5 (3.5)	43.5 (2.5)	46.4 (3.2)	43.2 (2.8)	**42.4** (2.8)
0.8125	30.8	**40.0** (3.7)	41.9 (2.6)	44.8 (3.3)	41.0 (2.8)	40.1 (3.0)
0.875	26.9	**36.1** (3.6)	40.2 (2.6)	42.6 (3.7)	38.0 (3.0)	36.9 (3.0)
0.9375	20.9	**32.3** (3.1)	38.0 (2.8)	39.9 (3.5)	33.7 (2.7)	32.5 (2.7)
1	0.0	**26.5** (2.8)	35.9 (2.9)	36.0 (3.5)	28.4 (2.7)	27.6 (2.7)

Table 2.2: Percentage of misclassification for cut Brownian Motion and Brownian Bridge.

Scenario A (Gaussian)							
n	mean	sd	M_α	OB	d^k_{FM}	knn3	knn5
50	same	diff	35.9 (3.5)	**19.0** (4.0)	47.0 (3.1)	45.6 (2.2)	46.2 (2.0)
	diff	same	42.3 (3.8)	47.3 (6.8)	43.7 (3.7)	42.9 (3.6)	**42.0** (3.6)
	diff	diff	**29.1** (5.0)	36.4 (10.1)	40.0 (5.4)	39.7 (3.0)	40.0 (3.1)
100	same	diff	34.2 (3.0)	**9.3** (2.1)	45.8 (3.5)	44.6 (1.9)	45.4 (1.8)
	diff	same	**34.6** (4.5)	45.1 (8.2)	37.0 (4.4)	42.1 (3.0)	41.0 (3.0)
	diff	diff	**22.0** (4.9)	35.7 (11.3)	34.2 (6.2)	38.3 (2.4)	38.6 (2.5)
Scenario B (exponential)							
n	mean	sd	M_α	OB	d^k_{FM}	knn3	knn5
50	same	diff	**24.2** (5.2)	30.2 (10.4)	37.0 (6.6)	37.6 (2.6)	38.0 (2.7)
	diff	same	41.8 (3.9)	49.1 (5.5)	42.3 (4.1)	38.0 (3.4)	**37.2** (3.6)
	diff	diff	**14.3** (4.8)	31.8 (12.8)	25.1 (9.0)	24.7 (3.1)	25.1 (3.5)
100	same	diff	**16.9** (3.1)	24.0 (9.6)	28.2 (6.1)	35.3 (2.4)	35.7 (2.3)
	diff	same	**34.5** (4.6)	48.3 (5.9)	36.7 (4.2)	36.5 (2.8)	35.6 (2.7)
	diff	diff	**7.7** (2.9)	30.1 (13.4)	17.8 (6.3)	21.6 (2.4)	21.8 (2.6)
Scenario C (dependent)							
n	mean	sd	M_α	OB	d^k_{FM}	knn3	knn5
50	same	diff	**30.0** (5.4)	33.3 (8.1)	40.1 (5.9)	39.9 (2.7)	39.9 (2.7)
	diff	same	43.6 (4.1)	48.8 (4.8)	42.9 (4.2)	38.1 (3.6)	**37.5** (3.8)
	diff	diff	**19.9** (4.9)	36.2 (11.0)	30.3 (7.7)	26.4 (3.1)	26.6 (3.3)
100	same	diff	**21.7** (3.0)	28.0 (7.5)	29.4 (5.7)	37.6 (2.4)	37.5 (2.4)
	diff	same	38.0 (4.3)	48.8 (5.0)	38.9 (3.8)	36.5 (2.7)	**35.6** (2.8)
	diff	diff	**13.3** (3.2)	34.6 (11.0)	23.2 (6.1)	23.4 (2.4)	23.3 (2.4)

Table 2.3: Percentage of misclassification for the experimental setting of Dai et al. (2017).

mean	$\alpha = 5 \cdot 10^{-5}$	$\alpha = 10^{-4}$	$\alpha = 10^{-3}$	$p = 10^{-1}$	$p = 10^{0}$	$p = 10^{1}$
$j = 1$	0.757	0.749	0.747	0.854	0.852	0.82
$j \leq 3$	0.953	0.95	0.948	0.919	0.928	0.953
$j = 4$	1	1	0.166	0.69	0.947	1
$j \geq 5$	1	1	0.065	0.974	1	1

Table 2.4: Power of the test $H_0 : m_0 = m_1$ at significance level 0.05.

Chapter 3

Variable selection in functional regression

The study of regression models is clearly among the leading topics in statistics. In particular, these models play a central role in the theory of statistics with functional data. As in the previous chapters, we consider "functional data" consisting of independent $x_1 = x_1(s), \ldots, x_n = x_n(s)$ observations (trajectories) drawn from a second-order (L^2) stochastic process $X = X(s)$, $s \in [0,1]$, with continuous trajectories and continuous mean and covariance functions, denoted by $m = m(s)$ and $K(s,t)$, respectively. All the involved random variables are supposed to be defined on a common probability space $(\Omega, \mathcal{F}, P)$.

More specifically, we are concerned with variable selection issues; see, (Berrendero et al., 2016, Sec. 1), Fan and Lv (2010) for additional information and references. Basically, a variable selection functional method is an automatic procedure that takes a function $\{x(s),\ s \in [0,1]\}$ to a finite-dimensional vector $(x(t_1), \ldots, x(t_p))'$. The overall idea of variable selection is to choose the variables $x(t_j)$ (or, equivalently, the "impact points" $t_1, \ldots, t_p \in [0,1]$; see Kneip et al. (2016)), in an "optimal way" so that the original functional problem (regression, classification, clustering,...) is replaced with the corresponding multivariate version, based on the selected variables.

Some notation

A vector containing the possible "impact" points $t_1, \ldots, t_p \in [0,1]$ will be denoted T (sometimes S) or T_p when we want to stress the dimension of T. Sometimes it may denote the set consisting of the same points. Also, $X(T_p)$ will stand for $(X(t_1), \ldots, X(t_p))'$. The superindex "*" will be used to denote that the points t_j^* are the "true" ones, or the "optimal" ones according to some criterion.

Given a random variable Z (with finite variance) the notation Z_{T_p} will refer to the orthogonal projection of Z on the space spanned by the components of $X(T_p) - m(T_p)$.

If $p^* < p$, the notation $T_{p^*} \prec T_p$ will indicate that all the points in T_{p^*} belong also to T_p.

Finally, as usual in statistics, we use an upper hat sign to denote the estimated quantities (or the predicted variables). For instance, $\widehat{T}_p$ will denote a data-driven estimator of T_p and $\widehat{Y}_{\widehat{T}_p}$ will stand for the corresponding (fully data-driven) prediction of the response Y_{T_p}. The halfway notation $Y_{\widehat{T}_p}$ will represent the orthogonal projection of the response variable onto the space spanned by the marginal variables indexed by the estimated points $\widehat{T}_p$.

Organization of the chapter

In the first part of the chapter we address the problem of functional regression with scalar response, which is subsequently extended to functional response. Finally we adapt the proposed methodology to the prediction of functional time series.

Thus, the problem with scalar response is presented in Section 3.1. In Section 3.2 we introduce and motivate (in population terms) our variable selection procedure. The asymptotic properties of the empirical version (when the parameters are estimated) are considered in Section 3.3. The problems associated with the choice of the number p of selected variables are analyzed in Sections 3.4 and 3.5. The empirical results (simulations and real data examples) for scalar response are presented in Section 3.6.

The last two sections correspond to the extension of this methodology to functional linear regression with functional response. Section 3.7 includes the extensions of the results derived in the first part of the chapter. In Section 3.8 this methodology is applied to the prediction of functional time series.

3.1 Introduction to scalar response: statement of the problem and motivation

The problem under study: variable selection in functional regression

In this first part of the chapter we are interested in functional regression models with scalar response, of type $y_i = g(x_i) + \varepsilon_i$, where g is a real function defined on a suitable space $\mathcal{X}$ where the trajectories of our process are supposed to live. The random variables ε_i are independent errors (and also independent from the x_i) with mean zero and common variance σ^2. As we just mentioned, we are interested in variable selection. In the regression setting, this would amount to replace the functional model $y_i = g(x_i)+\varepsilon_i$ with a finite dimensional version of type $y_i = \phi\big(x_i(t_1), \ldots, x_i(t_p)\big) + e_i$. Nevertheless, note that still the problem is of a functional nature, since the methods to select the t_j are generally based upon the full data trajectories.

Some motivation. Drawbacks of the classical linear L^2-model for variable selection purposes

It is quite natural to assume that the explanatory functional variables $x_i = x_i(t)$ are members of the space $L^2[0,1]$, endowed with the usual inner product $\langle x_1, x_2\rangle_2 = \int_0^1 x_1(s)x_2(s)\mathrm{d}s$, for $x_1, x_2 \in L^2[0,1]$. In this setting, the most popular choice for g is, by far, a linear (or affine) operator from $L^2[0,1]$ to $\mathbb{R}$ which leads to a model of type

$$y_i = \alpha_0 + \langle x_i, \beta\rangle_2 + \varepsilon_i,\ i = 1, \ldots, n, \tag{3.1}$$

where $x = x(s)$ is the explanatory functional variable, $\alpha_0 \in \mathbb{R}$ is the intercept constant and $\beta \in L^2[0,1]$ denotes the slope function. As in the standard multivariate regression model, the aim here is to estimate α_0 and β in order to be able to make accurate predictions of the response variable Y.

The corresponding theory is outlined in several places; see, e.g., the article Cardot and Sarda (2010) or the books Ferraty and Vieu (2006); Horváth and Kokoszka (2012). The Hilbert structure of the $L^2[0,1]$ space allows us to keep ourselves as close as possible to the usual least squares framework in multivariate regression; for example, the projection $P(x)$ of an element x on a closed subspace H is characterized by the orthogonality condition $\langle x - P(x), a\rangle_2 = 0$, for all $a \in H$. However, other crucial differences with the finite-dimensional case (mostly associated with the non-invertibility of the covariance operator of the process) make the functional L^2 regression theory far from trivial. Most of these difficulties are intrinsic to the infinite-dimensional nature of the data, so that they cannot be overcome by just replacing $L^2[0,1]$ with another function space. However, when it comes to variable selection applied to linear regression, it would be useful to have the finite dimensional linear model (based on the selected variables)

$$y_i = \alpha_0 + \sum_{j=1}^{p} \beta_j x_i(t_j) + \varepsilon_i,\ i = 1, \ldots, n \tag{3.2}$$

as a particular case of our general model. Notice that (3.2) cannot be established in the L^2 framework, since a transformation of type $x \in L^2[0,1] \mapsto \sum_{i=1}^{p} \beta_i x(t_j)$ is not a linear continuous functional in L^2. In heuristic terms, one would need to look for the slope function β in a suitable space, for which a "finite-dimensional" model such as (3.2) could make sense. More precisely, we will change the "habitat" space for the function β: instead of assuming $\beta \in L^2[0,1]$ we will assume that β belongs to the Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}(K)$, associated with K.

As we will see below, the assumed membership of β to $\mathcal{H}(K)$ entails some additional restrictions of regularity on the slope function β (when compared to the simple assumption $\beta \in L^2[0,1]$). In any case, such a situation is not unusual: some restrictions on β appear in different ways even when the classical L^2-model (3.1) is considered. The reason is that the space $L^2[0,1]$ is in fact too large from several points of view. Hence,

in spite of the advantages of the L^2-model commented above, one typically uses penalization or projection methods to exclude extremely rough solutions in the estimation of β.

Our proposal here, as presented in the next section, aims at reconciling two targets: first, we look for a functional linear model, wide enough to include finite-dimensional versions, such as (3.2), as particular cases. Second, we would like to achieve such a goal with a minimal change in the space where β lives.

Some related literature

A quite general RKHS-based approach to the problem of dimension reduction in functional regression has been proposed by Hsing and Ren (2009). These authors follow the inverse regression methodology to deal with a model of type $Y = \ell(\xi_1, ...\xi_d) + \varepsilon$ where ℓ is a link function and the ξ_j are linear functionals of the explanatory variable X, defined in RKHS terms. This pioneering reference shows very clearly the huge potential of the RKHS approach. However, as the authors point out, there are still many aspects not considered in that paper and worth of attention. Variable selection is one of them. In fact, the whole point of the present chapter is to show that things become particularly simple when the RKHS machinery is applied to variable selection. A recent use of RKHS methods in the problem of functional binary classification is developed in Berrendero et al. (2017). See also Berlinet and Thomas-Agnan (2004); Hsing and Eubank (2015) for a broader perspective of the applicability of RKHS methods in statistics.

Other variable selection methods, always aimed at selecting the "best points" $t_1, \ldots, t_d$ (or the "best variables" $X(t_1), \ldots, X(t_p)$) have been proposed as well, with no explicit reference of RKHS tools. Thus, the selection of the "best impact point" t_1 in a model of type (3.2) with $p = 1$ is addressed in McKeague and Sen (2010). Different variable selection methods have been suggested by Aneiros and Vieu (2014); Ferraty et al. (2010); Delaigle et al. (2012) for prediction and classification purposes. In addition, a non-parametric approach for non-linear functional regression models is presented in Aneiros and Vieu (2016). See the references therein for more information on non-linear models. Also, a criterion for "optimal design" in trajectory recovery is considered in Ji and Müller (2017).

A recent general proposal for dimension reduction (beyond variable selection and regression models) is Fraiman et al. (2016).

3.2 An RKHS-based linear model suitable for variable selection

Our choice of the ambient space for the slope function β is, in some sense, "customized" for the problem at hand, since we will consider the Reproducing Kernel Hilbert Space

(RKHS) associated with the process $\{X(s),\ s \in [0,1]\}$. The theory of RKHS's goes back to the 1950s and has found a surprisingly large number of applications in different fields, including statistics, see Berlinet and Thomas-Agnan (2004). These spaces have been introduced in detail in Section 1.2.1.

3.2.1 The RKHS functional regression model

We propose to replace the standard L^2 functional regression model (3.1) with the following RKHS counterpart

$$y_i = \alpha_0 + \langle x_i, \beta \rangle_K + \varepsilon_i,\ i = 1, \ldots, n, \tag{3.3}$$

where $\beta \in \mathcal{H}(K)$ and $\langle \cdot, \cdot \rangle_K$ denotes the inner product in $\mathcal{H}(K)$.

Since the estimation of the intercept term α_0 is straightforward from those of β and m, we will assume, without loss of generality, that $\alpha_0 = 0$ in what follows.

As mentioned at the end of the previous subsection, it is important to keep in mind that the trajectories of the process X do not belong to $\mathcal{H}(K)$. Thus, the expression $\langle x_i, \beta \rangle_K$ has no direct meaning, unless it is appropriately interpreted: in what follows, $\langle x_i, \beta \rangle_K$ for $x_i = X(\omega)$ must be understood as $\Psi_{x_i}^{-1}(\beta) := (\Psi_X^{-1}(\beta))(\omega)$, where Ψ_X is the Loève's isometry defined in (1.3) (Chapter 1). Then, with this definition, we might replace (for a given $\beta \in \mathcal{H}(K)$) the random process X with a specific trajectory x and in that case $\langle x, \beta \rangle_K$ would be well defined (as a constant) even if $x \notin \mathcal{H}(K)$.

Such an interpretation of $\langle X, \beta \rangle_K$ arises in the classical paper by Parzen (Parzen, 1961b, Th. 7A), aiming at different statistical purposes. In addition, note that in the context of the linear model (3.3), we assume $E[X(s)\varepsilon] = 0$ and $E[\varepsilon] = 0$, so that $\beta(s) = \operatorname{cov}\big(Y, X(s)\big)$; hence, $\langle X, \beta \rangle_K$ might be also defined as the solution $Z \in \mathcal{L}(X)$ of the functional equation $\Psi_X(Z)(s) = \operatorname{cov}\big(Y, X(s)\big)$.

The above commented problems to give a proper definition of $\langle X, \beta \rangle_K$ are reminiscent of those arising when defining Itô's stochastic integral. In fact, when $X(s)$ is a standard Brownian Motion in $[0,1]$, model (3.3) with $\alpha_0 = 0$ can be expressed as

$$Y = \int_0^1 \beta'(s) dX(s) + \varepsilon,\ \text{with } \beta \in \mathcal{H}(K),\ \text{and } K(s,t) = \min(s,t),$$

where $\int_0^1 \beta'(s) dX(s)$ is Itô's integral and $\mathcal{H}(K)$ is the space of all real absolutely continuous functions β on $[0,1]$ with $\beta' \in L^2[0,1]$ and $\beta(0) = 0$ (Janson, 1997, Example 8.19, p. 122).

3.2.2 Variable selection in the RKHS functional regression model

Consider the RKHS functional regression model (3.3) introduced in the previous paragraph, where $\mathbb{E}[\varepsilon] = \mathbb{E}\big[\big(X(t) - m(t)\big)\varepsilon\big] = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$.

Our goal

Under this model, for fixed p, we aim at selecting p values $t_1, \ldots, t_p$ in order to use the p dimensional vector $(X(t_1), \ldots, X(t_p))'$ instead of the whole process $\{X(s) : s \in [0,1]\}$ in our regression problem. Formally, we want to establish a transformation

$$\{X(s) : s \in [0,1]\} \mapsto (X(t_1), \ldots, X(t_p))',$$

which should be "optimal" in the sense that the points $t_1, \ldots, t_p$ are chosen according to an optimality criterion, oriented to minimize the information loss in the passage from infinite to finite dimension.

In this section, we address this problem at the population level, that is, we assume that the parameters defining the model (the slope function β, the covariance function K of the process X, the mean function m and the variance of the error variable, σ^2) are known. Of course, the practical implementation will require using suitable estimators of the unknown parameters. This raises several questions concerning the sample behavior of the method which will be addressed in subsequent sections.

The optimality criterion Q_1

The first obvious question to address in such strategy is the choice of the optimality criterion. We will see that, in fact, different criteria can be used but, fortunately, they are all equivalent.

One of the basic goals of a functional regression model is to predict the value of the response variable Y for a given trajectory of the input process X. Then, a sensible approach for variable selection is to choose the p points $X(t_1), \ldots, X(t_p)$ that give the best linear prediction (in the sense of the L^2 norm) of Y. This implies to find the vector T_p that minimizes the function

$$Q_1(T_p) := \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p} \Big\| Y - \sum_{j=1}^{p} \beta_j \big(X(t_j) - m(t_j)\big) \Big\|^2. \tag{3.4}$$

This natural criterion has been considered elsewhere, sometimes in slightly different contexts, see e.g., Ji and Müller (2017). The contribution here is to interpret (3.4) in RKHS terms and, as a consequence, to show that the problem of finding the optimal value of p can be addressed in a meaningful way.

Where to look for the optimum

An important technical aspect is the choice of an appropriate subset $\Theta_p \subset [0,1]^p$ to look for the optimum of the continuous function Q_1. This subset must be compact in order to guarantee the existence of the optimum. Moreover, if we want to get a meaningful optimal value of $T_p = (t_1, \ldots, t_p)$ we should rule out those points including repeated values in the coordinates t_i. To this end, we will fix an arbitrarily small value $\delta > 0$, and will look for our optimum in the space

$$\Theta_p = \Theta_p(\delta) = \{T_p = (t_1, \ldots, t_p) \in [0,1]^p : t_{i+1} - t_i \geq \delta, \text{ for } i = 0, \ldots, p\}, \tag{3.5}$$

where $t_0 = 0$, $t_{p+1} = 1$. In practice, the restriction to the subset Θ_p is not relevant, since we observe the functions in a finite grid, and we can set $\delta > 0$ as small as required so that all the points in the grid belong to Θ_p.

The reason for the choice (3.5) of Θ_p is technical, very much in the same spirit of (Ji and Müller, 2017, Eq. (9)). We need to work on a compact set and, at the same time, to avoid degeneracy problems in the choice of the points $(t_1, \ldots, t_p)$ that could lead to a singular covariance matrix in $(X(t_1), \ldots, X(t_p))$. Other choices are possible for Θ_p. For example, one could think of defining $t_0 = 0$, $t_{p+1} = 1$ and

$$\Theta_p \;=\; \{T_p = (t_1, \ldots, t_p) \in [0,1]^p : t_i \leq t_{i+1}, \text{ for } i = 0, \ldots, p\}. \tag{3.6}$$

This could lead to "degenerate" options with $t_i = t_{i+1}$ for some values of i. However, the theory we develop below using (3.5) could be carried out alternatively with (3.6) as long as we adopt the criterion of reducing the dimension of those vectors $(t_1, \ldots, t_p)$ with ties in the coordinate values by keeping just one coordinate for each different value. In this way, for example, $(0.2, 0.2, 0.5, 0.7)$ would be interpreted just as $(0.2, 0.5, 0.7)$.

Two additional, equivalent optimality criteria

A second optimality criterion, equivalent to that based on Q_1, arises if we take into account that, from the reproducing property, when the slope function is a finite linear combination of the form $\sum_{j=1}^{p} \beta_j K(t_j, \cdot)$, model (3.3) reduces to the usual finite dimensional multiple regression model:

$$Y = \sum_{j=1}^{p} \beta_j \big(X(t_j) - m(t_j)\big) + \varepsilon. \tag{3.7}$$

Then, another sensible approach for variable selection is to choose those points $t_1, \ldots, t_p$ giving the best approximation of the true slope function β in terms of a finite linear combination of the form $\sum_{j=1}^{p} \beta_j K(t_j, \cdot)$. It is quite natural to use the norm in $\mathcal{H}(K)$ to assess this approximation since both β and these finite linear combinations live in this

RKHS. This approach amounts to find the vector $T_p \in \Theta_p$ that minimizes the function

$$Q_2(T_p) := \min_{(\alpha_1,\ldots,\alpha_p)\in\mathbb{R}^p} \Big\|\beta - \sum_{j=1}^{p} \alpha_j K(t_j,\cdot)\Big\|_K^2. \tag{3.8}$$

Proposition 3.1 shows that the variable selection procedures defined by (3.4) and (3.8), although apparently different, are indeed equivalent. Moreover, in the proof of Proposition 3.1 we will see that the minimum in the expressions of Q_1 and Q_2 is achieved at the value

$$(\alpha_1^*,\ldots,\alpha_p^*)' = \Sigma_{T_p}^{-1} c_{T_p},$$

where $c_{T_p} = \big(\mathrm{cov}\big(X(t_1),Y\big),\ldots,\mathrm{cov}\big(X(t_p),Y\big)\big)'$ and Σ_{T_p} is the covariance matrix of $X(T_p)$, for $T_p = (t_1,\ldots,t_p)$.

In addition, we show that the Q_1 and Q_2-based criteria are also both equivalent to a third criterion, defined in terms of a functional Q_0, which only depends on the covariances $K(t_i,t_j)$ and $\mathrm{cov}\big(X(t_i),Y\big)$ for $i,j = 1,\ldots,p$. This Q_0 criterion turns out to be especially useful to implement the method in practice.

Proposition 3.1. *Assume that Y and X fulfill the RKHS functional regression model in* (3.3)*. Then,*

$$\operatorname*{argmin}_{T_p\in\Theta_p} Q_1(T_p) = \operatorname*{argmin}_{T_p\in\Theta_p} Q_2(T_p) = \operatorname*{argmax}_{T_p\in\Theta_p} Q_0(T_p), \tag{3.9}$$

where Q_1 and Q_2 are defined in (3.4) *and* (3.8) *respectively, and*

$$Q_0(T_p) := {c_{T_p}}'\Sigma_{T_p}^{-1} c_{T_p}, \tag{3.10}$$

with c_{T_p} and Σ_{T_p} as in the previous paragraphs.

Proof. Since $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}\big[\varepsilon\big(X(t) - m(t)\big)\big] = 0$ and $\langle X,\beta\rangle_K \in \mathcal{L}(X)$,

$$\Big\|Y - \sum_{j=1}^{p} \alpha_j\big(X(t_j) - m(t_j)\big)\Big\|^2 \;=\; \Big\|\langle X,\beta\rangle_K - \sum_{j=1}^{p} \alpha_j\big(X(t_j) - m(t_j)\big)\Big\|^2 + \sigma^2.$$

On the other hand, Loève's isometry implies

$$\Big\|\langle X,\beta\rangle_K - \sum_{j=1}^{p} \alpha_j\big(X(t_j) - m(t_j)\big)\Big\|^2 = \Big\|\beta - \sum_{j=1}^{p} \alpha_j K(t_j,\cdot)\Big\|_K^2.$$

From the last two equations, it follows that $Q_1(T_p) = Q_2(T_p) + \sigma^2$ and hence the first equality in (3.9).

By the reproducing property,

$$\Big\|\beta - \sum_{j=1}^{p} \alpha_j K(t_j, \cdot)\Big\|_K^2 = \|\beta\|_K^2 - 2\sum_{j=1}^{p} \alpha_j \beta(t_j) + \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j K(t_i, t_j). \tag{3.11}$$

The function K is positive semidefinite so that the expression in (3.11) defines a convex function in $\alpha = (\alpha_1, \ldots, \alpha_p)$. By computing its gradient (with respect to α) it is very easy to see that the minimum is achieved at $\alpha^* = (\alpha_1^*, \ldots, \alpha_p^*)' = \Sigma_{T_p}^{-1}\beta(T_p)$, where $\beta(T_p) = (\beta(t_1), \ldots, \beta(t_p))'$. Then,

$$Q_2(T_p) = \Big\|\beta - \sum_{j=1}^{p} \alpha_j^* K(t_j, \cdot)\Big\|_K^2 = \|\beta\|_K^2 - \beta(T_p)'\Sigma_{T_p}^{-1}\beta(T_p). \tag{3.12}$$

Finally, using $\mathbb{E}\big[\big(X(t) - m(t)\big)\varepsilon\big] = 0$ and Equation (1.3) we get

$$\operatorname{cov}\big(Y, X(t)\big) = \mathbb{E}\big[\langle X, \beta\rangle_K \big(X(t) - m(t)\big)\big] = \Psi_X(\langle X, \beta\rangle_K)(t) = \beta(t).$$

To obtain the last equality, recall that $\langle X, \beta\rangle_K = \Psi_X^{-1}(\beta)$. Therefore $\beta(T_p) = c_{T_p}$ and, by (3.12), $Q_2(T_p) = \|\beta\|_K^2 - Q_0(T_p)$. This implies the second equality in (3.9). □

The criterion provided by Q_0 (or Q_1, Q_2) for variable selection was already considered by McKeague and Sen (2010), for $p = 1$, when $X(t)$ is a fractional Brownian Motion with Hurst exponent $H \in (0, 1)$, and by Ji and Müller (2017) for $p \geq 1$ in the usual L^2 functional regression model. The RKHS formalism we incorporate here provides a simple way to describe the scenario under which variable selection would lead to the optimal solution (with no loss of information). Variable selection is specially suitable when the true regression model is sparse, meaning that the response depends on the explanatory variables through their values at a finite small number of p^* points. As it was mentioned before, this is the case under (3.3) when

$$\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, t). \tag{3.13}$$

Let $T_{p^*}^* = (t_1^*, \ldots, t_{p^*}^*) \in \Theta_{p^*}$. Then, it is clear that, under (3.13),

$$Q_2(T_{p^*}^*) = 0 \leq Q_2(T_{p^*}), \quad \text{for all } T_{p^*} \in \Theta_{p^*}.$$

As a consequence, the true set of relevant variables $T_{p^*}^*$ is the one selected by the optimization of the functions in Proposition 3.1. In this reasoning we have considered the case when we know the actual number of points p^* to be selected. In practice, this is not usually the case. However, notice that if we make a conservative choice, taking a number of variables p larger than the true one ($p > p^*$), the true relevant variables $T_{p^*}^*$ will always be included among the selected ones. Indeed, if the true

model is $Y = \sum_{j=1}^{p^*} \beta_j^* X(t_j^*) + \varepsilon$, this means that the orthogonal projection of Y on the space $\mathcal{L}(X)$ is $\sum_{j=1}^{p^*} \beta_j^* X(t_j^*)$. Then, assume that we try to fit (in the β_j's and the t_j's) an "overparameterized" model of type $Y = \sum_{j=1}^{p} \beta_j X(t_j) + \varepsilon$ with $p > p^*$. Since the projection is unique, the optimal fit under the second model must coincide with the optimum of the "true" model. So, it must necessarily include the variables $t_1^*, \ldots, t_p^*$ and the coefficients β_i for the remaining variables must be zero. Therefore the optimal set T_{p^*} of t_i^*'s in the first model must be included in the optimal set T_p for the second one.

Recall that we use the notation $T^* \prec T \in \Theta_p$ meaning that T^* is a sub-vector of T, that is, that the components of T^* are included within those of T. With this notation, what we have shown is that, under (3.13), $T^* = (t_1^*, \ldots, t_{p^*}^*) \prec \operatorname{argmax} Q_0(T_p)$, for $p \geq p^*$. In Section 3.4 we address the problem of estimating p^* when it is unknown.

3.2.3 A recursive expression

The function Q_0 defined in (3.10) can be rewritten in an alternative way, which is useful to analyze the gain when we add a new variable to a set of variables already selected. Moreover, this alternative expression paves the way for a sequential implementation of the variable selection method. Besides the notation c_{T_p} and Σ_{T_p}, introduced earlier, we will also use c_j to denote $\operatorname{cov}\big(X(t_j), Y\big)$, σ_j^2 to denote $\operatorname{var}\big(X(t_j)\big)$, and $c_{T_p,j}$ to denote the vector $\big(\operatorname{cov}\big(X(t_1), X(t_j)\big), \ldots, \operatorname{cov}\big(X(t_p), X(t_j)\big)\big)'$.

Proposition 3.2. *Given $T_{p+1} = (t_1, \ldots, t_{p+1}) \in \Theta_{p+1}$, $p \geq 1$, and $T_p \prec T_{p+1}$, for some $p \geq 1$ such that the covariance matrices $\Sigma_{T_{p+1}}$ of the process are invertible for all $T_{p+1} \in \Theta_{p+1}$,*

$$Q_0(T_{p+1}) \;=\; Q_0(T_p) \;+\; \frac{\Big({c_{T_p}}'\Sigma_{T_p}^{-1} c_{T_p,p+1} - c_{p+1}\Big)^2}{\sigma_{p+1}^2 - {c_{T_p,p+1}}'\Sigma_{T_p}^{-1} c_{T_p,p+1}}. \tag{3.14}$$

Proof. We have to rewrite the expression ${c_{T_{p+1}}}'\Sigma_{T_{p+1}}^{-1} c_{T_{p+1}}$, where $p \geq 1$. We can write the matrix $\Sigma_{T_{p+1}}$ in block form as

$$\Sigma_{T_{p+1}} = \left(\begin{array}{ccc|c} & \Sigma_{T_p} & & \begin{array}{c}\operatorname{cov}\big(X(t_1), X(t_{p+1})\big)\\ \vdots \\ \operatorname{cov}\big(X(t_p), X(t_{p+1})\big)\end{array} \\ \hline \operatorname{cov}\big(X(t_1), X(t_{p+1})\big) & \ldots & \operatorname{cov}\big(X(t_p), X(t_{p+1})\big) & \operatorname{cov}\big(X(t_{p+1}), X(t_{p+1})\big) \end{array}\right)$$

$$\equiv \left(\begin{array}{c|c} \Sigma_{T_p} & c_{T_p,p+1} \\ \hline {c_{T_p,p+1}}' & \sigma_{p+1}^2 \end{array}\right).$$

Then its inverse matrix is

$$\Sigma_{T_{p+1}}^{-1} \;=\; \left(\begin{array}{c|c} \Sigma_{T_p}^{-1} + \frac{1}{a}\Sigma_{T_p}^{-1} c_{T_p,p+1} c_{T_p,p+1}{}' \Sigma_{T_p}^{-1} & -\frac{1}{a}\Sigma_{T_p}^{-1} c_{T_p,p+1} \\ \hline -\frac{1}{a} c_{T_p,p+1}{}' \Sigma_{T_p}^{-1} & \frac{1}{a} \end{array}\right),$$

where $a = \sigma_{p+1}^2 - c_{T_p,p+1}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1}$. We can also write the vector of covariances as

$$c_{T_{p+1}}{}' \;=\; \big(\mathrm{cov}\big(X(t_1), Y\big), \ldots, \mathrm{cov}\big(X(t_{p+1}), Y\big)\big) \;=\; (c_{T_p} \mid c_{p+1}).$$

Using this notation we can rewrite the original expression as follows,

$$\begin{aligned} c_{T_{p+1}}{}' \Sigma_{T_{p+1}}^{-1} c_{T_{p+1}} =& c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p} + \frac{1}{a} c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1} c_{T_p,p+1}{}' \Sigma_{T_p}^{-1} c_{T_p} - \frac{c_{p+1}}{a} c_{T_p,p+1}{}' \Sigma_{T_p}^{-1} c_{T_p} \\ & - \frac{c_{p+1}}{a} c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1} + \frac{c_{p+1}^2}{a} \\ =& c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p} + \frac{1}{a}\left[\left(c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1}\right)^2 - 2 c_{p+1} c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1} + c_{p+1}^2\right] \\ =& c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p} + \frac{\left(c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1} - c_{p+1}\right)^2}{\sigma_{p+1}^2 - c_{T_p,p+1}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1}}, \end{aligned}$$

since the product $c_{T_p}{}' \Sigma_{T_p}^{-1} c_{T_p,p+1}$ is actually a real number. □

This last proposition is useful to simplify other derivations in the chapter. Equation (3.14) already appears in the well-known forward selection method for variable selection in multiple regression (see, e.g., Miller (2002), Section 3.2). A modification of the resulting expression is also used in the variable selection method proposed by Yenigün and Rizzo (2015), still in the multivariate regression setting. In such alternative version, the usual covariance is replaced by the distance covariance, defined in Székely et al. (2007).

The quotient in Equation (3.14) can be written in a more insightful way, as shown in the following result.

Proposition 3.3. *In the above defined setup, denoting* $X(T_p) = \big(X(t_1), \ldots, X(t_p)\big)'$, *and being* Y_{T_p} *and* $X(t_{p+1})_{T_p}$ *the projections on the closed subspace* $\mathrm{span}\{X(t_i) - m(t_i), t_i \in T_p\}$ *of* Y *and* $X(t_{p+1})$ *respectively,*

$$Q_0(T_{p+1}) \;=\; Q_0(T_p) \;+\; \frac{\mathrm{cov}^2\big(Y - Y_{T_p},\, X(t_{p+1})\big)}{\mathrm{var}\big(X(t_{p+1}) - X(t_{p+1})_{T_p}\big)}. \tag{3.15}$$

Proof. Using the notation of the statement, if $\widetilde{X}(t)$ is the centered process, we have $Y_{T_p} = c_{T_p}{}' \Sigma_{T_p}^{-1} \widetilde{X}(T_p)$. Thus we can rewrite the numerator of the quotient of Equation (3.14) as

$$\mathrm{cov}\big(Y - Y_{T_p}, X(t_{p+1})\big) \;=\; c_{p+1} - \mathrm{cov}\big(Y_{T_p}, X(t_{p+1})\big)$$

$$\begin{aligned} &= c_{p+1} - {c_{T_p}}'\Sigma_{T_p}^{-1}\mathrm{cov}\big(\widetilde{X}(T_p), X(t_{p+1})\big) \\ &= c_{p+1} - {c_{T_p}}'\Sigma_{T_p}^{-1}c_{T_p,p+1}, \end{aligned}$$

since the covariances are not affected by the centering. For the denominator, we have (taking into account that $X(t_{p+1})_{T_p} = {c_{T_p,p+1}}'\Sigma_{T_p}^{-1}\widetilde{X}(T_p)$),

$$\begin{aligned} &\mathrm{var}\big(X(t_{p+1}) - X(t_{p+1})_{T_p}\big) \\ &\quad = \mathrm{var}\big(X(t_{p+1})\big) + \mathrm{var}\big(X(t_{p+1})_{T_p}\big) - 2\mathrm{cov}\big(X(t_{p+1}), X(t_{p+1})_{T_p}\big) \\ &\quad = \sigma_{p+1}^2 + {c_{T_p,p+1}}'\Sigma_{T_p}^{-1}c_{T_p,p+1} - 2{c_{T_p,p+1}}'\Sigma_{T_p}^{-1}\mathrm{cov}\big(X(t_{p+1}), \widetilde{X}(T_p)\big) \\ &\quad = \sigma_{p+1}^2 - {c_{T_p,p+1}}'\Sigma_{T_p}^{-1}c_{T_p,p+1}. \end{aligned}$$

From these two expressions the conclusion follows straightforwardly. □

The quotient of Equation (3.15) is known as *part correlation coefficient* or *semi-partial correlation coefficient*, a quantity which appears in several techniques dealing with multivariate data. Actually, it can be proved that this quotient tends to zero when t_{p+1} tends to one of the points in T_p, so that selecting a point too close to one of those already selected is redundant and non-informative according to this criterion.

Proposition 3.4. *Given a process with continuous covariance function and under the same hypotheses of Proposition 3.2, $Q_0(T_{p+1})$ converges to $Q_0(T_p)$ when $t_{p+1} \to t_j$ for some $1 \leq j \leq p$.*

Proof. It is equivalent to see that the quotient of Equation (3.15) tends to zero. As mentioned in the last proof, the numerator of this quotient equals $\mathrm{cov}\big(Y - Y_{T_p}, X(t_{p+1}) - X(t_{p+1})_{T_p}\big)$. Since $0 \leq \mathrm{var}(x-y) = \mathrm{var}(x) + \mathrm{var}(y) - 2\mathrm{cov}(x,y)$, the quotient can be bounded as

$$\frac{\mathrm{cov}^2\big(Y - Y_{T_p},\, X(t_{p+1})\big)}{\mathrm{var}\big(X(t_{p+1}) - X(t_{p+1})_{T_p}\big)} \leq \frac{1}{4}\frac{\big(\mathrm{var}\big(Y - Y_{T_p}\big) + \mathrm{var}\big(X(t_{p+1}) - X(t_{p+1})_{T_p}\big)\big)^2}{\mathrm{var}\big(X(t_{p+1}) - X(t_{p+1})_{T_p}\big)}. \quad (3.16)$$

Without loss of generality we can assume that $t_{p+1} \to t_p$. We first check that the variance of the denominator goes to zero. On the first hand, by hypothesis all the variables $\{X(t_1), \ldots, X(t_p)\}$ are linearly independent, which implies $X(t_p)_{T_p} = X(t_p)$. On the other hand,

$$\begin{aligned} \big\|X(t_{p+1}) - X(t_p)\big\|^2 &= \mathrm{var}\big(X(t_{p+1})\big) + \mathrm{var}\big(X(t_p)\big) - 2\mathrm{cov}\big(X(t_{p+1}), X(t_p)\big) \\ &= K(t_{p+1}, t_{p+1}) + K(t_p, t_p) - 2K(t_{p+1}, t_p) \;\to\; 0 \end{aligned}$$

since K is continuous in $[0,1]^2$. Thus,

$$\mathrm{var}\big(X(t_{p+1}) - X(t_{p+1})_{T_p}\big) \;\leq\; \big(\big\|X(t_{p+1}) - X(t_p)_{T_p}\big\| + \big\|X(t_p)_{T_p} - X(t_{p+1})_{T_p}\big\|\big)^2,$$

and the second term also goes to zero as $t_{p+1} \to t_p$ since the projection on $\mathrm{span}\{X(t_i) - m(t_i), t_i \in T_p\}$ is a continuous function in $L^2(\Omega)$. Then, the limit of the right hand side of Equation (3.16) is equal to $\lim_{x\to 0}(y+x)^2x^{-1} = 0$. □

3.3 Sample properties of the variable selection method

3.3.1 The proposed method

In order to carry out the variable selection in practice, we have to estimate the function Q_0 from a sample $(y_1, x_1), \ldots, (y_n, x_n)$ of independent observations drawn from the model (3.3). The most natural estimator is given by $\widehat{Q}_0(T_p) = \widehat{c}_{T_p}{}'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p}$, where $\widehat{c}_{T_p}$ and $\widehat{\Sigma}_{T_p}$ are the sample versions of c_{T_p} and Σ_{T_p}, respectively, based on the sample mean $\bar{x}(t) = n^{-1}\sum_{i=1}^{n} x_i(t)$ and the sample covariances

$$\widehat{\operatorname{cov}}\big(X(s), X(t)\big) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i(s)x_i(t) - \bar{x}(s)\bar{x}(t)$$

of the trajectories. Then, if we want to select p variables we propose to use $\widehat{T}_{p,n}$, where

$$\widehat{T}_{p,n} := \operatorname*{argmax}_{T_p \in \Theta_p} \widehat{Q}_0(T_p) = \operatorname*{argmax}_{T_p \in \Theta_p} \widehat{c}_{T_p}{}'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p}. \tag{3.17}$$

In practice, the number of combinations of variables is usually too large to carry out an exhaustive search to find the optimal p^* variables, even for small values of p^*. Then, we need to define a search strategy to perform the selection. That is, we must decide how to explore the space of all possible combinations of variables. We propose to use the sequential approach we describe below.

Observe that a proof analogous to that of Equations (3.14) and (3.15) also gives their corresponding sample versions:

$$\begin{aligned}
\widehat{Q}_0(T_{p+1}) &= \widehat{Q}_0(T_p) + \frac{\big(\widehat{c}_{T_p}{}'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p,p+1} - \widehat{c}_{p+1}\big)^2}{\widehat{\sigma}^2_{p+1} - \widehat{c}_{T_p,p+1}{}'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}_{T_p,p+1}}, \\
\widehat{Q}_0(T_{p+1}) &= \widehat{Q}_0(T_p) + \frac{\widehat{\operatorname{cov}}^2\big(Y - \widehat{Y}_{T_p},\, X(t_{p+1})\big)}{\widehat{\operatorname{var}}\big(X(t_{p+1}) - \widehat{X}(t_{p+1})_{T_p}\big)}.
\end{aligned} \tag{3.18}$$

These equations suggest a sequential way to carry out the variable selection. Initially it is selected the point $t_1 \in [\delta, 1]$ which maximizes the previous quotient for $p = 1$ (which equals $\widehat{\operatorname{cov}}^2\big(Y, X(t_1)\big)\widehat{\operatorname{var}}\big(X(t_1)\big)^{-1}$). Then, in each step, we find the variable $t_{p+1} \in [\delta, 1]$, $p > 1$, maximizing the equation above. In this way, we obtain nested subsets of variables, since $T_p \prec T_{p+1}$. This greedy method does not guarantee the convergence to the global maximum of $\widehat{Q}_0$, but it shows a good behavior in practice, as we will show later on.

3.3.2 Asymptotic results

In the following results, we will analyze the asymptotic behavior of the estimator proposed in (3.17). We start with three preliminary results that may be of some interest by themselves. First we prove that, under some moment conditions, the sample mean and covariance functions of X converge uniformly a.s. to their population counterparts:

Lemma 3.5. *Assume that the process X has continuous trajectories with continuous mean and covariance functions and that it fulfills that $\mathbb{E}\big[\sup_{t\in[\delta,1]} X(t)^2\big] < \infty$, for a certain $\delta \geq 0$. Then,*

$$\sup_{s,t\in[\delta,1]} |\widehat{\mathrm{cov}}\big(X(s),X(t)\big) - \mathrm{cov}\big(X(s),X(t)\big)| \overset{\text{a.s.}}{\to} 0. \tag{3.19}$$

Proof. Note that the assumption implies $\mathbb{E}\big[\sup_{t\in[\delta,1]} |X(t)|\big] < \infty$ and the stochastic process $\{X(t) : t \in [\delta,1]\}$ has finite strong expectation with trajectories in $\mathcal{C}[\delta,1]$, which is a separable Banach space. Then, we can apply Mourier's SLLN (see, e.g., Theorem 4.5.2 of Laha and Rohatgi (1979), p. 452) to conclude

$$\sup_{t\in[\delta,1]} |\bar{x}(t) - m(t)| \overset{a.s.}{\to} 0, \tag{3.20}$$

Similarly, the process $Z(s,t) := X(s)X(t)$, with trajectories in $\mathcal{C}([\delta,1]^2)$, is such that its strong expectation exists. Indeed, since

$$0 \leq \big(|X(s)| - |X(t)|\big)^2 = |X(s)|^2 + |X(t)|^2 - 2|Z(s,t)|,$$

it holds

$$\mathbb{E}\big[\sup_{s,t\in[\delta,1]} |Z(s,t)|\big] \leq \mathbb{E}\big[\sup_{t\in[\delta,1]} |X(t)|^2\big] < \infty.$$

Moreover, $\mathcal{C}([\delta,1]^2)$ is separable since $[\delta,1]^2$ is compact. Then, Mourier's SLLN and (3.20) imply (3.19). □

Next, we prove that both $\widehat{Q}_0$ and Q_0 are continuous functions for any $p \geq 1$:

Lemma 3.6. *Assume that the process $X(t)$ has continuous mean and covariance functions. Let $p \geq 1$ and $\Theta_p = \Theta_p(\delta)$ be such that the assumptions of Lemma 3.5 hold. In addition, assume that the covariance matrix Σ_{T_p} is invertible for all $T_p \in \Theta_p$. Then, the functions $\widehat{Q}_0$ and Q_0 are continuous on Θ_p.*

Proof. Fix $p \geq 1$. First, we prove that Q_0 is continuous. Since the process $X(t)$ has continuous mean and covariance functions we have that

$$c_{T_p} = \big(\mathrm{cov}\big(X(t_1),Y\big),\ldots,\mathrm{cov}\big(X(t_p),Y\big)\big)'$$

is continuous on Θ_p. On the other hand, since the entries of Σ_{T_p} are continuous on $[0,1]^2$, $\det(\Sigma_{T_p})$ is also continuous on Θ_p, where $\det(\Sigma)$ stands for the determinant of Σ. By assumption, $\det(\Sigma_{T_p}) > 0$ for all $T_p \in \Theta_p$. Since Θ_p is compact, $\inf_{T_p \in \Theta_p} \det(\Sigma_{T_p}) > 0$. Observe that

$$\Sigma_{T_p}^{-1} = \frac{\operatorname{adj}(\Sigma_{T_p})}{\det(\Sigma_{T_p})},$$

where $\operatorname{adj}(\Sigma)$ denotes the adjugate of Σ. As a consequence, the entries of $\Sigma_{T_p}^{-1}$ are continuous on Θ_p, and hence the function Q_0 is also continuous.

The proof for $\widehat{Q}_0$ is analogous with the only difference that in this case we must ensure that $\inf_{T_p \in \Theta_p} \det(\widehat{\Sigma}_{T_p}) > 0$ with probability 1. On the other hand, from (3.19) it follows that

$$\sup_{T_p \in \Theta_p} |\det(\widehat{\Sigma}_{T_p}) - \det(\Sigma_{T_p})| \xrightarrow{\text{a.s.}} 0. \tag{3.21}$$

We have seen before that $\inf_{T_p \in \Theta_p} \det(\Sigma_{T_p}) > 0$. Then, with probability 1, there exists n_0 such that if $n \geq n_0$, $\inf_{T_p \in \Theta_p} \det(\widehat{\Sigma}_{T_p}) > 0$. □

The two previous lemmas allow us to prove the uniform convergence on Θ_p of the empirical criterion for variable selection to the theoretical one.

Lemma 3.7. *Under the assumptions of Lemma 3.6, it holds that*

$$\sup_{T_p \in \Theta_p} |\widehat{Q}_0(T_p) - Q_0(T_p)| \xrightarrow{\text{a.s.}} 0.$$

Proof. It is enough to establish the uniform convergence a.s. of the coordinates of $\widehat{c}_{T_p}$ and the entries of $\widehat{\Sigma}_{T_p}^{-1}$ to those of c_{T_p} and $\Sigma_{T_p}^{-1}$ respectively.

Equation (3.20) and the same argument leading to (3.20) but applied to the process $Z(t) = X(t)Y$ yield

$$\sup_{t \in [\delta,1]} \Big| \frac{1}{n} \sum_{i=1}^{n} \big(x_i(t) - \bar{x}(t)\big)(y_i - \bar{y}) - \operatorname{cov}\big(X(t), Y\big) \Big| \xrightarrow{\text{a.s.}} 0, \tag{3.22}$$

and hence the uniform convergence a.s. of the coordinates of $\widehat{c}_{T_p}$ to those of c_{T_p}.

Finally, observe that $\widehat{\Sigma}_{T_p}^{-1} = \det(\widehat{\Sigma}_{T_p})^{-1} \operatorname{adj}(\widehat{\Sigma}_{T_p})$. Then, since $\inf_{T_p \in \Theta_p} \det(\Sigma_{T_p}) > 0$ and from (3.19), (3.21) we conclude the uniform convergence a.s. of the entries of $\widehat{\Sigma}_{T_p}^{-1}$ to those of $\Sigma_{T_p}^{-1}$. □

Now assume that the sparsity condition (3.13) holds. Then, $T_{p^*}^* = (t_1^*, \ldots, t_{p^*}^*) \in \Theta_{p^*}$ is "sufficient" in the sense that the response only depends on the regressor variable through the values $X(t_1^*), \ldots, X(t_{p^*}^*)$. We have already seen that $T_{p^*}^*$ is a global maximum of

Q_0 (see the remark below Equation (3.13)). In fact, we are going to prove that under mild conditions it is the only global maximum of Q_0 on Θ_{p^*} and that the estimator $\widehat{T}_{p^*,n}$ (defined in (3.17) with $p = p^*$) converges a.s. to $T^*_{p^*}$. Then, our proposal is able to identify consistently the true relevant points.

Theorem 3.8. *Assume* (3.13) *holds, that the process $X(t)$ has continuous mean and covariance functions and that the covariance matrix $\Sigma_{T_{p^*} \cup S_{p^*}}$ is invertible for all $T_{p^*}, S_{p^*} \in \Theta_{p^*}$, with $T_{p^*} \neq S_{p^*}$. Let $\Theta_{p^*} = \Theta_{p^*}(\delta)$ be such that the assumptions of Lemma 3.5 hold. Then,*

(a) The point $T^*_{p^*} \in \Theta_{p^*}$, given by (3.13)*, is the only global maximum of Q_0 on Θ_{p^*}.*

(b) If $\widehat{T}_{p^*,n} = \operatorname{argmax}_{T_{p^*} \in \Theta_{p^*}} \widehat{Q}_0(T_{p^*})$, then $\widehat{T}_{p^*,n} \to T^*_{p^*}$ a.s. as $n \to \infty$.

(c) $\widehat{T}_{p^*,n}$ converges to $T^*_{p^*}$ in quadratic mean, that is, $\mathbb{E}\|\widehat{T}_{p^*,n} - T^*_{p^*}\|_2^2 \to 0$, as $n \to \infty$, where $\|\cdot\|_2$ stands for the usual Euclidean norm in $\mathbb{R}^p$.

Proof. (a) In view of (3.9), it is enough to prove that $T^* := T^*_{p^*}$ is the unique global minimum of

$$Q_1(T_{p^*}) = \|Y - Y_{T_{p^*}}\|^2 = \|Y_{T^*} - Y_{T_{p^*}}\|^2 + \operatorname{var}(\varepsilon).$$

The expression above readily shows that T^* minimizes Q_1. Suppose that there exists another minimum $S^* \in \Theta_{p^*}$ such that $S^* \neq T^*$. Then, we must have $\|Y_{T^*} - Y_{S^*}\|^2 = 0$ and hence $Y_{T^*} - Y_{S^*} = 0$ a.s. As a consequence, using the notation $\widetilde{X}(t) = X(t) - m(t)$, there exist coefficients β_j and α_j such that $\sum_{j=1}^{p^*} \beta_j \widetilde{X}(t^*_j) - \sum_{j=1}^{p^*} \alpha_j \widetilde{X}(s^*_j) = 0$ a.s., for $T^*, S^* \in \Theta_{p^*}$ with $S^* \neq T^*$. This fact contradicts the assumption that the covariance matrix $\Sigma_{T^* \cup S^*}$ must be invertible. Therefore, $T^* = S^*$.

(b) Since the functions $\widehat{Q}_0$ and Q_0 are continuous on Θ_{p^*} (by Lemma 3.6) and the sequence of functions $\widehat{Q}_0$ tends uniformly a.s. to Q_0 on Θ_{p^*} (by Lemma 3.7) the fact that Q_0 has a unique maximum on Θ_{p^*} (part (a)) implies that $\widehat{T}_{p^*,n}$ converges almost surely to $T^*_{p^*}$.

(c) From part (b), we have $\|\widehat{T}_{p^*,n} - T^*_{p^*}\|_2 \to 0$ (Euclidean norm of the vector) a.s. as $n \to \infty$. Moreover, since both $\widehat{T}_{p^*,n}$ and $T^*_{p^*}$ belong to Θ_{p^*},

$$\|\widehat{T}_{p^*,n} - T^*_{p^*}\|_2 \leq \|\widehat{T}_{p^*,n}\|_2 + \|T^*_{p^*}\|_2 \leq 2p^*.$$

The result follows from dominated convergence theorem (using $2p^*$ as the integrable dominating function). □

Once we have selected p^* points, we can use them to predict the response variable. The optimal predictions (in a square mean sense) are given by:

$$\widehat{Y}_{\widehat{T}_{p^*}} = \widehat{\beta}_1\big(X(\widehat{t}_1) - \bar{x}(\widehat{t}_1)\big) + \cdots + \widehat{\beta}_{p^*}\big(X(\widehat{t}_{p^*}) - \bar{x}(\widehat{t}_{p^*})\big),$$

where $(\widehat{\beta}_1, \ldots, \widehat{\beta}_{p^*})' = \widehat{\Sigma}^{-1}_{\widehat{T}_{p^*}} \widehat{c}_{\widehat{T}_{p^*}}$. On the other hand, the prediction we would use under condition (3.13) if we knew the true relevant points and the true values of the parameters of the model would be

$$Y_{T^*} = \beta^*_1\big(X(t^*_1) - m(t^*_1)\big) + \cdots + \beta^*_{p^*}\big(X(t^*_{p^*}) - m(t^*_{p^*})\big),$$

where now $(\beta^*_1, \ldots, \beta^*_{p^*})' = \Sigma^{-1}_{T^*_{p^*}} c_{T^*_{p^*}}$. The following result refers to the asymptotic behavior of the data-driven predictions $\widehat{Y}_{\widehat{T}_{p^*}}$. It is shown that they converge a.s. and in quadratic mean to the oracle values $Y_{T^*_{p^*}}$.

Theorem 3.9. *Under the assumptions of Theorem 3.8,* $\widehat{Y}_{\widehat{T}_{p^*}} \overset{\text{a.s.}}{\to} Y_{T^*_{p^*}}$. *If, in addition, there exists* $\eta > 0$ *such that* $\mathbb{E}[\sup_{t\in[\delta,1]} |X(t)|^{2+\eta}] < \infty$ *then* $\widehat{Y}_{\widehat{T}_{p^*}} \xrightarrow{L^2} Y_{T^*_{p^*}}$, *as* $n \to \infty$.

Proof. For simplicity, denote $\widehat{T} := \widehat{T}_{p^*}$, $T^* := T^*_{p^*}$ and $\widetilde{X} = X - m$. Observe that

$$\begin{aligned}
|\widehat{Y}_{\widehat{T}} - Y_{T^*}| &= \big|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}}\big(X(\widehat{T}) - \bar{x}(\widehat{T})\big) - c_{T^*}{}'\Sigma^{-1}_{T^*}\big(X(T^*) - m(T^*)\big)\big| \\
&\le |\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T}) - c_{T^*}{}'\Sigma^{-1}_{T^*}\widetilde{X}(T^*)| + \big|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}}\big(\bar{x}(\widehat{T}) - m(\widehat{T})\big)\big| \\
&\le |\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T}) - c_{\widehat{T}}{}'\Sigma^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T})| + |c_{\widehat{T}}{}'\Sigma^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T}) - c_{T^*}{}'\Sigma^{-1}_{T^*}\widetilde{X}(T^*)| \\
&\quad + \big|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}}\big(\bar{x}(\widehat{T}) - m(\widehat{T})\big)\big| \\
&\le \|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}} - c_{\widehat{T}}{}'\Sigma^{-1}_{\widehat{T}}\|_2 \, \|\widetilde{X}(\widehat{T})\|_2 + |c_{\widehat{T}}{}'\Sigma^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T}) - c_{T^*}{}'\Sigma^{-1}_{T^*}\widetilde{X}(T^*)| \\
&\quad + \|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}^{-1}_{\widehat{T}}\|_2 \, \big\|\bar{x}(\widehat{T}) - m(\widehat{T})\big\|_2 \\
&\le \sup_{T\in\Theta_{p^*}} \|\widehat{c}_T{}'\widehat{\Sigma}^{-1}_T - c_T{}'\Sigma^{-1}_T\|_2 \sup_{T\in\Theta_{p^*}} \|\widetilde{X}(T)\|_2 \\
&\quad + |c_{\widehat{T}}{}'\Sigma^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T}) - c_{T^*}{}'\Sigma^{-1}_{T^*}\widetilde{X}(T^*)| \\
&\quad + \sup_{T\in\Theta_{p^*}} \|\widehat{c}_T{}'\widehat{\Sigma}^{-1}_T\|_2 \sup_{T\in\Theta_{p^*}} \big\|\bar{x}(T) - m(T)\big\|_2,
\end{aligned}$$

where $\|\cdot\|_2$ stands for the Euclidean vectorial norm.

Then, to prove $\widehat{Y}_{\widehat{T}} \to Y_{T^*}$ a.s. it is enough to see that the three addends of the last expression go to 0 a.s. Observe that $\sup_{T\in\Theta_{p^*}} \|\widehat{c}_T{}'\widehat{\Sigma}^{-1}_T - c_T{}'\Sigma^{-1}_T\|_2 \to 0$ a.s., as $n \to \infty$, by (3.19) and (3.22) (from proof of Lemma 3.7). Moreover, since $X(t)$ has continuous trajectories and continuous mean function, and Θ_{p^*} is compact, we have $\sup_{T\in\Theta_{p^*}} \|\widetilde{X}(T)\|_2 < \infty$, a.s. For the second addend, the continuity of c_T, Σ_T and $\widetilde{X}(T)$, together with Theorem 3.8(b), imply that $|c_{\widehat{T}}{}'\Sigma^{-1}_{\widehat{T}}\widetilde{X}(\widehat{T}) - c_{T^*}{}'\Sigma^{-1}_{T^*}\widetilde{X}(T^*)| \to 0$ a.s., as $n \to \infty$. Finally, $\sup_{T\in\Theta_{p^*}} \|\bar{x}(T) - m(T)\|_2 \le \sqrt{p^*} \sup_{s\in[\delta,1]} |\bar{x}(s) - m(s)|$, which goes to zero a.s. by Equation (3.20), and the same argument used for the first addend

leads to $\sup_{T\in\Theta_{p^*}} \|\widehat{c}_T{}'\widehat{\Sigma}_T^{-1}\|_2 \leq \sup_{T\in\Theta_{p^*}} \|c_T{}'\Sigma_T^{-1}\|_2 + \epsilon < \infty$ for some $\epsilon > 0$, since it is a continuous function (proof of Lemma 3.6) over the compact space Θ_{p^*}.

In order to prove $\widehat{Y}_{\widehat{T}} \xrightarrow{L^2} Y_{T^*}$, as $n \to \infty$, we will check that there exists $\eta > 0$ such that $\sup_n \mathbb{E}|\widehat{Y}_{\widehat{T}}|^{2+\eta} < \infty$, which in turn implies that the sequence $\widehat{Y}_{\widehat{T}}^2$ is uniformly integrable. Then, we can apply that a uniformly integrable sequence of random variables which converges in probability, also converges in L^1 (see, e.g., Proposition 6.3.2 and the corollary of Proposition 6.3.3 in Laha and Rohatgi (1979)).

Observe that

$$|\widehat{Y}_{\widehat{T}}|^{2+\eta} = \big|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}_{\widehat{T}}^{-1}\big(X(\widehat{T}) - \bar{x}(\widehat{T})\big)\big|^{2+\eta} \leq \|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}_{\widehat{T}}^{-1}\|_2^{2+\eta} \big\|X(\widehat{T}) - \bar{x}(\widehat{T})\big\|_2^{2+\eta}, \text{ a.s.}$$

We have seen that $\sup_{T\in\Theta_{p^*}} \|\widehat{c}_T{}'\widehat{\Sigma}_T^{-1} - c_T{}'\Sigma_T^{-1}\|_2 \to 0$ a.s., as $n \to \infty$. Then, given $\epsilon > 0$, for large enough n,

$$\|\widehat{c}_{\widehat{T}}{}'\widehat{\Sigma}_{\widehat{T}}^{-1}\|_2^{2+\eta} \leq \epsilon + \sup_{T\in\Theta_{p^*}} \|c_T{}'\Sigma_T^{-1}\|_2^{2+\eta} := C < \infty, \text{ a.s.}$$

By assumption, there exists $\eta > 0$ such that $\mathbb{E}\big[\sup_{t\in[\delta,1]} |\widetilde{X}(t)|^{2+\eta}\big] < \infty$. From the last two displayed equations and Equation (3.20), for large enough n,

$$\mathbb{E}|\widehat{Y}_{\widehat{T}}|^{2+\eta} \leq C\,\mathbb{E}\|X(\widehat{T}) - \bar{x}(\widehat{T})\|_2^{2+\eta} \leq C'\big(\epsilon + \mathbb{E}\big[\sup_{t\in[\delta,1]} |\widetilde{X}(t)|^{2+\eta}\big]\big) < \infty,$$

where $\epsilon > 0$ and $C' = C(p^*)^{(2+\eta)/2}$. Since the last upper bound does not depend on n, we get that the supremum $\sup_n \mathbb{E}|\widehat{Y}_{\widehat{T}}|^{2+\eta}$ is finite. □

3.4 Estimating the number of variables

As discussed above, our method for variable selection works, whenever the slope function β belongs to the RKHS associated with the covariance function K. It is based on the idea of asymptotically minimizing (on $T_p = (t_1, \ldots, t_p)$) the residuals

$$Q_1(T_p) := \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p} \Big\|Y - \sum_{j=1}^p \beta_j \widetilde{X}(t_j)\Big\|^2 = \Big\|Y - \sum_{j=1}^p \beta_j^* \widetilde{X}(t_j)\Big\|^2,$$

where $\widetilde{X}(t_j) = X(t_j) - m(t_j)$ and $(\beta_1^*, \ldots, \beta_p^*)' = \Sigma_T^{-1} c_T$, Σ_T being the covariance matrix of $(X(t_1), \ldots, X(t_p))'$ and c_T the vector whose j-th component is $\operatorname{cov}\big(X(t_j), Y\big)$. As proved in Proposition 3.1, this amounts to maximize the function Q_0 defined in (3.10), which in turn is equivalent to minimize the function Q_2, defined in (3.8). Also, the functions, Q_1 and Q_2 agree up to an additive constant and both agree with Q_0 up to a change of sign plus an additive constant.

Throughout this section we assume the validity of the sparsity assumption (3.7), that is, we assume that the slope function β has the form $\beta = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$, as stated in Equation (3.13), for some constants $\beta_1, \ldots, \beta_{p^*} \in \mathbb{R}$ and for $T_{p^*}^* = (t_1^*, \ldots, t_{p^*}^*)$. In this case, we can properly speak of a specific target set of "true" variables $T^* = T_{p^*}^* = (t_1^*, \ldots, t_{p^*}^*)$ to be selected and, in particular, of a "true" number p^* of variables to select.

Keeping in mind these facts, the following comments provide some clues and motivation for the data-based selection of p^*. They will be formalized in the statement and proof of Lemma 3.10.

(a) On the one hand, any selection of type $T_p = (t_1, \ldots, t_p)$ with $p < p^*$ is clearly sub-optimal, since it would lack some relevant information, contributed by the variables in T^* not in T_p.

(b) Likewise, a choice T_p "by excess" with $T^* \prec T_p$ would not provide any benefit. To see this note that, under (3.7), the minimum of Q_2 is obviously attained at T^* and the value of Q_2 at such minimum is 0, which cannot be improved.

(c) As a consequence, the maximum value of Q_2 for points with $p^* + 1$ coordinates is attained at some T_{p^*+1} such that $T^* \prec T_{p^*+1}$ (that is, T^* is a sub-vector of T_{p^*+1}) but, in any case, $Q_0(T_{p^*+1}) - Q_0(T^*) = 0$.

(d) Then, the optimal p^* is such that the maximum value of $Q_0(T_p)$ agrees with that of $Q_0(T_{p^*})$ for any T_p such that $T_{p^*} \prec T_p$. Thus p^* is in fact the "elbow" value in the plot of $p \mapsto Q_0(T_p^*)$ from which on the increase of the maximum values of Q_0 stops.

The following lemma will set the theoretical basis of our procedure of estimation of p^*. As a consequence of this result, a procedure to estimate p^* is proposed below.

Lemma 3.10. *Let us consider the model* (3.3) *under the assumption that β can be expressed as $\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$, where p^* is the minimal integer for which such representation holds. Define $\widehat{Q}_0^{\max}(p) = \max_{T_p \in \Theta_p} \widehat{Q}_0(T_p)$. Then, under the assumptions of Lemma 3.6 we have*

(a)

$$\widehat{Q}_0^{\max}(p^* + 1) - \widehat{Q}_0^{\max}(p^*) \to 0, \ \textit{a.s.}, \tag{3.23}$$

(b) for all $p < p^*$,

$$\lim_n \left(\widehat{Q}_0^{\max}(p + 1) - \widehat{Q}_0^{\max}(p)\right) > 0, \ \textit{a.s.}$$

Proof. (a) Let us first prove

$$Q_0(T_{p^*+1}^*) - Q_0(T_{p^*}^*) = 0. \tag{3.24}$$

To see this, note that in the proof of Proposition 3.1 we have proved $Q_0(T_p) = \|\beta\|_K^2 - Q_2(T_p)$ with $Q_2(T_p) = \min_{(\beta_1,\ldots,\beta_p)\in\mathbb{R}^p} \|\beta - \sum_{j=1}^p \beta_j K(t_j,\cdot)\|_K^2$. Also, under (3.13), $Q_2(T^*_{p^*}) = 0$ so that $Q_0(T^*_{p^*}) = \|\beta\|_K^2$, which is the maximum possible value of Q_0. On the other hand, it is clear that $Q_2(T^*_{p^*+1}) \leq Q_2(T^*_{p^*})$ so that we must also have $Q_2(T^*_{p^*+1}) = 0$ and $Q_0(T^*_{p^*+1}) = \|\beta\|_K^2$. This proves (3.24). Now conclusion (3.23) follows directly from the uniform convergence of $\widehat{Q}_0$ to Q_0, as established in Lemma 3.7.

(b) Similarly to part (a) we only need to prove

$$Q_0(T^*_{p+1}) - Q_0(T^*_p) > 0 \text{ for all } p < p^*. \tag{3.25}$$

Indeed, assume we have $Q_0(T^*_{p+1}) - Q_0(T^*_p) = 0$ for some $p < p^*$. Then, since the prediction error $Q_1(T_p)$ defined in (3.4) satisfies $Q_1(T_p) = -Q_0(T_p) + \|\beta\|_K^2 + \sigma^2$ we would have that the prediction error $Q_1(T^*_p)$ obtained with p variables in the sparse model $Y = \sum_{j=1}^q \beta_j X(t_j) + \varepsilon$, with $\text{var}(\varepsilon) = \sigma^2$ would be the same, for $q = p$ and $q = p + 1$. Then, by recurrence, we get that the error would be in fact the same, irrespective of the number of q explanatory variables in the range $p, \ldots, p^*$ (note that, in a linear regression model, if the incorporation of no individual variable reduces the residual error, no linear combination of such variables does). Thus, the linear model $Y = \langle \beta, X\rangle_K + \varepsilon$ holds for a slope function of type $\beta = \sum_{j=1}^p \beta_j K(t^*_j,\cdot)$ with $p < p^*$. This is a contradiction with the assumption that p^* is the minimal value for which such model holds.

Now the result follows from (3.25) and the a.s. uniform convergence of $\widehat{Q}_0$ to Q_0. □

This result suggests the following method to estimate p^*:

1. Define
$$\Delta = \min_{p<p^*} \left\{Q_0(T^*_{p+1}) - Q_0(T^*_p)\right\}. \tag{3.26}$$
Assume we are able to fix a value $\epsilon > 0$ such that $\epsilon < \Delta$.

2. Define
$$\widehat{p} = \min_p \left\{\widehat{Q}_0^{\max}(p+1) - \widehat{Q}_0^{\max}(p) < \epsilon\right\}, \tag{3.27}$$

In view of Equation (3.18), this difference can be rewritten as the quotient involved in that equation.

Theorem 3.11. *Under the assumptions of Lemma 3.10 the estimator $\widehat{p}$ defined in* (3.27) *fulfills $\widehat{p} \to p^*$, almost surely.*

Proof. This result is a direct consequence of Lemma 3.10. □

In practice, the calculation of Δ defined in Equation (3.26) is not feasible, since it is merely a theoretical bound. Thus, the restriction $\epsilon < \Delta$ should be understood as choosing a value ϵ small enough. In order to fix this value from the data, different approaches could be used. For instance in Delaigle et al. (2012), where empirical methods to select both p and T_p in functional classification are given, the authors suggest to set ϵ equal to $\rho\widehat{Q}_0^{\max}(p-1)$ for a pre-determined small ρ. Nevertheless, we suggest to use techniques inspired in the change point detection methodology in time series, which avoid the need of an additional parameter. We could interpret the collection of values $L_n(p) = \ln\left(\widehat{Q}_0^{\max}(p+1) - \widehat{Q}_0^{\max}(p)\right)$ for $p = 1, \ldots$ as a time series and apply the usual k-means clustering algorithm to these values, with $k = 2$. Then, ϵ is fixed as $L_n(p)$, for p the largest value such that $L_n(p)$ belongs to the same cluster as $L_n(1)$. This is equivalent to estimate $\widehat{p}$ directly as the minimum value of p such that all the values $L_n(p)$ with $p \geq \widehat{p}$ belong to a different cluster than that of $L_n(1)$. This is the approach used in the experimental setting exposed in Section 3.6 but, as it was the case with the sequential search of the points, this technique does not guarantee that the true number p^* is selected.

3.5 When p^* is not estimated: the conservative oracle property

Under the sparseness assumption (3.13), where p^* is unknown, another sensible approach for the choice of the number p of selected variables is to take a conservative, large enough value of p.

The basic idea of this section is easy to state: suppose that a "conservative oracle" gives us a value p such that $p > p^*$. Accordingly, we perform our variable selection procedure for such value p. This yields p variables $\widehat{t}_1, \ldots, \widehat{t}_p$. Then, we can be sure that the "true" variables $t_1^*, \ldots, t_{p^*}^*$ are very close to p^* variables in $\{\widehat{t}_1, \ldots, \widehat{t}_p\}$.

The next result formalizes this property.

Theorem 3.12. *Let us consider the model* (3.3) *under the assumption that β can be expressed as $\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$, where p^* is the minimal integer for which such representation holds. Let $\widehat{t}_1, \ldots, \widehat{t}_p$ be the variables selected by the method* (3.17)*, where p is a given value larger than p^*. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\Big(t_i^* \in \bigcup_{j=1}^{p} \big(\widehat{t}_j - \epsilon, \widehat{t}_j + \epsilon\big),\ i = 1, \ldots, p^*\Big) = 1, \ \textit{eventually, as } n \to \infty. \tag{3.28}$$

Proof. Recall that the choice of the variables $T_p = (t_1, \ldots, t_p)$ is performed by asymptotically maximizing the function $Q_0(T_p)$, defined in (3.10). More precisely, as Q_0 depends on unknown population quantities, we in fact maximize the estimator $\widehat{Q}_0$ defined in Subsection 3.3.1.

Now, let us note that the maximum of Q_0 is not unique. Indeed, we assume that the "minimal" sparse representation of β has the form $\beta(t) = \sum_{j=1}^{p^*} \beta_j K(t_j^*, \cdot)$ but, of course, if $p > p^*$, we may formally put $\beta(t) = \sum_{i=1}^{p} \beta_i K(s_i, \cdot)$ as long as the "true" optimal points $t_i^*, \ldots, t_{p^*}^*$ are among the s_i's and all the coefficients β_i not matching with such t_i^*'s are null. On the other hand, from the uniqueness of the function β, all the maxima of $Q_0(T_p)$ must have an expression of this type.

Let us assume that conclusion (3.28) does not hold. Then, with positive probability, there exists a subsequence of maxima $\widehat{T}_{p,n}^*$ of $\widehat{Q}_0$ such that the point t_1^*, for instance, is not contained in the union of $(\widehat{t}_j - \epsilon, \widehat{t}_j + \epsilon)$, $j = 1, \ldots, p$. Thus, with positive probability, there is a further subsequence (denoted again $\widehat{T}_{p,n}^*$) converging to some T_p^{**} whose coordinates are all at a distance of, at least, ϵ from t_1^*. According to Lemma 3.7, $\widehat{Q}_0(T_p)$ converges to $Q_0(T_p)$ uniformly a.s. in T_p. In particular this entails that, with positive probability, $\widehat{Q}_0(\widehat{T}_{p,n}^*)$ converges to $Q_0(T_p^{**})$, which contradicts the fact that T_p^{**} cannot be a maximum of Q_0. □

This result is reminiscent of the *Sure Screening Property* defined in Fan and Lv (2008), which is used to quantify the efficiency of multivariate variable selection methods. But, obviously, property (3.28) is adapted to cope with the functional nature of the data and the fact that the values t_i range on a continuous domain.

3.6 Experiments

The purpose of this section is to give some insights into the practical behavior of our proposal for variable selection, both in simulations and real data examples. We are aware that the design of these experiments is largely discretionary, as the range of possible models for simulation is potentially unlimited (especially in the case of functional data models) and there is also a considerable amount of real data examples currently available in the FDA literature. Still, our choices have not been completely arbitrary. We have tried to follow some objective criteria. First, the theoretical models chosen for the simulations must obviously include some situations in which our crucial "sparseness" assumption $\beta = \sum_j \beta_j K(t_j, \cdot)$ is fulfilled. As discussed above, such models are quite natural if we are willing to use variable selection techniques. Also, it looks reasonable to include at least one model in which this assumption is not valid. Regarding the real data, we have just chosen two examples used in the recent literature for the purpose of checking other variable selection methods in functional regression settings.

In any case, we would like to emphasize that we make here no attempt to draw any definitive conclusion on the performance of our method when compared with others. In our view, no unique empirical study can lead to safe, objective conclusions in this regard: the only reliable verdict should be given by the users community, after some

time of practice with real data problems. Our purposes here are far more unassuming; we just want to provide some hints suggesting that our proposal

(a) has a satisfactory performance in the "sparse" models for which it has been designed,

(b) can be implemented in practice, with an affordable computational cost,

(c) could be hopefully competitive under other theoretical models, far from the ideal assumption $\beta = \sum_j \beta_j K(t_j, \cdot)$,

(d) has also a satisfactory practical performance in a couple of real data examples commonly used in the literature of variable selection.

The R code used in the experiments can be provided on request.

3.6.1 Simulation experiments

Keeping in mind the above general lines, we next define the simulation models under study. In our context a "model" is defined by three elements: a stochastic process (from which the functional data are generated), a regression equation, of type $Y = \langle X, \beta \rangle_K + \varepsilon$ (or, more generally, $Y = g(X) + \varepsilon$) and an error variable ε. In what follows, ε has been chosen in all cases as $\varepsilon \sim \mathcal{N}(0, \sigma)$ with $\sigma = 0.2$.

We have considered several processes, covering a broad range of different situations.

1. *Standard Brownian Motion* (Bm) $\{B(t),\ t \in [0,1]\}$.
2. *Geometric Brownian Motion* (gBm). This non-Gaussian process is also known as exponential Brownian motion. It can be defined just by $X(t) = e^{B(t)}$.
3. *Integrated Brownian Motion* (iBm): it is obtained as $X(t) = \int_0^t B(s)\mathrm{d}s$. Note that the trajectories of this non-Markovian process are smooth.
4. *Ornstein-Uhlenbeck process* (OU). This is a Gaussian process $\{X(t)\}$ which satisfies the stochastic differential equation $\mathrm{d}X(t) = -\mu X(t)\mathrm{d}t + \sigma \mathrm{d}B(t)$. In our simulations we have chosen $\mu = \sigma = 1$.
5. *Fractional Brownian Motion* (fBM). This process is a generalization of the Brownian motion $B(t)$ but, unlike $B(t)$, it does not have (in general) independent increments. The mean function of this Gaussian process is identically 0 and its covariance function is $K(t,s) = 0.5(|t|^{2H} + |s|^{2H} - |t-s|^{2H})$, where $H \in (0,1)$ is the so-called Hurst exponent. Note that for $H = 0.5$, this process coincides with the standard Brownian Motion. Also, the trajectories of this process are still not differentiable at every point but the index H is closely related to the Hölder continuity properties of these trajectories. In particular, when $H > 0.5$, the trajectories look "more regular" than those of the Brownian motion, having

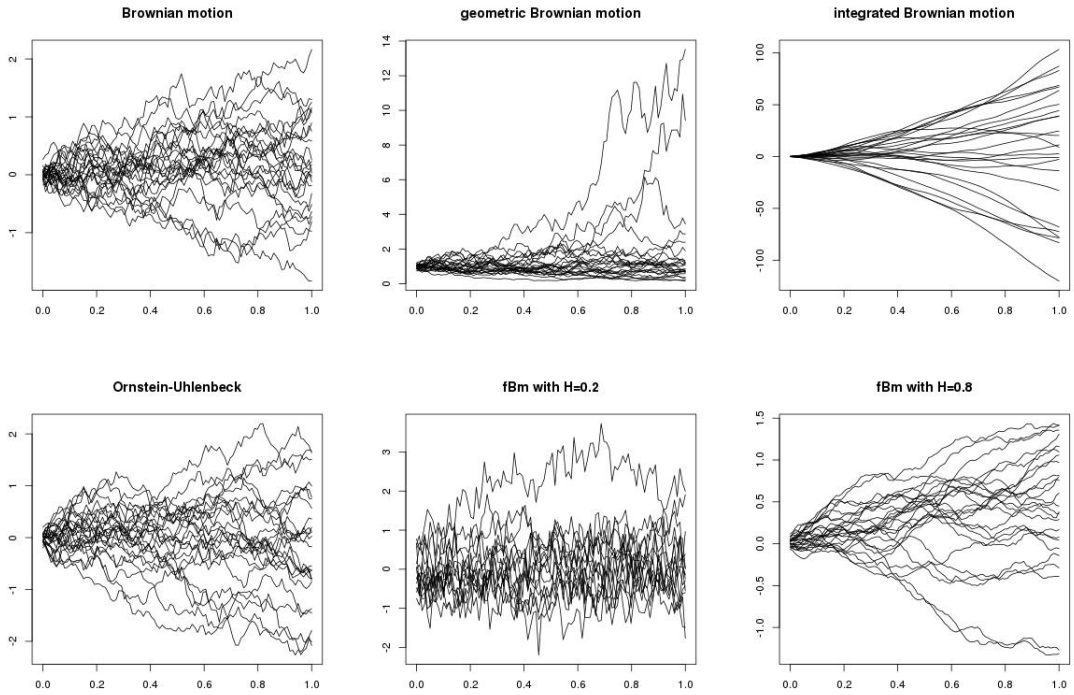


Figure 3.1: 25 trajectories of each of the processes used in the simulations.

a wilder appearance for $H < 0.5$. To cover both cases we have used $H = 0.2$ and $H = 0.8$ in our simulations.

Figure 3.1 shows some trajectories of each of these six processes, where the variables $X(t)$ for t in a neighborhood of 0 have been omitted to satisfy the non-degeneracy requirements of the method.

As for the regression function g we have considered the following three choices.

1. Two functions β in (3.3) of type $\beta(t) = \sum_j \beta_j K(t_j^*, \cdot)$ so that the regression model reduces to the "sparse version" (3.7). We have considered two different regression functions. For the first one, we have used the set of points $T^* = (0.2, 0.4, 0.9)$ with weights $(\beta_1, \beta_2, \beta_3) = (2, -5, 1)$; this is "Regression model 1" in the tables. For the second one, we have used $T^* = (0.16, 0.47, 0.6, 0.85, 0.91)$ and weights $(2.1, -0.2, -1.9, 5, 4.2)$ ("Regression model 2" in the tables). Therefore, the response variables in both cases are given, respectively, by

$$\begin{aligned} Y_1 &= 2X(0.2) - 5X(0.4) + X(0.9) + \varepsilon, \\ Y_2 &= 2.1X(0.16) - 0.2X(0.47) - 1.9X(0.67) + 5X(0.85) + 4.2X(0.91) + \varepsilon. \end{aligned}$$

2. $\beta(t) = \ln(1+t)$ and the regression model is (3.1) with $\alpha_0 = 0$. Thus, the sparse RKHS model (3.7) does not hold in this case. This is "Regression model 3" in the tables. It has been already used in Cuevas et al. (2002). Therefore, the corresponding response variable is generated by

$$Y_3 = \int_0^1 \ln(1+t)X(t)\mathrm{d}t + \varepsilon.$$

3.6.2 Real data

We have also checked the different methods when applied to two real data sets. Since these data have been already considered in other recent papers of the FDA literature, we will give only brief descriptions of them.

1. *Tecator*. This data set has been widely used in the literature. The trajectories $X(t)$ are 100 channel spectrum of absorbances of 215 meat samples, and the response variable Y is the fat content. However, (in most versions of this data set) 15 of these observations are repeated, so we have removed them. The remaining 200 trajectories are discretized on a grid of 100 points. As usual when working with these measurements, we use the second derivative of the curves instead of the original ones. One version of this data set (including repeated curves) can be found as part of the `fda.usc` R-package (see Febrero-Bande and de la Fuente (2012)). The version we have used in our experiment is available along with the R-code.

2. *Ash content in sugar samples*. This data set has been used, for example, in Aneiros and Vieu (2014). The version we use corresponds in fact to a subset of the whole data set, available in `http://www.models.kvl.dk/Sugar_Process`. The response variable Y is the percentage of ash content in 266 sugar samples. The trajectories $X(t)$ are the fluorescence spectra from 275 to 560 nm at excitation wavelength 290. These curves are discretized on a grid of 100 equispaced points.

Figure 3.2 shows some trajectories of both original data sets, although we work with centered version of the curves.

3.6.3 Methods under study and methodology

We compare our proposal with other methods for variable selection recently considered in the literature. We now list the methods under study along with the notation used in the tables below. For the first three methods we have used our own R implementations.

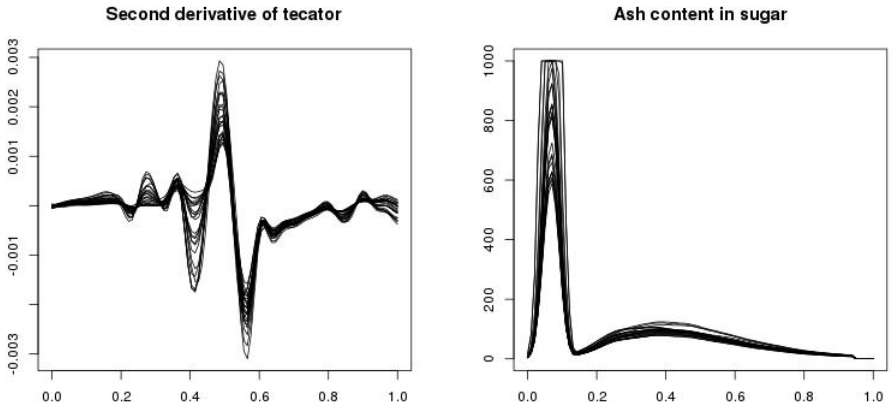


Figure 3.2: 25 trajectories of each of the real data sets.

1. The method proposed in this chapter (RKHS). It has been implemented using the iterative approximation described in equation (3.18). The number of relevant points is chosen as explained at the end of Section 3.4, by doing clustering on the values $L_n(p)$. Therefore, no validation technique is required.

2. The variable selection procedure proposed in Kneip et al. (2016) (KPS): in the original article, a mixed method for standard functional linear regression and variable selection technique is proposed. Since we are here concerned with variable selection, we have implemented just the corresponding part of the proposal. Essentially, the idea is to select the points (called "impact points" in Kneip et al. (2016)) maximizing the covariance between the response variable Y and a "decorrelated" version, $Z(t)$, of the original process. By construction, the decorrelated process $Z(t)$ is such that $Z(t)$ and $Z(s)$ are almost uncorrelated whenever $|t-s| \geq \delta$. The value δ and the number of selected variables are chosen using the BIC criterion, as proposed in the original paper.

3. *Partitioning Variable Selection* (PVS) with ML penalization, as proposed in Aneiros and Vieu (2014). The original sample must be split into two independent subsamples, which should be asymptotically of the same sizes. The basic idea is to apply some multivariate variable selection technique in this context, but taking advantage of the functional structure of the data. The procedure works in two steps. In the first step, one constructs an equispaced subgrid of variables among all the variables in the original grid (see below). Then a variable selection technique for multivariate data is applied on this subgrid, using the first subsample of the original data. For instance, we might use LASSO with ML penalization, as proposed in the original paper. Then, in the second step, this variable selection technique is applied again to an enlarged grid, constructed by taking all variables

in an interval around those selected in step 1 (using the second subsample). We have used the R function "cv.glmnet" of the package `glmnet` (see Friedman et al. (2010)). This function fits a generalized linear model via maximum likelihood with LASSO penalization and the amount of the penalization is fixed by 10-fold cross validation. Although the default implementation of the function standardizes the variates, we have decided not to do it in this case. This version of LASSO selects automatically the number of variables, thus the only smoothing parameter to be selected here is the grid step for the points used in first stage of the method. This parameter is also set by 10-fold cross-validation.

4. *Maxima Hunting* (MH), proposed in Berrendero et al. (2016): the original method was used in the setting of variable selection in supervised classification, but there is no conceptual restriction to apply the same procedure in a regression setting. The basic idea is to select the local maxima of the "distance covariance" (association measure for random variables introduced in Székely et al. (2007)) between the response and the marginal variables of the process. In practice, the numerical identification of these maxima depends on a smoothing parameter h which is chosen by 10-fold cross-validation. The number of variables is also set by cross-validation. The code of this method was provided by the authors Berrendero et al. (2016).

5. *Partial Least Squares* (PLS). This technique is well-known among the functional data practitioners. The goal of PLS is not to pick up a few variables but to select some appropriate linear functionals of the original data (very much in the spirit of principal components analysis). So PLS is not a variable selection procedure, but a dimension reduction method. This means that PLS is not directly comparable to the variable selection methods considered here, since its aims are not exactly the same. When we choose to use a variable selection procedure, it is understood that we want to perform some kind of dimension reduction still keeping the interpretability of the information directly given in terms of the original variables. By contrast, PLS might perhaps provide some gains in efficiency but at the expense of doing a dimensionality reduction with a more difficult interpretation. Anyway, we have included PLS in our study as a useful reference for comparisons, aimed at checking how much do we lose by restricting ourselves to variable selection methods. We have used the function "fregre.pls.cv" of the `fda.usc` R-package to compute the predictions. The number of components in the model is chosen using the "Akaike information criterion" (AIC). Moreover, for the real data sets, which are smoother than the simulated ones, we have found that it is better to penalize the norm of the second derivative of the slope function. The amount of penalization in these cases is also fixed using AIC model selection criterion.

6. *Base.* We denote by "base" the prediction methodology derived from the standard L_2 linear regression model (3.1). No variable selection or dimension reduction procedure is done. Hence, this method is incorporated again just for the sake

of comparison, to assess the accuracy of the predictions based on some previous variable selection procedure with those provided by the standard functional regression model (3.1). For the real data sets and the integrated Brownian motion, which are quite smooth, we have used the function "fregre.basis.cv" of the `fda.usc` R-package (which relies on a spline basis representation of the trajectories) to compute the predictions. In this function the number of basis elements to retain is set by generalized cross-validation. As for PLS, we allow a penalization in the second derivative. However, for the remaining examples, we have found that it is better to use a data-derived basis, in order to capture the irregularities of the data. For these sets we have used the function "fregre.pc.cv" with no penalization, in which the number of components is chosen by the "Akaike information criterion".

For each model, all methods are checked for the sample size $n = 150$, which has been split on 100 observations used as training data and 50 employed as test data. As usual, the functional simulated data are discretized to $(x(t_1), \ldots, x(t_{100}))$, where t_i are equispaced points in $[0, 1]$, starting from $t_1 = 1/100$. The real data sets are already provided in a discretized fashion, so we use the corresponding grids. For all the experiments we obtain the Relative Mean Squared Error (RMSE) of each method, as defined by

$$\text{RMSE}\left(\widehat{Y}; Y\right) = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}{\sum_{i=1}^{n} y_i^2}. \tag{3.29}$$

Moreover, in those cases where β has a "sparse" form of type $\beta(t) = \sum_j \beta_j K(t_j, \cdot)$ we obtain two measures of the accuracy in the variable selection procedure. Namely, we calculate the Hausdorff distance between the set of estimated points and the set of the true ones T^* as well as the number of points selected $(\widehat{p})$ in order to compare it with the true number (p^*). We have imposed a maximum of 10 selected points to the methods, except to PVS, since our implementation of this method does not permit to decide the number of selected variables. The Hausdorff distance gives us an idea of the precision of the method since it takes into account the separation to the true points as well as the number of selected points. Each experiment has been replicated 100 times.

3.6.4 Numerical outputs

Tables 3.1, 3.2 and 3.3 provide the performance of the methods measured in terms of the Relative Mean Squared Error (Eq. (3.29)) of the predictions, for each of the three regression models tested and for the two real data sets. Methods are presented in columns. Models appear in rows. Table 3.4 contains the Hausdorff distance between the selected and the "true" relevant points for the four variable selection methods and the two models with sparse β function. For these same experiments, Table 3.5 provides

Regression model 1

	RKHS	KPS	PVS	MH	PLS	Base
Bm	**1.09** (0.394)	2.38 (1.94)	**1.2** (0.36)	5.2 (3.08)	4.62 (1.4)	5.9 (1.56)
gBm	**0.42** (0.292)	5.5 (5.53)	**0.55** (0.343)	6.55 (6.24)	3.71 (1.81)	5.02 (2.62)
iBM	**0.0166** (0.0162)	0.106 (0.162)	0.13 (0.0553)	48.5 (10.9)	0.139 (0.0523)	**0.0303** (0.00813)
OU	**1.07** (0.433)	2.25 (1.51)	**1.23** (0.462)	5.59 (3.57)	4.61 (1.63)	6 (2.07)
fBm 0.2	**0.422** (0.134)	1.76 (2.52)	**0.532** (0.185)	14.6 (5.27)	11.3 (3.78)	29.7 (7.83)
fBm 0.8	**2.91** (0.929)	3.54 (1.41)	**3.08** (0.99)	11.4 (7.04)	3.75 (1.03)	3.79 (1.04)

Regression model 2

	RKHS	KPS	PVS	MH	PLS	Base
Bm	**0.0976** (0.0481)	1.18 (0.595)	**0.171** (0.0779)	2.48 (1.04)	0.596 (0.158)	0.693 (0.201)
gBm	**0.0206** (0.0128)	1.03 (0.75)	**0.295** (0.329)	2.17 (1.61)	0.429 (0.164)	0.462 (0.236)
iBm	**0.000312** (0.000491)	0.00102 (0.000744)	0.102 (0.0368)	0.0525 (0.015)	0.00405 (0.00165)	**0.000221** (0.0000625)
OU	**0.0846** (0.0292)	1.19 (0.715)	**0.163** (0.0783)	2.28 (1.25)	0.54 (0.166)	0.613 (0.167)
fBm 0.2	**0.0831** (0.0215)	5.25 (4.01)	**0.17** (0.11)	4.7 (4.26)	3.14 (0.999)	7.16 (2.162)
fBm 0.8	**0.105** (0.0303)	0.168 (0.0592)	0.373 (0.183)	0.442 (0.132)	0.138 (0.0382)	**0.122** (0.0294)

Table 3.1: Mean and standard deviation of the RMSE for the response variable for simulated data sets with models 1 and 2 (scale of 10^{-2}).

the number of selected points. In all the tables of the simulated data appear the mean and the standard deviation of each of the measured quantities.

In view of these tables, for the simulated data the proposed method seems to outperform the other variable selection procedures, according to all the considered criteria metrics (RMSE, Hausdorff distance and number of points) whenever a sparse model of type (3.2) holds. The PVS method also performs quite well. When the model is not satisfied, and the response variable depends on the whole trajectory, PLS and the base method are the best in general, as expected. The Maxima Hunting method outperforms these two methods in some cases. The order of magnitude of the error for the remaining methods is the same as that of PLS in most cases. That is, using a few number of variables instead of the whole trajectory does not significantly affect the prediction error, even if the response depends on the whole trajectory.

The exact points selected with each of the methods for the real data sets are plotted in Figures 3.3 and 3.4. It seems that our method and Maxima Hunting are the ones that better manage the continuity of the data, in the sense that they do not choose close and highly correlated points. We can see also that there are some points in common among the ones selected by MH and our proposal. For the Tecator set, KPS obtains the best results, followed by our method. The results presented here for this data set might not coincide with those of previous works, since (as explained before) we have removed the repeated observations. For the sugar data, our proposal is slightly outperformed, but it is better than the estimators that use the whole trajectories. In addition, RKHS method uses only 2 points in this case, in contrast with KPS and PVS (10 points each of them, which is the fixed maximum).

On the other hand, we also provide a couple of results regarding execution time. We have measured the execution time of the six methods for the third regression model when the functional data are drawn from the Ornstein-Uhlenbeck process and the

Regression model 3

	RKHS	KPS	PVS	MH	PLS	Base
Bm	4.44 (1.29)	4.25 (1.08)	4.1 (1.1)	**3.88** (1.05)	4.05 (1.13)	**3.79** (0.969)
gBm	1.37 (1.25)	0.987 (0.23)	1.01 (0.343)	1.84 (0.603)	**0.887** (0.323)	**0.884** (0.324)
iBm	0.00379 (0.00118)	0.0033 (0.00109)	0.0127 (0.00477)	0.042 (0.0116)	**0.00299** (0.000997)	**0.00305** (0.00106)
OU	4.77 (1.2)	4.45 (1.13)	4.11 (0.982)	**3.97** (0.928)	4.28 (1.24)	**4.01** (0.998)
fBm 0.2	4.41 (1.01)	4.28 (0.992)	4.12 (1)	**4.09** (0.95)	4.49 (1.18)	**3.71** (0.835)
fBm 0.8	4.89 (1.15)	4.45 (0.988)	4.23 (1.1)	**4.03** (0.867)	**4.08** (0.897)	4.11 (0.921)

Table 3.2: Mean and standard deviation of the RMSE for the response variable for simulated data sets with model 3 (scale of 10^{-1}).

	RKHS	KPS	PVS	MH	PLS	Base
2^{nd} derivative of tecator	**0.032**	0.034	**0.030**	0.056	0.039	0.048
Ash content in sugar	0.321	**0.185**	**0.222**	0.246	0.401	0.465

Table 3.3: RMSE for the response variable for real data sets.

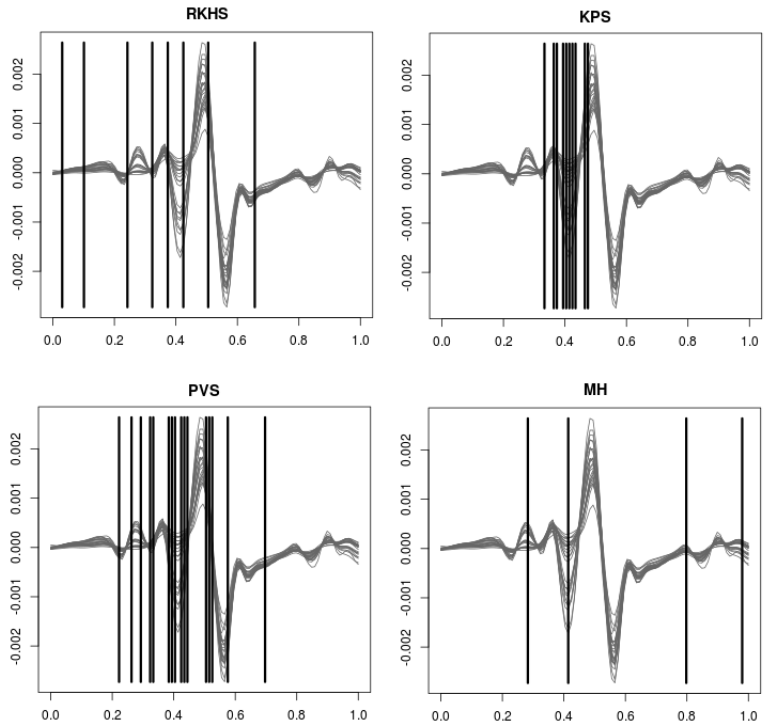


Figure 3.3: Original curves of Tecator data set with the selected points for each of the methods.

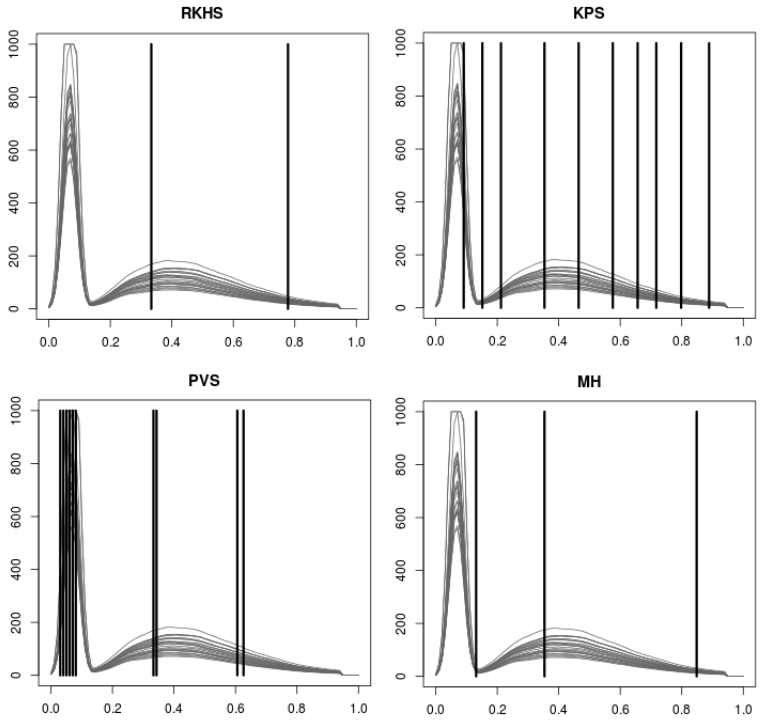


Figure 3.4: Original curves of the ash content in sugar set with the selected points for each of the methods.

Regression model 1

	RKHS	KPS	PVS	MH
Bm	**0.089** (0.141)	2.1 (0.374)	1.63 (0.648)	2.04 (0.431)
gBm	**0.135** (0.219)	2.17 (0.304)	1.71 (0.692)	2.13 (0.637)
iBm	**0.998** (0.0141)	2.14 (0.313)	1.05 (0.828)	6.01 (0.54)
OU	**0.139** (0.238)	2.03 (0.354)	1.64 (0.625)	2.03 (0.41)
fBm 0.2	**0** (0)	2.08 (0.362)	1.69 (0.592)	2.01 (0.621)
fBm 0.8	**0.264** (0.201)	2.09 (0.28)	1.6 (0.69)	2.9 (1.74)

Regression model 2

	RKHS	KPS	PVS	MH
Bm	1.29 (0.0731)	1.27 (0.181)	**0.994** (0.312)	1.51 (0.733)
gBm	1.25 (0.134)	1.43 (0.875)	**1.21** (0.869)	1.62 (0.929)
iBm	**0.775** (0.297)	1.26 (0.181)	7.25 (0.355)	6.53 (1.44)
OU	1.27 (0.127)	1.21 (0.238)	**1.03** (0.321)	1.44 (0.674)
fBm 0.2	1.3 (0)	1.27 (0.19)	**0.946** (0.419)	1.33 (0.272)
fBm 0.8	**1.18** (0.268)	1.2 (0.218)	4.29 (2.9)	2.77 (1.12)

Table 3.4: Mean and standard deviation of the Hausdorff distance to the actual relevant points (Scale of 10^{-1}).

Regression model 1 ($p^* = 3$)

	RKHS	KPS	PVS	MH
Bm	**3.14** (0.427)	8.3 (2.05)	11.1 (3.95)	6.42 (1.7)
gBm	**3.59** (0.753)	8.22 (2.01)	10.9 (3.61)	6.2 (1.75)
iBm	6.09 (0.473)	9.12 (1.35)	11.6 (4.32)	**1.17** (0.378)
OU	**3.23** (0.489)	8.34 (2.01)	10.9 (3.88)	5.91 (1.48)
fBm 0.2	**3** (0)	8.56 (1.78)	11.5 (3.79)	5.6 (2)
fBm 0.8	3.29 (0.478)	8.92 (1.73)	10.6 (4.21)	**2.92** (1.32)

Regression model 2 ($p^* = 5$)

	RKHS	KPS	PVS	MH
Bm	**4.88** (0.743)	9.61 (0.975)	8.75 (2.07)	6.15 (1.98)
gBm	**5.45** (1.09)	9.55 (0.869)	8.89 (2.74)	6.14 (2.2)
iBm	**5.68** (1.1)	9.28 (1.08)	3.03 (0.964)	1.3 (0.459)
OU	**5.05** (0.84)	9.69 (0.704)	8.11 (1.95)	5.77 (1.84)
fBm 0.2	**4** (0.086)	9.72 (0.727)	6.96 (1.88)	7.41 (1.92)
fBm 0.8	**5.07** (0.742)	9.29 (1.11)	6.96 (3.35)	3.17 (1.16)

Table 3.5: Mean and standard deviation of the number of selected points ($\widehat{p}$).

	RKHS	KPS	PVS	MH	PLS	Base
OU	**0.00566** (0.00136)	11.6 (1.11)	14.1 (1.95)	1.15 (0.0717)	**0.173** (0.0447)	0.342 (0.0372)
fBm 0.8	**0.00806** (0.00254)	8.39 (2.21)	35.9 (11.1)	1.65 (0.335)	**0.24** (0.0831)	0.49 (0.123)

Table 3.6: Mean and standard deviation of the execution time.

fractional Brownian motion with $H = 0.8$. The results can be seen in Table 3.6. As we have already mentioned, the RKHS-based method does not require any validation step to determine the number of selected variables. Therefore, the execution time is significantly smaller than that of the other variable selection methods. We can also see that the execution time for the PVS method is much bigger than the others. Note however that this method has in general a good behavior in terms of prediction error.

3.7 Extension to functional response

In the previous sections we have presented several results for an RKHS regression model with functional predictors and scalar response. In this section we will show how to extend both the model and most of the previous results to the problem with functional response. The two main difficulties with this extension are listed below.

- The methodology can not be directly extended by just adjusting the model for every $Y(s)$, where Y is the response process. With this procedure one would obtain different sets of points $t_1^s, \ldots, t_p^s$ for each $s \in [0,1]$. The optimal number of variables p also may not necessarily be the same for every s. Thus, we need to define an optimality criterion that can be directly optimized in Θ_p. We have decided to remove this dependence by integrating over s. However, different approaches could be analyzed. For instance, all the theory remains valid (with slight changes) if one takes the supremum over s of the residuals between the responses $Y(s)$ and the projections $Y(s)_{T_p}$. We choose the first approach since it provided better empirical results in the tested examples.

- The cross-covariance function $\mathrm{cov}(X(s), Y(t))$, $s, t \in [0,1]$, which was not needed previously in the chapter, plays an important role in this setting. The new optimality criteria depend on this function, so we need uniform almost sure convergence of the sample counterpart in order to extend the previous results on variable selection. This entails some restrictions on the slope function $\beta(s,t)$ that generates the responses. In particular, it is necessary that $\beta(s,\cdot) \in \mathcal{H}(K)$ for every $s \in [0,1]$ and that the stochastic process $U(s) = \Psi_X^{-1}(\beta(s,\cdot))$ satisfies $\mathbb{E}[\sup_s U(s)^2] < \infty$. This latter condition is true for $\beta(s,\cdot)$ in $\mathcal{H}_0(K)$, but it is not clear when it is satisfied for a general function in the RKHS. When this condition is not fulfilled, we can not assure consistent estimation of the selected variables.

3.7.1 Extended regression model

Let us start defining the regression model with functional response. For a recent overview on functional regression see Morris (2015) (the functional-functional problem can be found in Section 6). In the literature, the classical functional regression model with functional response, for $m = \mathbb{E}[X]$, is

$$Y(s) \;=\; \int_0^1 \beta(s,t)\big(X(t) - m(t)\big)\mathrm{d}t + \varepsilon(s) \;=\; \langle \beta(s,\cdot), X - m\rangle_2 + \varepsilon(s),\;\; s \in [0,1],$$

which is the direct extension of (3.1). This model is analyzed, among others, by Cai et al. (2006) and Kokoszka et al. (2008), or Yao et al. (2005), where FDA techniques are applied to longitudinal data. Müller and Yao (2008) generalize it to an additive structure and Müller (2005) gives a review of some FDA techniques, including functional regression with functional response. For a non-parametric approach to this problem see Ferraty et al. (2012).

As we made in Section 3.2.1 when defining the model of Equation (3.3), we suggest to replace the L^2 product with the inner product of the RKHS associated with X. That is, we define for every $s \in [0,1]$,

$$Y(s) \;=\; \langle \beta(s,\cdot), X\rangle_K + \varepsilon(s) \;=\; \Psi_X^{-1}(\beta(s,\cdot)) + \varepsilon(s), \tag{3.30}$$

where $\beta(s,\cdot) \in \mathcal{H}(K)$ and ε is a zero mean L^2 stochastic process, point-wisely uncorrelated with X. This model is the extension to functional response of model (3.3) studied in the previous sections. Using Loève's isometry, the slope function β can be rewritten as

$$\beta(s,\cdot) \;=\; \Psi_X\left(Y(s) - \varepsilon(s)\right) \;=\; \mathbb{E}\left[\big(Y(s) - \varepsilon(s)\big)\big(X(\cdot) - m(\cdot)\big)\right] \;=\; c(s,\cdot), \tag{3.31}$$

where $c(s,t)$ stands in this section for $\mathrm{cov}\big(Y(s), X(t)\big)$.

The natural question now is how to extend the sparse representation (3.13) of the slope function β, in order to perform variable selection in this setting. Since in the previous model we assume that $\beta(s,\cdot) \in \mathcal{H}(K)$ for every $s \in [0,1]$, the most direct extension is to replace the coefficients $\beta_j \in \mathbb{R}$ in Equation (3.13) with continuous functions $\beta_j(\cdot) \in C[0,1]$. That is, the slope function β in (3.30) is defined by a single set of points $T^* = (t_1^*, \ldots, t_p^*)$ as

$$\beta(\star,\cdot) = \sum_{j=1}^{p} \beta_j(\star) K(t_j^*, \cdot), \tag{3.32}$$

where β_j are continuous functions. It means that all the evaluations $\beta(s,\cdot)$ depend on the same set of points T^*. This allows us to write the whole response Y in terms of the same set of variables $\{\widetilde{X}(t_1^*), \ldots, \widetilde{X}(t_p^*)\}$, where $\widetilde{X} = X - \mathbb{E}[X]$. It could seem a rather strict assumption but the approximation error of the whole function can be reduced

by increasing the number of points p. In fact, in the practical examples presented in Section 3.8.3 for time series forecasting we see that a small number of points is typically enough (usually less than ten). Besides, at the end of this current section we present some simulations to quantify how good this finite-dimensional approximation is when increasing the number of elements p in this sum.

Whenever the slope function follows this sparse representation, the response Y is, pointwise for each $s \in [0,1]$,

$$Y(s) \;=\; \Psi_X^{-1}\Big(\sum_{j=1}^{p} \beta_j(s) K(t_j^*, \cdot)\Big) + \varepsilon(s) \;=\; \sum_{j=1}^{p} \beta_j(s)\big(X(t_j^*) - m(t_j^*)\big) + \varepsilon(s). \quad (3.33)$$

It is clear that if the functions $\beta_j(\cdot)$ and the trajectories of the error process $\varepsilon(\cdot)$ are continuous, then the realizations of the response Y are also continuous. Besides, in this case the covariance function $\operatorname{cov}\big(Y(s), X(t)\big)$ is continuous in $[0,1]^2$.

Optimality criteria

In order to select the relevant points $\{t_1^*, \ldots, t_p^*\}$ we translate the optimality criteria Q_0, Q_1 and Q_2 (Equations (3.10), (3.4) and (3.8)) to this context, which would be proven also to be equivalent (in the sense that the set of optimal points T_p^* is the same). Let us start with Q_1, as we did in the previous sections. Since each evaluation $Y(s)$ is a random variable in $L^2(\Omega)$, we minimize pointwisely the distances

$$q_1(T_p \,;\, \alpha_1, \ldots, \alpha_p)(s) \;=\; \Big\| Y(s) - \sum_{j=1}^{p} \alpha_j(s)\big(X(t_j) - m(t_j)\big) \Big\|^2,$$

where the coefficients $\alpha_j(s)$ (which are just real numbers) depend on the points $t_1, \ldots, t_p$. We have to find the functions $\alpha_j(\cdot)$ such that $\alpha_j(s)$, $s \in [0,1]$, give the best approximation of $Y(s)$ for a given set of points T_p. Since we are assuming that the relevant points $(t_1^*, \ldots, t_p^*)$ are the same for every s, one can not merely minimize this function for every s. Then, integrating q_1 over s leads to

$$Q_1(T_p) \;:=\; \int_0^1 \min_{\alpha_j(s) \in \mathbb{R}} q_1(T_p \,;\, \alpha_1, \ldots, \alpha_p)(s)\, \mathrm{d}s, \quad (3.34)$$

which can now be minimized with respect to T_p in Θ_p (Equation (3.5)) and ensures that the optimal α_j are continuous, as we will see. Different extensions are suitable. For instance, one could take the supremum over $[0,1]$ instead of the integral. From Proposition 3.1 we know that $q_1(s)$ is a convex function in $\alpha_j(s)$ for each $s \in [0,1]$. Thus, we can obtain an explicit expression of the minimizing functions, denoted by $\alpha_j^*(s)$, pointwise for each $s \in [0,1]$. We will see in the proof of Proposition 3.13 that these optimal functions are given by

$$(\alpha_1^*(s), \ldots, \alpha_p^*(s)) \;=\; \Sigma_{T_p}^{-1}\, c(s, T_p),$$

where now $c(s,T_p) = \big(\mathrm{cov}\big(Y(s),X(t_1)\big),\ldots,\mathrm{cov}\big(Y(s),X(t_p)\big)\big)'$ is the vector of evaluations of the cross-covariance function.

As before, we can also formulate the optimization problem using the norm of the RKHS using

$$q_2(T_p;\alpha_1,\ldots,\alpha_p)(s) \;=\; \Big\|\beta(s,\cdot)-\sum_{j=1}^{p}\alpha_j(s)K(t_j,\cdot)\Big\|_K^2$$

for each $s\in[0,1]$. Then we define

$$Q_2(T_p) := \int_0^1 \min_{\alpha_j(s)\in\mathbb{R}} q_2(T_p;\alpha_1,\ldots,\alpha_p)(s)\mathrm{d}s. \tag{3.35}$$

For the sake of clarity, we use the following notation: given two sets of real valued functions $\{f_i\}$ and $\{g_i\}$ we write, using vector notation,

$$\sum_{i=1}^{N}(f_ig_i)(\cdot) \;=\; \sum_{i=1}^{N} f_i(\cdot)g_i(\cdot) \;=\; (f_1(\cdot),\ldots,f_N(\cdot))'(g_1(\cdot),\ldots,g_N(\cdot)).$$

In the following result we extend Proposition 3.1 to the new definitions of Q_1 and Q_2 and redefine the function Q_0, which has an easily computable expression.

Proposition 3.13. ***(Extension of Prop. 3.1)*** *Let Y and X be two stochastic processes with continuous trajectories fulfilling the regression model* (3.30) *with $\mathbb{E}\|\varepsilon\|_2^2<\infty$. Then,*

$$\arg\min_{T_p\in\Theta_p} Q_1(T_p) \;=\; \arg\min_{T_p\in\Theta_p} Q_2(T_p) \;=\; \arg\max_{T_p\in\Theta_p} Q_0(T_p),$$

where

$$Q_0(T_p) := \int_0^1 c(s,T_p)'\Sigma_{T_p}^{-1}c(s,T_p)\mathrm{d}s$$

and $c(s,T_p) = \big(\mathrm{cov}\big(Y(s),X(t_1)\big),\ldots,\mathrm{cov}\big(Y(s),X(t_p)\big)\big)'$.

Proof. Since $\mathbb{E}[\varepsilon(s)]=0$, $\mathbb{E}\big[\varepsilon(s)\big(X(t)-m(t)\big)\big]=0$ and $\langle X,\beta(s,\cdot)\rangle_K\in\mathcal{L}(X)$ for every $s,t\in[0,1]$, using the definition of Loève's isometry we have that

$$\Big\|Y(s)-\sum_{j=1}^{p}\alpha_j(s)\big(X(t_j)-m(t_j)\big)\Big\|^2 \;=\; \Big\|\beta(s,\cdot)-\sum_{j=1}^{p}\alpha_j(s)K(t_j,\cdot)\Big\|_K^2+\sigma^2(s),$$

where $\sigma^2(s)=\mathrm{var}(\varepsilon(s))<\infty$, so the minimizing values $\alpha_j(s)$ are the same for both sides of the equality. Then integrating over s, $Q_1(T_p)=Q_2(T_p)+\|\sigma^2\|_2=Q_2(T_p)+\mathbb{E}\|\varepsilon\|_2^2$ (by Fubini's theorem) and we get the first equality of the statement.

Again pointwisely for each $s \in [0,1]$, using the reproducing property of $\mathcal{H}(K)$,

$$\left\|\beta(s,\cdot)-\sum_{j=1}^{p}\alpha_j(s)K(t_j,\cdot)\right\|_K^2 = \|\beta(s,\cdot)\|_K^2+\sum_{i,j=1}^{p}\alpha_i(s)\alpha_j(s)K(t_i,t_j)-2\sum_{j=1}^{p}\alpha_j(s)\beta(s,t_j).$$

Since K is positive-definite, this latter function is convex in $\alpha_j(s)$ for each $s \in [0,1]$. Therefore we can compute its minimum pointwisely, which is achieved at $(\alpha_1^*(\cdot),\ldots,\alpha_p^*(\cdot))' = \Sigma_{T_p}^{-1}c(\cdot,T_p)$, since $c(s,t) = \beta(s,t)$ for each s (Equation (3.31)). Then if we substitute this optimum in the previous equation we get

$$\min_{\alpha_j(s)\in\mathbb{R}} \left\|\beta(s,\cdot)-\sum_{j=1}^{p}\alpha_j(s)K(t_j,\cdot)\right\|_K^2 \;=\; \|\beta(s,\cdot)\|_K^2 - c(s,T_p)'\Sigma_{T_p}^{-1}c(s,T_p).$$

Hence, integrating over $s \in [0,1]$,

$$Q_1(T_p) = \int_0^1 \sigma^2(s)\mathrm{d}s + \int_0^1 \|\beta(s,\cdot)\|_K^2\mathrm{d}s - Q_0(T_p) = C - Q_0(T_p).$$

This constant C is finite since the integral of σ^2 is equal to $\mathbb{E}\|\varepsilon\|_2^2 < \infty$ and the integral $\int_0^1 \|\beta(s,\cdot)\|_K^2\mathrm{d}s$ is bounded by $\sup_{s\in[0,1]}\|\beta(s,\cdot)\|_K^2$, where $\|\beta(s,\cdot)\|_K^2$ is a continuous function on $[0,1]$ (it is the composition of two continuous functions, $s \mapsto \beta(s,\cdot) = \mathrm{cov}(Y(s),X(\cdot))$ and $f \mapsto \|f\|_K$). □

3.7.2 Sample estimation

As mentioned, the optimality criterion defined by Q_0 is simple to implement in practice. In this section we study the asymptotic properties of the natural estimator of Q_0. These results are useful when studying asymptotic results on the selected points and the estimated trajectories. We work with a sample $(x_1,y_1),\ldots,(x_n,y_n)$ of size n. Then, for a given number of points p, the natural estimator for the function $Q_0(T_p)$ is

$$\widehat{Q}_0(T_p) = \int_0^1 \widehat{c}'(s,T_p)\;\widehat{\Sigma}_{T_p}^{-1}\;\widehat{c}(s,T_p)\mathrm{d}s, \tag{3.36}$$

where $\widehat{c}(\cdot,T_p) = (\widehat{c}(\cdot,t_1),\ldots,\widehat{c}(\cdot,t_p))'$ and $\widehat{c}$ is the usual estimator of the covariance function based on the sample means $\bar{x},\bar{y}$,

$$\widehat{c}(s,t) \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i(s)x_i(t) - \bar{y}(s)\bar{x}(t).$$

The entries of the sample covariance matrix, $\widehat{K}(t_i,t_j)$, $t_i,t_j \in T_p$, are computed equivalently. According to this criterion, we propose to select as the most relevant points as

$$\widehat{T}_p \;=\; \arg\max_{T_p\in\Theta_p} \widehat{Q}_0(T_p). \tag{3.37}$$

Further down we prove some consistence results for this estimator, under the assumption that the finite dimensional model defined by Equation (3.33) holds. First we prove an additional result which was not needed for scalar response.

Lemma 3.14. *Let model* (3.30) *hold with* $\mathbb{E}[\sup_s |X(s)|^2]$, $\mathbb{E}[\sup_s |\varepsilon(s)|^2] < \infty$ *for* $s \in [0,1]$ *and the slope function* β *such that the corresponding stochastic process* $U(s) = \Psi_X^{-1}\big(\beta(s,\cdot)\big)$ *fulfills* $\mathbb{E}[\sup_s |U(s)|^2] < \infty$*, then*

$$\sup_{t,s\in[0,1]} |\widehat{c}(s,t) - c(s,t)| \overset{\text{a.s.}}{\to} 0. \tag{3.38}$$

Proof. The first addend in the definition of $\widehat{c}(s,t)$ is the sample mean of the product process $Z(s,t) = Y(s)X(t)$, whose trajectories fall in the separable Banach space $C[0,1]^2$. Then, whenever the strong expectation of Z exists, the sample mean converges uniformly a.s. to $\mathbb{E}[Z]$ by Mourier's SLLN. The requirement is equivalent to $\mathbb{E}\|Z\|_\infty < \infty$. As in the proof of Lemma 3.5,

$$\begin{aligned}
\sup_{(s,t)\in[0,1]^2} \big|Y(s)X(t)\big| &\leq \frac{1}{2} \sup_{s\in[0,1]} \big|Y(s)\big|^2 + \frac{1}{2} \sup_{s\in[0,1]} \big|X(s)\big|^2 \\
&\leq \sup_{s\in[0,1]} \big|\Psi_X^{-1}\big(\beta(s,\cdot)\big)\big|^2 + \sup_{s\in[0,1]} |\varepsilon(s)|^2 + \frac{1}{2} \sup_{s\in[0,1]} \big|X(s)\big|^2,
\end{aligned}$$

and the expectations of these supremums are finite by hypothesis. In the proof of Lemma 3.5 we proved that the sample expectation $\bar{x}$ converges uniformly a.s. to $\mathbb{E}[X]$. For the convergence of $\bar{y}$ we can apply again Mourier's SLLN since

$$\begin{aligned}
\mathbb{E}\Big[\sup_{s\in[0,1]} \big|Y(s)\big|\Big] &\leq \mathbb{E}\Big[\sup_{s\in[0,1]} \big|\Psi_X^{-1}\big(\beta(s,\cdot)\big)\big|\Big] + \mathbb{E}\Big[\sup_{s\in[0,1]} \big|\varepsilon(s)\big|\Big] \\
&\leq \Big(\mathbb{E}\Big[\sup_{s\in[0,1]} \big|\Psi_X^{-1}\big(\beta(s,\cdot)\big)\big|^2\Big]\Big)^{\frac{1}{2}} + \Big(\mathbb{E}\Big[\sup_{s\in[0,1]} \big|\varepsilon(s)\big|^2\Big]\Big)^{\frac{1}{2}} < \infty.
\end{aligned}$$

□

The condition in the statement over the slope function is fulfilled whenever the sparse representation (3.32) holds. But it may not be the case for limits of these finite linear combinations. We next extend the results about the continuity and convergence of the optimality functions $\widehat{Q}_0$ and Q_0.

Lemma 3.15. ***(Extension of Lemmas 3.6 and 3.7)*** *Let* X *be a stochastic process in* $C[0,1]$ *and* $p \geq 1$ *be such that the covariance matrices* Σ_{T_p} *are invertible for all* $T_p \in \Theta_p$*. Then* Q_0 *and* $\widehat{Q}_0$ *are continuous on* Θ_p *and* $\sup_{T_p\in\Theta_p} |\widehat{Q}_0(T_p) - Q_0(T_p)| \overset{\text{a.s.}}{\to} 0$.

Proof. Rewriting the functions as

$$Q_0(T_p) \;=\; \int_0^1 c(s,T_p)'\Sigma_{T_p}^{-1}c(s,T_p)\;\mathrm{d}s \;=\; \int_0^1 q_0(T_p;s)\mathrm{d}s,$$

$$\widehat{Q}_0(T_p) \;=\; \int_0^1 \widehat{c}(s,T_p)'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}(s,T_p)\;\mathrm{d}s \;=\; \int_0^1 \widehat{q}_0(T_p;s)\mathrm{d}s,$$

we can use the pointwise properties of the integrands of Q_0 and $\widehat{Q}_0$ previously showed.

Regarding the continuity of the functions, $q_0(T_p;s)$ and $\widehat{q}_0(T_p;s)$ are continuous in T_p for each $s \in [0,1]$ by Lemma 3.6. In particular, they are uniformly continuous, which implies the equicontinuity of the family $\{q_0(s,\cdot)\}_{s\in[0,1]}$. Then, for every $\epsilon > 0$, if $\|T_p - S_p\|_2 < \delta$, where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^p$,

$$|Q_0(T_p)-Q_0(S_p)| \;=\; \Big|\int_0^1 \big(q_0(T_p;s)-q_0(S_p;s)\big)\mathrm{d}s\Big| \;\leq\; \int_0^1 \big|q_0(T_p;s)-q_0(S_p;s)\big|\mathrm{d}s \;<\; \epsilon.$$

And equivalently to see that $\widehat{Q}_0$ is continuous with probability one.

The uniform convergence is straightforward in view of the definition of Q_0. By Lemma 3.14, the vector $\widehat{c}(s,T_p)$ converge uniformly a.s. to $c(s,T_p)$, and in the proof of Lemma 3.7 we proved the uniform convergence of the inverse covariance matrices. Then, we get

$$\sup_{(s,T_p)\in[0,1]\times\Theta_p} \big|\widehat{c}(s,T_p)'\widehat{\Sigma}_{T_p}^{-1}\widehat{c}(s,T_p) - c(s,T_p)'\Sigma_{T_p}^{-1}c(s,T_p)\big| \;\overset{\text{a.s.}}{\to}\; 0,$$

which implies the convergence of the integral over $s \in [0,1]$. □

If we now assume that the curves are drawn from the finite dimensional model of Equation (3.33), additional asymptotic results for the estimator $\widehat{T}_p$ and for the estimated curves can be derived. It is clear from the expression of Q_1 (Equation (3.34)) that, whenever the slope function β has an sparse representation as in Equation (3.32), the set T^* is a global minimum of Q_1 on Θ_{p^*}, and therefore a global maximum of Q_0. Assuming that p^* is known, we can prove that this optimum is unique and the estimated points $\widehat{T}_{p^*}$ converge to the true ones T^*.

Theorem 3.16. ***(Extension of Th. 3.8)*** *Let X be a stochastic process with continuous mean and covariance functions. Under the assumptions of Lemma 3.15 for $p = p^*$, whenever* (3.32) *holds and covariance matrices $\Sigma_{T_{p^*}\cup S_{p^*}}$ are invertible for all $T_{p^*}, S_{p^*} \in \Theta_{p^*}$ with $T_{p^*} \neq S_{p^*}$, then:*

(a) The vector $T^* \in \Theta_{p^*}$ is the only global maximum of Q_0 on this space.

(b) $\widehat{T}_{p^*} \overset{a.s.}{\to} T^*$ with the sample size $n \to \infty$, where $\widehat{T}_{p^*}$ is given in Eq.(3.37) *with $p = p^*$.*

(c) $\widehat{T}_{p^*}$ converges to T^* in quadratic mean when $n \to \infty$.

Proof. (a) As before, U_{T_p} denotes the orthogonal projection of the random variable U in the closed subspace $\text{span}\{X(t_j) - m(t_j),\, t_j \in T_p\}$ of $L^2(\Omega)$. Because of the equivalence of the criteria proved in Proposition 3.13, it is enough to see that T^* is the only global minimum of $Q_1(T_{p^*})$ in Θ_{p^*}. From Equation (3.34) it is clear that T^* minimizes Q_1 since

$$Q_1(T_{p^*}) \;=\; \int_0^1 \|Y(s) - Y(s)_{T_{p^*}}\|^2 \,\mathrm{d}s \;=\; \int_0^1 \|Y(s)_{T^*} - Y(s)_{T_{p^*}}\|^2 \,\mathrm{d}s + \|\text{var}(\varepsilon)\|_2,$$

where $\|\cdot\|_2$ is the norm in $L^2[0,1]$, and therefore its minimum value is $\|\text{var}(\varepsilon)\|_2$. If there exists another vector $S^* \neq T^*$ which also achieves this value, one must have $\|Y(s)_{T^*} - Y(s)_{S^*}\|^2 = 0$ for almost every $s \in [0,1]$ (except on a set of measure zero with regard to the Lebesgue measure). However, it is enough to have one point s_0 in which this equality holds. For this point we have that $Y(s_0)_{T^*} = Y(s_0)_{S^*}$ a.s., which contradicts the assumption that the covariance matrix $\Sigma_{T^* \cup S^*}$ is invertible, and then $T^* = S^*$.

(b) The result follows from the fact that $\widehat{Q}_0$ and Q_0 are continuous functions such that $\widehat{Q}_0$ tends uniformly a.s. to Q_0 (Lemma 3.15) and Q_0 has a unique maximum in Θ_{p^*} (part (a)).

(c) The same argument as in part (c) of the proof of Theorem 3.8. □

Once we have selected the most relevant points from the sample, we want to estimate the trajectories of the response process. That is, we want to approximate

$$Y(\cdot)_{T^*} \;=\; \alpha_1(\cdot)\big(X(t_1^*) - m(t_1^*)\big) + \ldots + \alpha_{p^*}(\cdot)\big(X(t_{p^*}^*) - m(t_{p^*}^*)\big). \tag{3.39}$$

In the proof of Proposition 3.13 we have seen that the functions $(\alpha_1(\cdot), \ldots, \alpha_{p^*}(\cdot))'$ used to carry out this projection are given by $\Sigma_{T_{p^*}}^{-1}(c(\cdot, t_1^*), \ldots, c(\cdot, t_{p^*}^*))$. Therefore, we can construct the estimated curves as $\widehat{\alpha}_1(\cdot)(X(\widehat{t}_1) - \bar{x}(\widehat{t}_1)) + \ldots + \widehat{\alpha}_{p^*}(\cdot)(X(\widehat{t}_{p^*}) - \bar{x}(\widehat{t}_{p^*}))$, where now the functions $(\widehat{\alpha}_1(\cdot), \ldots, \widehat{\alpha}_{p^*}(\cdot))'$ are computed using the sample version of the covariances as $\widehat{\Sigma}_{\widehat{T}_{p^*}}^{-1}(\widehat{c}(\cdot, \widehat{t}_1), \ldots, \widehat{c}(\cdot, \widehat{t}_{p^*}))$. Thus our proposed estimator for $Y(\cdot)_{T^*}$ is

$$\widehat{Y}(\cdot)_{\widehat{T}_{p^*}} \;=\; \widehat{c}(\cdot, \widehat{T}_{p^*})'\widehat{\Sigma}_{\widehat{T}_{p^*}}^{-1}\big(X(\widehat{T}_{p^*}) - \bar{x}(\widehat{T}_{p^*})\big). \tag{3.40}$$

Under the same conditions of the previous theorem, we can see that this estimator converges to $Y(\cdot)_{T^*}$ uniformly a.s. and in quadratic mean.

Theorem 3.17. ***(Extension of Th. 3.9)*** *Under the same assumptions of Theorem 3.16 and when the sample size n increases,*

(a) $\widehat{Y}(\cdot)_{\widehat{T}_{p^*}}$ converges to $Y(\cdot)_{T^*}$ a.s. in $C[0,1]$, that is, $\sup_s |\widehat{Y}(s)_{\widehat{T}_{p^*}} - Y(s)_{T^*}| \overset{\text{a.s.}}{\to} 0$.

(b) If, in addition, there exists $\eta > 1$ such that $\mathbb{E}[\sup_s |X(s)|^{2\eta}] < \infty$, then it also holds $\mathbb{E}[(\sup_s |\widehat{Y}(s)_{\widehat{T}_{p^*}} - Y(s)_{T^*}|)^2] \to 0$.

Proof. (a) We derive a proof analogous to the one of Theorem 3.9. For simplicity, let us denote $\widehat{T}_{p^*} = \widehat{T}$ and $\widetilde{X} = X - m$. We can bound the norm in the statement as

$$
\begin{aligned}
\|\widehat{Y}(\cdot)_{\widehat{T}} - Y(\cdot)_{T^*}\|_\infty =& \big\|\widehat{c}(\cdot,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\big(X(\widehat{T}) - \bar{x}(\widehat{T})\big) - c(\cdot,T^*)'\Sigma_{T^*}^{-1}\big(X(T^*) - m(T^*)\big)\big\|_\infty \\
\leq& \big\|\widehat{c}(\cdot,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c(\cdot,T^*)'\Sigma_{T^*}^{-1}\widetilde{X}(T^*)\|_\infty \\
&+ \big\|\widehat{c}(\cdot,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\big(\bar{x}(\widehat{T}) - m(\widehat{T})\big)\big\|_\infty \\
\leq& \big\|\widehat{c}(\cdot,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c(\cdot,\widehat{T})'\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T})\big\|_\infty \\
&+ \big\|c(\cdot,\widehat{T})'\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c(\cdot,T^*)'\Sigma_{T^*}^{-1}\widetilde{X}(T^*)\big\|_\infty \\
&+ \big\|\widehat{c}(\cdot,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\big(\bar{x}(\widehat{T}) - m(\widehat{T})\big)\big\|_\infty \\
\leq& \sup_{s\in[0,1]} \big\|\widehat{c}(s,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1} - c(s,\widehat{T})'\Sigma_{\widehat{T}}^{-1}\big\|_2 \, \|\widetilde{X}(\widehat{T})\|_2 \\
&+ \big\|c(\cdot,\widehat{T})'\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c(\cdot,T^*)'\Sigma_{T^*}^{-1}\widetilde{X}(T^*)\big\|_\infty \\
&+ \sup_{s\in[0,1]} \|\widehat{c}(s,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\|_2 \, \big\|\bar{x}(\widehat{T}) - m(\widehat{T})\big\|_2 \\
\leq& \sup_{(s,T)\in[0,1]\times\Theta_{p^*}} \big\|\widehat{c}(s,T)'\widehat{\Sigma}_T^{-1} - c(s,T)'\Sigma_T^{-1}\big\|_2 \sup_{T\in\Theta_{p^*}} \|\widetilde{X}(T)\|_2 \quad (3.41) \\
&+ \sup_{s\in[0,1]} \big|c(s,\widehat{T})'\Sigma_{\widehat{T}}^{-1}\widetilde{X}(\widehat{T}) - c(s,T^*)'\Sigma_{T^*}^{-1}\widetilde{X}(T^*)\big| \quad (3.42) \\
&+ \sup_{(s,T)\in[0,1]\times\Theta_{p^*}} \|\widehat{c}(s,T)'\widehat{\Sigma}_T^{-1}\|_2 \sup_{T\in\Theta_{p^*}} \big\|\bar{x}(T) - m(T)\big\|_2, \quad (3.43)
\end{aligned}
$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^{p^*}$. The first terms in Equation (3.41) goes a.s. to zero due to Lemma 3.14 (for the cross-covariance vector) and the argument in the proof of Lemma 3.7 (for the inverse of the auto-covariance matrix). Besides, the second term is a.s. finite due to the continuity of both the trajectories of X and the mean function. For the term in Equation (3.42), the convergence follows from the continuity of c, Σ, $\widetilde{X}$ and the norm in $C[0,1]$ together with Theorem 3.16(b). Finally, the first term in Equation (3.43) is bounded due to the a.s. uniform convergence of the covariances and the second term converges a.s. to zero by Equation (3.20).

(b) As in the previous part, we denote $\widehat{T}_{p^*} = \widehat{T}$. The statement is equivalent to see that the real valued random variables $\sup_s(|\widehat{Y}(s)_{\widehat{T}} - Y(s)_{T^*}|)^2$, indexed by the sample size n and denoted as Z_n, converge to zero in $L^1(\Omega)$ (the space of random variables such that $\mathbb{E}|Z| < \infty$). From part (a) we know that they converge a.s. to zero, so it only remains to check that the sequence Z_n is uniformly integrable (Vitali's convergence theorem). In order to do this, it is enough to check that $\sup_n \mathbb{E}[Z_n^\eta] < \infty$ for some $\eta > 1$ (de la

Vallée-Poussin's theorem). Since the function $f(y) = y^{2\eta}$ is convex in this case,

$$\begin{aligned} Z_n^\eta &\leq \Big(\sup_{s\in[0,1]} \big|\widehat{Y}(s)_{\widehat{T}}\big| + \sup_{s\in[0,1]} \big|Y(s)_{T^*}\big| \Big)^{2\eta} \\ &\leq 0.5\Big[\Big(2 \sup_{s\in[0,1]} \big|\widehat{Y}(s)_{\widehat{T}}\big| \Big)^{2\eta} + \Big(2 \sup_{s\in[0,1]} \big|Y(s)_{T^*}\big| \Big)^{2\eta} \Big]. \end{aligned}$$

Thus, we have to verify that

$$2^{2\eta-1}\Big(\sup_n \mathbb{E}\Big[\Big(\sup_{s\in[0,1]} \big|\widehat{Y}(s)_{\widehat{T}}\big| \Big)^{2\eta} \Big] + \mathbb{E}\Big[\Big(\sup_{s\in[0,1]} \big|Y(s)_{T^*}\big| \Big)^{2\eta} \Big] \Big) < \infty. \tag{3.44}$$

Let us start with the first addend of this equation. The supremum of $\|\widehat{c}(s,T)'\widehat{\Sigma}_T^{-1} - c(s,T)'\Sigma_T^{-1}\|_2$ for $(s,T) \in [0,1] \times \Theta_{p^*}$ goes a.s to zero by Lemma 3.14. Then we can bound the supremum as

$$\begin{aligned} \Big(\sup_{s\in[0,1]} \big|\widehat{Y}(s)_{\widehat{T}}\big| \Big)^{2\eta} &= \Big(\sup_{s\in[0,1]} \big|\widehat{c}(s,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\big(X(\widehat{T}) - \bar{x}(\widehat{T})\big)\big| \Big)^{2\eta} \\ &\leq \big\|X(\widehat{T}) - \bar{x}(\widehat{T})\big\|_2^{2\eta} \Big(\sup_{s\in[0,1]} \big\|\widehat{c}(s,\widehat{T})'\widehat{\Sigma}_{\widehat{T}}^{-1}\big\|_2 \Big)^{2\eta} \\ &\leq \big\|X(\widehat{T}) - \bar{x}(\widehat{T})\big\|_2^{2\eta} \Big(\sup_{s\in[0,1], T\in\Theta_{p^*}} \big\|c(s,T)'\Sigma_T^{-1}\big\|_2 + \epsilon \Big)^{\eta} \\ &\leq C\big\|X(\widehat{T}) - \bar{x}(\widehat{T})\big\|_2^{2\eta}, \end{aligned}$$

where $C < \infty$, since the function involved in the supremum inside brackets is a continuous function on the compact space $[0,1] \times \Theta_{p^*}$. To conclude we can use the same reasoning as in the proof of Theorem 3.9.

For the second addend in (3.44), which does not depend on the sample size,

$$\begin{aligned} \mathbb{E}\Big[\Big(\sup_{s\in[0,1]} \big|Y(s)_{T^*}\big| \Big)^{2\eta} \Big] &= \mathbb{E}\Big[\Big(\sup_{s\in[0,1]} \big|c(s,T^*)'\,\Sigma_{T^*}^{-1}\big(X(T^*) - m(T^*)\big)\big| \Big)^{2\eta} \Big] \\ &\leq \Big(\sup_{s\in[0,1]} \big\|c(s,T^*)'\,\Sigma_{T^*}^{-1}\big\|_2 \Big)^{2\eta}\ \mathbb{E}\Big[\big\|X(T^*) - m(T^*)\big\|_2^{2\eta} \Big], \end{aligned}$$

where the expectation is finite by hypothesis. □

3.7.3 Number of relevant points

We are left with deriving an estimator for p^*, the number of points to select. As we did before, we can rewrite the new function Q_0 in an iterative way as

$$Q_0(T_{p+1}) = \int_0^1 c(s,T_{p+1})'\Sigma_{T_{p+1}}^{-1} c(s,T_{p+1})\ \mathrm{d}s$$

$$= \int_0^1 c(s,T_p)'\Sigma_{T_p}^{-1}c(s,T_p)\ \mathrm{d}s + \int_0^1 \frac{\mathrm{cov}\big(Y(s)-Y(s)_{T_p},X(t_{p+1})\big)^2}{\mathrm{var}\big(X(t_{p+1}),X(t_{p+1})_{T_p}\big)}\ \mathrm{d}s$$

$$= Q_0(T_p) + \int_0^1 \frac{\mathrm{cov}\big(Y(s)-Y(s)_{T_p},X(t_{p+1})\big)^2}{\mathrm{var}\big(X(t_{p+1}),X(t_{p+1})_{T_p}\big)}\ \mathrm{d}s. \tag{3.45}$$

This derivation can also be done using the sample counterpart of Q_0,

$$\widehat{Q}_0(T_{p+1}) \ = \ \widehat{Q}_0(T_p) + \frac{\int_0^1 \widehat{\mathrm{cov}}\big(Y(s)-Y(s)_{T_p},X(t_{p+1})\big)^2\ \mathrm{d}s}{\widehat{\mathrm{var}}\big(X(t_{p+1}),X(t_{p+1})_{T_p}\big)}.$$

In order to compute this quotient in practice we can use a similar reasoning as in the proof of Proposition 3.2 and rewrite the previous equation as,

$$\widehat{Q}_0(T_{p+1}) \ = \ \widehat{Q}_0(T_p) \ + \ \frac{\int_0^1 (\widehat{c}(s,T_p)'\ \widehat{\Sigma}_{T_p}^{-1}\ \widehat{K}(t_{p+1},T_p) - \widehat{c}(s,t_{p+1}))^2\ \mathrm{d}s}{\widehat{K}(t_{p+1},t_{p+1}) - \widehat{K}(t_{p+1},T_p)'\ \widehat{\Sigma}_{T_p}^{-1}\ \widehat{K}(t_{p+1},T_p)}, \tag{3.46}$$

where $\widehat{c}(s,T_p)$ is the vector with entries $\widehat{\mathrm{cov}}(Y(s),X(t_j))$, $t_j \in T_p$, and equivalently for $\widehat{K}$.

Notice that the quotient of the integral in Equation (3.45) is zero under (3.32) only when $p = p^*$ (if not, it would mean that some α_j is identically zero), and $Q_0(T_{p+1}) = Q_0(T_p)$ for $p \geq p^*$. That is, the optimality criterion Q_0 does not reach its maximum value for $p < p^*$. Additionally, there is no room for improvement using $p > p^*$. Therefore, the number p^* would be the smallest p such that the maximum value of $Q_0(T_p)$ (or the minimum of Q_1) remains unchanged when increasing p.

As in the previous part of the chapter (Equation (3.27)), the sample version of this idea is as follows. Defining

$$\Delta = \min_{p<p^*} \big(Q_0(T_{p+1}^*) - Q_0(T_p^*)\big) > 0$$

and fixing some $0 < \epsilon < \Delta$, we set

$$\widehat{p} \ = \ \min\{p \ : \ \widehat{Q}_0^{max}(p+1) - \widehat{Q}_0^{max}(p) < \epsilon\}, \tag{3.47}$$

where $\widehat{Q}_0^{max}(p) = \max_{T_p\in\Theta_p} \widehat{Q}_0(T_p)$. This estimator is a.s. consistent for the true number of relevant variables.

Theorem 3.18. *Suppose that assumptions of Lemma 3.15 hold for $p \leq p^*$ and that p^* is the smallest integer such that Equation* (3.32) *is satisfied. Then the estimator given by Equation* (3.47) *fulfills $\widehat{p} \xrightarrow{\text{a.s.}} p^*$.*

Proof. We can prove similar results as the ones given in Lemma 3.10 using the same reasoning, with the only difference that now $Q_0(T_{p^*}^*) = \int_0^1 \|\beta(s,\cdot)\|_K^2 \mathrm{d}s$ (as in the proof of Proposition 3.13). □

That is, we are able to consistently estimate the number of elements in the projection whenever the slope function β meets the sparse representation of Equation (3.32). However, this condition may not hold for real problems. In this regard, the following result ensures us that the response process Y can be approximated arbitrarily well by increasing the number of points p in the sparse representation, even if the β function that defines the model is not sparse. Unlike the previous results of the current section, this result did not appear in the problem with scalar response, but the proof can be easily adapted to that context to see that $\|(Y-\varepsilon) - Y_{T_p^*}\|$ goes to zero as $p \to \infty$.

Proposition 3.19. *Let (X, Y) follow model (3.30), with β a continuous function in $[0,1]^2$ and $\varepsilon(s)$ uncorrelated with $X(t)$ for every $s, t \in [0,1]$, then*

$$\mathbb{E}\big\|\big(Y(\cdot)-\varepsilon(\cdot)\big) - Y(\cdot)_{T_p^*}\big\|_2^2 \;\to\; 0, \quad \textit{as} \quad p \to \infty,$$

where the set T_p^* contains the points that maximize Q_0 in Θ_p.

Proof. In this proof we use the notation f_{T_p}, for $f \in \mathcal{H}(K)$, to denote the projection of f on the closed subspace $\operatorname{span}\{K(t_i,\cdot), t_i \in T_p\}$. This proof is based on Equation (5.4) of Cambanis (1985). It states that, whenever $\boldsymbol{q}_p \in [0,1]^p$ form a regular sequence of points generated by a density h (that is, $\boldsymbol{q}_p$ are the quantiles of h), for every function $f \in \mathcal{H}(K)$,

$$\|f\|_K^2 - \|f_{\boldsymbol{q}_p}\|_K^2 \;=\; \|f - f_{\boldsymbol{q}_p}\|_K^2 \;\to\; 0 \;\text{ as }\; p \to \infty.$$

The first equality is due to the fact that we are projecting onto a closed subspace of $\mathcal{H}(K)$, so $\|f\|_K^2 = \|f_{\boldsymbol{q}_p}\|_K^2 + \|f - f_{\boldsymbol{q}_p}\|_K^2$. We can change the order of the integrals in the statement,

$$\begin{aligned}
\mathbb{E}\big\|\big(Y(\cdot)-\varepsilon(\cdot)\big) - Y(\cdot)_{T_p^*}\big\|_2^2 &= \int_0^1 \big\|\big(Y(s)-\varepsilon(s)\big) - Y(s)_{T_p^*}\big\|^2 \mathrm{d}s \\
&= Q_1(T_p^*) - \mathbb{E}\|\varepsilon\|_2^2 \;=\; Q_2(T_p^*) \\
&\leq \quad Q_2(\boldsymbol{q}_p) \;=\; \int_0^1 \big\|\beta(s,\cdot) - \beta(s,\cdot)_{\boldsymbol{q}_p}\big\|_K^2 \mathrm{d}s.
\end{aligned}$$

We now check that the sequence of functions $g_p(s) = \|\beta(s,\cdot) - \beta(s,\cdot)_{\boldsymbol{q}_p}\|_K^2$ converges to zero in $L^1[0,1]$. Since the points $\boldsymbol{q}_p$ form a regular sequence, we know that $g_p(s) \to 0$ as $p \to \infty$ pointwisely. Besides, these functions are bounded,

$$|g_p(s)| \;=\; \|\beta(s,\cdot)\|_K^2 - \|\beta(s,\cdot)_{\boldsymbol{q}_p}\|_K^2 \;\leq\; \|\beta(s,\cdot)\|_K^2,$$

and this bound is an integrable function, $\int_0^1 \|\beta(s,\cdot)\|_K^2 \mathrm{d}s \leq \sup_s \|\beta(s,\cdot)\|_K^2 < \infty$ (because $s \mapsto \|\beta(s,\cdot)\|_K^2$ is the composition of two continuous functions, $s \mapsto \beta(s,\cdot)$ and $f \mapsto \|f\|_K^2$, over the compact $[0,1]$). Then, the result follows from the dominated convergence theorem. □

3.7.4 Convergence of the finite-dimensional approximations in a particular case

We check here the accurateness of the approximations based on finite linear combinations (see (3.32)) for an arbitrary slope function β. In order to avoid the use of samples, which introduces noise to the measurements, we will work with the RKHS associated to the standard Brownian motion, since in this case we explicitly know the space $\mathcal{H}(K)$:

$$\mathcal{H}(K) \;=\; \{f \in L^2[0,1] \;:\; f(0)=0,\; f \text{ absolutely continuous and } f' \in L^2[0,1]\},$$

where f' denotes the almost everywhere derivative of f. The inner product of this space is given by

$$\langle f, g\rangle_K = \int_0^1 f'(s)g'(s)\mathrm{d}s, \quad \text{for } f, g \in \mathcal{H}(K).$$

Since different norms in function spaces are not equivalent, it is not obvious which norm to use to measure the discrepancies. We decided to use the $\mathcal{H}(K)$-norm, since it is the one that appears in our theoretical results. In addition, the L^2-norm is always less than the RKHS-norm. Thus, the differences provided here are greater than the differences in $L^2[0,1]$.

We approximate different kernels in $\mathcal{H}(K)$, increasing p from 1 to 3. First, we use one of the slope functions that we will use in the forecasting experiments of the next section: a finite-dimensional slope function as those in Equation (3.32) with points $T^* = (0.3, 0.5, 0.9)$ and weight functions $\beta_j(s) = \ln((1+s)j^{-1})$ for $j = 1, 2, 3$ and $s \in [0,1]$. Then we also consider the functions

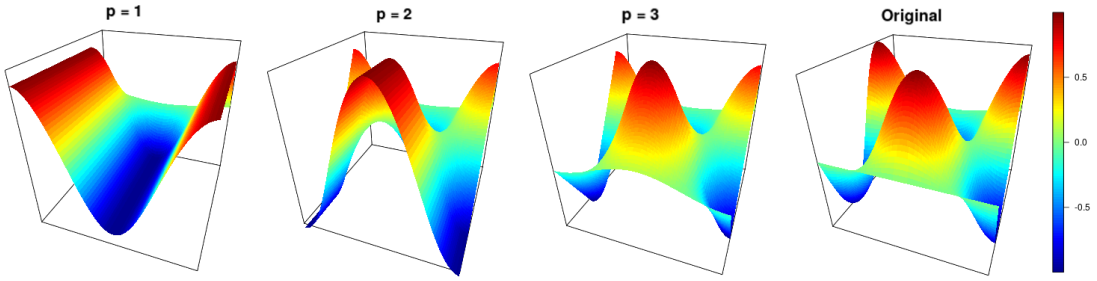
$$\beta^1(s,t) = \cos(2\pi s)\sin(2\pi t), \;\; \beta^2(s,t) = \sin(2\pi st), \text{ and } \beta^3(s,t) = -\ln(5st+1). \quad (3.48)$$

Other continuous functions could be considered as long as they fulfill $\beta(s,0) = 0$, for all $s \in [0,1]$, and the derivatives of $\beta(s,\cdot)$ lie in $L^2[0,1]$.

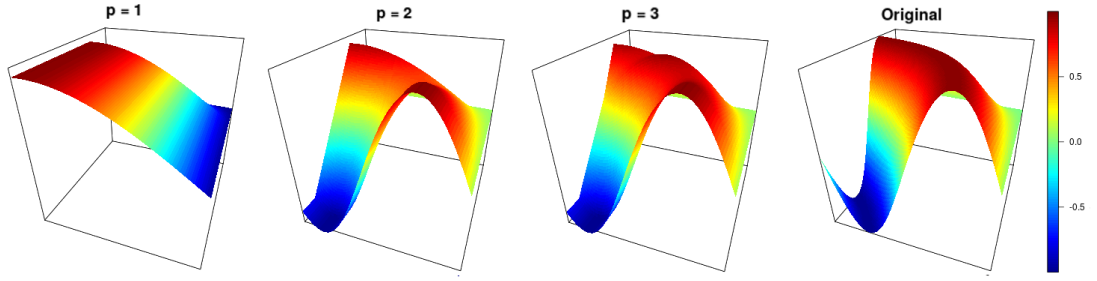
Figure 3.5 shows the approximated kernels. The distances $\|\beta(s,\cdot) - \beta(s,\cdot)_{T_p^*}\|_K$ for every $s \in [0,1]$, where $\beta(s,\cdot)_{T_p^*}$ denotes the projection of $\beta(s,\cdot)$ on $\mathrm{span}\{K(t_i,\cdot), t_i \in T_p^*\}$, for p from 1 to 20, are plotted in Figure 3.6. We can see that these distances go to zero for every s, although this convergence does not seem to be at the same rate for all the points. As expected, the distances for the sparse representation are zero for all $s \in [0,1]$ when $p = 3$ (the true model).

3.8 An application to functional time series forecasting

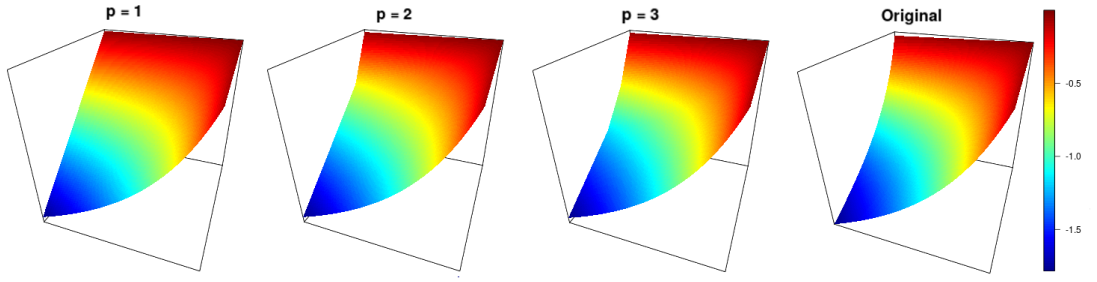
In the previous section the realizations of the process X are assumed to be independent. However, for some applications this assumption is not reasonable. Sometimes the data consists of a single curve $z(s)$ sequentially recorded in time, up to an instant point



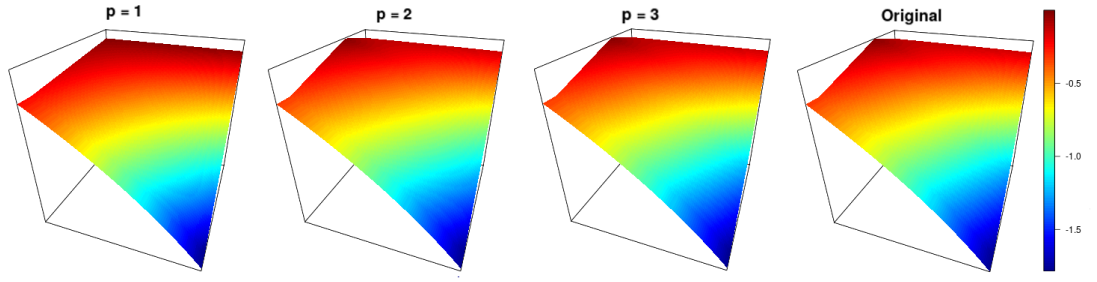
(a) Function β^1 of Equation (3.48).



(b) Function β^2 of Equation (3.48).

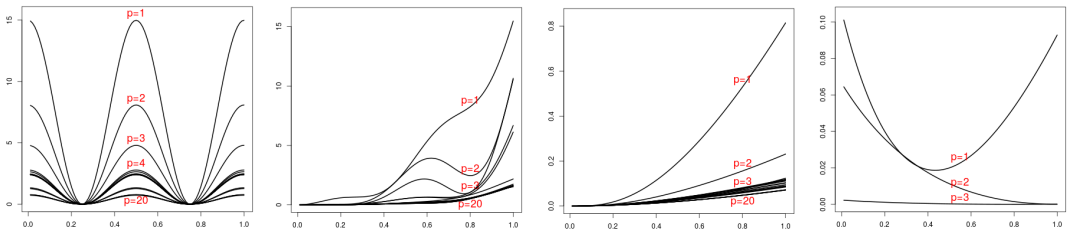


(c) Function β^3 of Equation (3.48).



(d) Sparse slope function with logarithms.

Figure 3.5: Approximations of the functions $\beta(s,\cdot) \in \mathcal{H}(K)$ when increasing p in Equation (3.32).



(a) β^1 of Eq. (3.48). (b) β^2 of Eq. (3.48). (c) β^3 of Eq. (3.48). (d) Sparse with logs.

Figure 3.6: Distances $\|\beta(s,\cdot) - \beta(s,\cdot)_{T_p^*}\|_K$, for $s \in [0,1]$.

$s \in (-\infty, N]$ (or $[0, N]$). Typically these curves present a periodical behavior. Common examples are daily financial or meteorological records. This directly suggests to split the curve z into pieces of the same length, sharing a common structure. These new curves are denoted as $x_n \equiv x_n(\cdot)$, for $n \leq N$ (where this n has nothing to do with the sample size), and are usually rescaled to the interval $[0,1]$ in order to make them independent of the time unit of measure. Each one of these curves x_n is understood to be a realization of the corresponding random process X_n. The set of these processes X_n, indexed by $n \in \mathbb{Z}$, is known as *functional time series* (see Álvarez-Liébana (2017) for details).

In this section we adapt the previous methodology to define a predictor based on a functional autoregressive (AR) model, especially suitable for variable selection purposes. Bosq (2000) gives a good introduction into the field of linear processes in function spaces and introduces functional autoregressive processes in depth. AR models have shown to be a valuable tool in functional time series analysis since they combine computational tractability with sufficient generality (e.g. Besse and Cardot (1996); Kargin and Onatski (2008); Didericksen et al. (2012); Aue et al. (2015)).

Classical approaches

The standard assumption in the literature is that $L^2[0,1]$, the space of square-integrable functions on $[0,1]$, is the space into which the curves fall. This is a sensible choice, since $L^2[0,1]$ is a separable Hilbert space and offers desirable geometric properties through the definition of the natural scalar product. However, considering our variable selection purpose, the main drawback of $L^2[0,1]$ is that, strictly speaking, it consists of equivalence classes of functions. That is, two functions represent the same L^2-function if the set where they differ has zero measure. In other words, for any particular point $s \in [0,1]$, the value $f(s)$ is not well-defined for a function f in this space.

Since we require pointwise evaluations $x_n(s)$ of the curves, the space of continuous functions on $[0,1]$, $C[0,1]$, which is a Banach space with the supremum-norm, is a more natural space to work in. In addition, the subsequent change of norm allows us

to obtain uniform convergence results. The problem of estimating AR models in $C[0,1]$ has been already addressed in the literature (e.g. Ruiz-Medina and Álvarez-Liébana (2018) and references therein). The usual methodology is to project the curves (via the inner product in $L^2[0,1]$) onto the finite dimensional subspace spanned by some eigenfunctions of the covariance operator of the data, as in Pumo (1998). Some limitations of this principal component approach have been discussed extensively in the literature. For instance, the resulting space is shown to be optimal in order to represent the variability of the process, but the dependence might be lost by the dimension reduction (Kargin and Onatski (2008) and Hörmann et al. (2015)). Furthermore, Bernard (1997) points out the sensitivity of the proposal to small errors in the estimation of small eigenvalues.

Some advantages of this methodology

In this work the projection on a finite dimensional space is replaced by the choice of some evaluations $x_n(t_1), \ldots, x_n(t_p)$. In the cases relevant for variable selection, it is shown that the new model falls into the wide class of Banach space–valued processes (ARB(q)) introduced in Bosq (2000), which directly gives us sufficient conditions for the existence of a unique stationary solution. Notably, for instance the well-known Ornstein-Uhlenbeck process satisfies our model. In these cases we are able to adapt the results of the previous section. Besides, our predictor coincides with the optimal one in this setting, in the sense that it is the best probabilistic predictor, as stated in Mokhtari and Mourid (2003).

The advantages of predicting autoregressive processes with this new approach are numerous, apart from the ones already mentioned. From an applied perspective, the proposed method is flexible concerning the structure of the data: whether it is recorded on a grid or available as continuous functions - the methodology remains similar with slight technical differences. Besides, the use of this RKHS based model avoids the need of inverting the covariance operator since this is carried out, in some sense, by the inner product of the space.

In order to show the practical relevance of the method, a simulation study is conducted. To evaluate the performance in the real world four data sets are studied. The execution times of the tested methods are also analyzed. Our proposal is competitive both in prediction accuracy and computational efficiency.

Related literature

The literature in the field of functional time series analysis is developing quickly. Recent publications include time-domain methods like Hörmann and Kokoszka (2010), where a weak dependence concept is introduced, Aue et al. (2015), Klepsch and Klüppelberg (2017) and Klepsch et al. (2017), where prediction methodologies based on linear models

are developed, and Aue and Klepsch (2017), where an estimator of functional linear processes based on moving average model fitting is derived. Besides, there are some proposals for feature selection on multivariate time series, like Fan and Lv (2010); Lam and Yao (2012); Liu (2014); Tran et al. (2015) and the references therein. However, as far as we know, there are no previous approaches to variable selection for continuous time series in the same sense as it is presented here.

Specific notation

We include here the specific notation for time series context, which has not been previously used in the chapter. We work with a time series $X_n(\cdot)$, $n \in \mathbb{Z}$, taking values in $C[0,1]$, the space of continuous functions over $[0,1]$, such that $\mathbb{E}\big[(\sup_s |X(s)|)^2\big] = \|\sup_s |X(s)|\|^2 < \infty$. That is, the supremum of the trajectories, as well as each $X_n(s)$, belong to $L^2(\Omega)$, the space of square integrable random variables. The supremum norm in $[0,1]$ is denoted by $\|\cdot\|_\infty$.

Given a second order time series X_n, we define its lagged covariance function $c_r(s,t)$, for $s,t \in [0,1]$, $r \in \{0\} \cup \mathbb{N}$, as $\mathrm{cov}(X_r(s), X_0(t))$. This time series is said to be weakly stationary (or just stationary) if its mean function is constant over time and $\mathrm{cov}(X_r(s), X_0(t)) = \mathrm{cov}(X_{r+n}(s), X_n(t))$ for all $s,t \in [0,1]$ and $n \in \mathbb{N}$. Thus, the covariance function $K(s,t)$ will be denoted here as $c_0(s,t)$. Likewise, $c_1(s,t)$ will play the role of the cross-covariance function $c(s,t) = \mathrm{cov}(Y(s), X(t))$ of the previous section. For the sake of clarity in the equations, and since the time series throughout this work are stationary, we make the following abuse of notation: we assume to work with the centered processes $X_n - \mathbb{E}[X_n]$, denoted simply by X_n.

A sample of the time series is denoted as $x_1, \ldots, x_m$, where each x_j is drawn from the corresponding process $X_1, \ldots, X_m$. The sample size is denoted as m in this section in order to avoid possible misunderstandings with the index of the time series.

3.8.1 Model definition

Given $x_n \in C[0,1], n \in \mathbb{Z}$, trajectories drawn from a functional time series X_n, the standard autoregressive model is of the form (see Chapter 6 Bosq (2000))

$$x_n = \rho(x_{n-1}) + \varepsilon_n, \qquad n \in \mathbb{Z}, \tag{3.49}$$

for some white noise process $\varepsilon_n, n \in \mathbb{Z}$, in $C[0,1]$ and some bounded linear operator ρ. In this section we propose a different functional autoregressive model "customized" to give a theoretical framework for variable selection.

We work with $X_n, n \in \mathbb{Z}$, a centered stationary process such that $\mathbb{E}\big[\big(\sup_s |X_n(s)|\big)^2\big] < \infty$. In this context we propose to replace (3.49) with

$$X_n(s) = \langle \phi(s,\cdot), X_{n-1}(\cdot)\rangle_{c_0} + \varepsilon_n(s), \qquad n \in \mathbb{Z}, \tag{3.50}$$

where $\langle\cdot,\cdot\rangle_{c_0}$ denotes the scalar product of the RKHS generated by the autocovariance function of $X_n, n \in \mathbb{Z}$, and for some appropriate kernel ϕ such that $\phi(s,\cdot) \in \mathcal{H}(c_0)$ for all $s \in [0,1]$. The error $\varepsilon_n, n \in \mathbb{Z}$, is a strong $C[0,1]$-white noise Bosq (2000, p. 148) pointwisely uncorrelated with X_n. Note that, since the process is stationary, its covariance structure remains invariant and then the space $\mathcal{H}(c_0)$ does not change with n. As already mentioned, the trajectories of the process do not belong to $\mathcal{H}(c_0)$, thus we understand the inner product of the previous equation as $\Psi^{-1}_{X_{n-1}}(\phi(s,\cdot))$, where $\Psi_{X_{n-1}}$ denotes the Loève's isometry (as suggested in Parzen (1961a), just as we have done previously in this chapter).

Equation (3.50) is simply the pointwise definition, for each $s \in [0,1]$, of a fully functional model. This definition can be applied for any process such that $\mathbb{E}\big[(\sup_s |X_n(s)|)^2\big] < \infty$ for $n \in \mathbb{Z}$, since then $X_n(s) \in L^2(\Omega)$ and Loève's isometry can be applied. Moreover, using the definition of $\Psi_{X_{n-1}}$ we rewrite Equation (3.31) in this context,

$$\phi(s,\cdot) \;=\; \Psi_{X_{n-1}}\big(X_n(s) - \varepsilon_n(s)\big) \;=\; \mathbb{E}\big[\big(X_n(s) - \varepsilon_n(s)\big)X_{n-1}(\cdot)\big] \;=\; c_1(s,\cdot), \tag{3.51}$$

and then

$$\|c_1(s,\cdot)\|_{c_0} = \|\phi(s,\cdot)\|_{c_0} = \|X_n(s) - \varepsilon_n(s)\| < \infty.$$

That is, the pointwise evaluations $X_n(s)$ of the process given in Equation (3.50) can be written as $\Psi^{-1}_{X_{n-1}}(c_1(s,\cdot)) + \varepsilon_n(s)$, which is always well-defined. However, it has to be carefully analysed whether or not the model (3.50) can be understood as a fully functional model (in the same vein as (3.49)). From the last displayed equations we also see that, when changing the working space from $L^2[0,1]$ to $\mathcal{H}(c_0)$, the solution of the model does not require to invert the covariance operator. It could be understood as if the "inversion" was intrinsically carried out by the inner product of $\mathcal{H}(c_0)$.

Whenever ϕ follows a sparse representation as those in Equation (3.32),

$$\phi(s,\cdot) \;=\; \sum_{j=1}^{p} \alpha_j(s) c_0(t_j,\cdot) \;\in\; \mathcal{H}(X), \tag{3.52}$$

model of Equation (3.50) reduces to

$$X_n(s) = \sum_{j=1}^{p} \alpha_j(s) X_{n-1}(t_j) + \varepsilon_n(s). \tag{3.53}$$

In this section we analyze only this finite family, since we are mainly interested in variable selection. In view of the previous discussion, we propose the following definition.

Definition 3.20. A sequence X_n such that $\mathbb{E}\big[(\sup_s |X_n(s)|)^2\big] < \infty$, $n \in \mathbb{Z}$, is called "sparse functional autoregressive process of order 1" (FCAR-sparse(1)) if it is stationary and such that

$$X_n(\cdot) \;=\; \sum_{j=1}^{p} \alpha_j(\cdot) X_{n-1}(t_j) + \varepsilon_n(\cdot) \;\equiv\; \Psi^{-1}_{X_{n-1}}(\phi(\cdot,\star)) + \varepsilon_n(\cdot), \quad n \in \mathbb{Z} \tag{3.54}$$

where $\phi(s,\cdot) \in \mathcal{H}(c_0)$ is as in Equation (3.52) and $\varepsilon_n, n \in \mathbb{Z}$, is a strong $C[0,1]$-white noise (which implies $\mathbb{E}(\sup_s |\varepsilon_0(s)|)^2 < \infty$) pointwisely uncorrelated with X_n.

In order to make sense of this functional definition and to be able to obtain some properties about the process, we make use of a more general family. In view of Equation (3.54), one can understand that each realization x_n equals $\rho(x_{n-1}) + \varepsilon_n$, where ρ is the operator:

$$\rho(f) = \sum_{j=1}^{p} \alpha_j(\cdot) f(t_j) \quad \text{ for } f \in C[0,1]. \tag{3.55}$$

We prove in Proposition 3.21 that this interpretation is well founded. Note that this operator depends on the covariance function $c_1(s,t)$, since it uses the same set of points t_j and functions α_j that define it (by Equation (3.51)).

Proposition 3.21. *Let X_n follow a FCAR-sparse(1) model of Definition 3.20 such that $\sum_{j=1}^{p} \|\alpha_j\|_\infty < 1$, then* (3.54) *has a unique strictly stationary solution given by*

$$x_n = \sum_{j=0}^{\infty} \rho^j(\varepsilon_{n-j}), \quad n \in \mathbb{Z},$$

where ρ is defined in Equation (3.55)*. The series converges almost surely and*

$$\mathbb{E}\Big[\Big(\sup_s \Big|X_n(s) - \sum_{j=0}^{p} \rho^j(\varepsilon_{n-j})(s)\Big|\Big)^2\Big] \to 0 \quad \text{as } p \to \infty.$$

Proof. This proof relies on the theory of Banach space valued autoregressive (ARB) processes (introduced in Definition 6.1 of Bosq (2000)) with $B = C[0,1]$. First we check that our model follows an ARB model as in (3.49). That is, that the operator of Equation (3.55) is bounded. It follows from the definition of the norm in the space of linear operators,

$$\begin{aligned}
\|\rho\|_{\mathcal{L}} \;&=\; \sup_{\|f\|_\infty \le 1} \Big\| \sum_{j=1}^{p} \alpha_j(\cdot) f(t_j) \Big\|_\infty \;=\; \sup_{\|f\|_\infty \le 1} \sup_{s \in [0,1]} \Big| \sum_{j=1}^{p} \alpha_j(s) f(t_j) \Big| \\
&\le\; \sup_{\|f\|_\infty \le 1} \sup_{s \in [0,1]} \sum_{j=1}^{p} |\alpha_j(s)|\, |f(t_j)| \;\le\; \Big(\sup_{s \in [0,1]} \sum_{j=1}^{p} |\alpha_j(s)| \Big) \Big(\sup_{\|f\|_\infty \le 1} \sup_{t \in [0,1]} |f(t)| \Big)
\end{aligned}$$

$$\leq \sum_{j=1}^{p} \sup_{s\in[0,1]} |\alpha_j(s)| = \sum_{j=1}^{p} \|\alpha_j\|_\infty < 1.$$

The operator norm satisfies $\|AB\|_{\mathcal{L}} \leq \|A\|_{\mathcal{L}}\|B\|_{\mathcal{L}}$, for A, B operators in $\mathcal{L}$. Then the result follows from the corollary of Theorem 6.1 in Bosq (2000) since

$$\|\rho^{j_0}\|_{\mathcal{L}} \leq \|\rho\|_{\mathcal{L}}^{j_0} < 1,$$

for all $j_0 \in \mathbb{N}$. □

The condition $\sum_{j=1}^{p} \|\alpha_j\|_\infty < 1$ may be relaxed, as it is deduced from the proof, since it is enough to have $\|\rho^{j_0}\|_{\mathcal{L}} < 1$. Proving this result for a general kernel function ϕ is not straightforward. Each $\phi(s,\cdot)$ can be written as a pointwise limit of functions in $\mathcal{H}_0(c_0)$, and then the operator ρ of Equation (3.55) would be defined as a limit of operators of finite rank. But then one needs to directly impose the artificial condition that this operator is bounded, which is something we usually cannot guarantee. However, the following example (taken from Bosq (2000)) shows that the finite-dimensional model defined in (3.53) still holds for some well-known processes.

Example 3.22. *Let z be a trajectory of a continuous version of the Ornstein-Uhlenbeck process,*

$$z(s) = \int_{-\infty}^{s} e^{-\theta(s-t)} \mathrm{d}B(t),$$

where B is a standard Brownian motion. If we define $x_n(s) = z(n+s)$ for $s \in [0,1]$, then x_n can be rewritten as in Equation (3.54). In this setting the operator ρ is given, for $f \in C[0,1]$, by $\rho(f)(s) = e^{-\theta s} f(1)$, $s \in [0,1]$, that is, $p = 1$, $t_1 = 1$ and $\alpha_j(s) = e^{-\theta s}$. Then $x_n = \rho(x_{n-1}) + \varepsilon_n$, where now $\varepsilon_n(s)$ is a white noise given by $\int_n^{n+s} e^{-\theta(n+s-t)} \mathrm{d}B(t)$.

As shown in Proposition 2 of Mokhtari and Mourid (2003), if the true kernel of the model is as in Equation (3.52), the best linear predictor of $(X_n - \varepsilon_n)$ based on $X_{n-1}(t_1), \ldots, X_{n-1}(t_p)$ is the best probabilistic predictor.

An extension of model FCAR-sparse(1) defined in Equation (3.54) to FCAR-sparse(q) can be carried out whenever $X_n(s) = Z(s+n)$ for $s \in [0,1]$ and Z is a stationary process with continuous trajectories. We define $Z_{n,q}(s) = Z(s+n-q+1)$, $s \in [0,q]$, which is basically the concatenation of $X_{n-(q-1)}, \ldots, X_n$. In this case we can write

$$X_n(s) = \Psi_{Z_{n-1,q}}(\phi(s,\cdot))^{-1} + \varepsilon_n(s), \ s \in [0,1]$$

where now each $\phi(s,\cdot)$ belongs to the RKHS associated with $Z_{n-1,q}$ and q is the minimum for which this model holds. In this case, Equation (3.51) is

$$\phi(s,t) = \mathbb{E}[(X_n(s) - \varepsilon_n(s))Z_{n-1,q}(t)] = c_i(s,t) \ \text{ for } t \in (q-i, q-i+1].$$

All the results and comments in the remainder of this section are valid in this case, with some additional assumptions. For the sake of simplicity we present here the case $q = 1$. Nevertheless, we include some comments where applicable to clarify the changes due to this extension.

3.8.2 Adaptation of the asymptotic results to FCAR-sparse

Analyzing the proofs of the previous section, all of them are sustained by the convergence of the sample covariance functions $\widehat{\text{cov}}(X(s), X(t))$ and $\widehat{\text{cov}}(Y(s), X(t))$ of Lemmas 3.5 and 3.14. With the next result (based on a result of Pumo (1998)) we adapt these two lemmas for AR processes, where the estimator of c_1 is given by

$$\widehat{c}_1(s,t) = \frac{1}{m-1} \sum_{i=1}^{m-1} x_{i+1}(s) x_i(t).$$

The "uniformly geometrically strong mixing" condition that appears in the statement is standard in functional time series context and can be found in Pumo (1998, Section 2.1) (see also Bosq (2000, p.58)). Then, all the proofs of the results presented in Section 3.7 are also valid for FCAR-sparse processes, merely reading the response Y as X_n. Besides, due to the stationarity of X_n, the optimality criteria do not depend on n.

Lemma 3.23. *Assume that $X_n, n \in \mathbb{Z}$, is a FCAR-sparse(1) process satisfying the same assumptions of Proposition 3.21 and:*

H1. The process $X_n(t)X_n(s)$ for $t, s \in [0,1]$ is uniformly geometrically strong mixing.

H2. (Cramer conditions) For every $t, s \in [0,1]$ there exist $d > 0$ and $D < \infty$ such that

- $d \leq \mathbb{E}[X_0^2(t)X_0^2(s)] \leq D$ *and*
- $\mathbb{E}|X_0(t)X_0(s)|^k \leq D^{k-2}k!\, \mathbb{E}[X_0^2(t)X_0^2(s)]$ *for* $k \geq 3$.

Then,

$$\sup_{t,s\in[0,1]} |\widehat{c}_0(s,t) - c_0(s,t)| \xrightarrow{\text{a.s.}} 0 \qquad \text{and} \tag{3.56}$$

$$\sup_{t,s\in[0,1]} |\widehat{c}_1(s,t) - c_1(s,t)| \xrightarrow{\text{a.s.}} 0. \tag{3.57}$$

Proof. By Lemma 1 of Pumo (1998) we know that for some positive constants A_1, A_2, A_3,

$$\begin{aligned} \mathbb{P}\left(\sup_{t,s\in[0,1]} |\widehat{c}_0(s,t) - c_0(s,t)| \geq \epsilon \right) \leq\ & (2\sqrt{m} + A_1)\ \exp\left(-A_2\epsilon^2\sqrt{m}\right) \\ & + A_3\epsilon^{\frac{2}{5}} m\ \exp\left(-\log(r^{-1})\sqrt{m}\right), \end{aligned}$$

where $0 < r < 1$ is given by assumption H1. By Borel-Cantelli, if the series in m whose terms are these probabilities is convergent for every $\epsilon > 0$, we get the almost sure convergence stated in Equation (3.56). The sum is of order

$$\sum_{m=1}^{\infty} \mathbb{P}\left(\sup_{t,s\in[0,1]} |\widehat{c}_0(s,t) - c_0(s,t)| \geq \epsilon \right) \sim 2\sum_{m=1}^{\infty} \frac{\sqrt{m}}{e^{C_\epsilon \sqrt{m}}} + \sum_{m=1}^{\infty} \frac{A_1}{e^{C_\epsilon \sqrt{m}}} + D_\epsilon \sum_{m=1}^{\infty} \frac{m}{e^{C_r \sqrt{m}}},$$

where $C_\epsilon, C_r, D_\epsilon > 0$ and these three series converge (for example by the limit comparison test with $\sum m^{-\gamma}, \gamma > 1$). Concerning (3.57), the same Lemma 1 of Pumo (1998) states that the bounds for these probabilities are equivalent but with $m-1$ in place of m. □

Cramer conditions appear often in the literature related with limit theorems for AR processes in Banach spaces. For instance, all bounded processes satisfy them, and also the Ornstein-Uhlenbeck process of Example 3.22. In the latter case, $|X_n(s)X_n(t)| = e^{-k(t+s)} X_{n-1}(1)^2$, then, $e^{-k(t+s)}\mathbb{E}[X_0(1)^{2k}] \leq D^{k-2}k! e^{-2(t+s)}\mathbb{E}[X_0(1)^4]$, where $X_0(1) \sim \mathcal{N}(0, 0.5)$. Using the expression for the moments of a Gaussian variable,

$$e^{-k(t+s)} \frac{(2k)!}{2^{2k}k!} \leq D^{k-2} e^{-2(t+s)} \frac{3k!}{4},$$

which is satisfied, for instance, for $D \geq 5e^{-2}/12$.

For the case of greater order FCAR-sparse(q) with $q > 1$, this Lemma 3.23, and therefore all the results of the previous section, hold whenever the process $Z_{n,q}$ fulfils assumptions H1 and H2. This is equivalent to suppose that all the products $X_i(t)X_j(s)$ satisfy H1 and H2 for $0 \leq i, j \leq q-1$. Besides, in order to check the equivalence of the optimality criteria Q_0, Q_1 and Q_2 of Proposition 3.13, we have to substitute the function $c_1(s,t)$ by the continuous picewise-defined function

$$c(s,t) = c_i(s, t-q+i) \text{ for } t \in (q-i, q-i+1] \text{ and } s \in [0,1]. \tag{3.58}$$

We need the additional assumption that the marginal variables of $Z_{n,q}$ are all linearly independent to ensure the invertibility of the covariance matrices of $Z_{n,q}$ evaluated in $(t_1^1, \ldots, t_{p^q}^q) \in [0,q]^p$. This assumption introduces some further restrictions to the model. For instance, the functions $\alpha_j^i(s)$ can not vanish for $s \in [0,1]$ and $1 < i \leq q$.

3.8.3 Experimental setting

In this section we introduce the data sets which appear along the experiments (both simulated and real), as well as other methods of the literature used for comparison. For the implementations we use the greedy approach described previously in this chapter (but using the adapted functional definitions of Q_0 given in Proposition 3.13). The number of points is fixed as explained at the end of 3.4 and also by cross-validation.

We compare the efficiency of the proposal with two other recent methods. Both of them carry out the dimension reduction using functional principal components. For the forecasting experiments we also compare with two "base" methods that do not reduce dimension. We indicate in brackets the names used in the tables for each method. The entries marked with bold letters in these tables correspond to the best performance in each case.

- The method proposed in this chapter (RKHS) has been implemented in four different ways. As mentioned, we use two approaches to select the number of relevant variables; doing clustering on the maximum values of the $\widehat{Q}_0$ functions (CL) and by cross-validation (CV). In addition, the points can be selected by using covariance vectors on a grid or computing purely functional lagged-covariance functions. We use one or the other depending on the nature of the data.
- The method proposed in Aue et al. (2015) (fFPE). This proposal uses a dimension reduction technique based on functional principal component analysis to find a finite dimensional space on which the prediction is performed using a vector autoregressive model. The model order and dimension of the finite dimensional space are chosen by the fFPE criterion. For details, see Aue et al. (2015), where the empirical properties of the approach are demonstrated in depth. The R-code of this method was provided by the authors.
- The method proposed in Bosq (2000) and Kokoszka and Reimherr (2013) (KR). This prediction method by Bosq is the one known as the standard prediction method for functional autoregressive processes. To determine the order of the functional autoregressive model to be fitted, we use the multiple testing procedure of Kokoszka and Reimherr (2013).
- Exact and Naive methods are implemented in order to provide some bounds on the errors. These methods are also used, for instance, in Horváth and Kokoszka (2012). The exact prediction consists in "predicting" the response directly as $\rho(x_{n-1})$. Therefore, it can be only applied for simulated data, since the operator ρ is unknown for real data sets. It is not really a prediction method but gives us an idea of the minimum error that we can achieve. The Naive approach simply predicts $\widehat{x}_n$ as x_{n-1}.

Both the maximum number of points to select and the number of principal components are limited to 10. For the simulated data all methods are tested using a sample of size $n = 115$, where 100 realizations are used for training and the remaining 15 for test. Each experiment has been replicated 100 times. For the real data sets we use a window moving approach with five blocks to obtain several measures of the errors. The size of the windows is adjusted depending on the sample size of each set. The order of the process is always limited to 3 for the methods. However, for our proposal we have to set it to order 1 whenever the curves can not be interpreted as $X_n(s) = Z(s+n)$, with Z continuous.

Usually the functional data sets are given in a discretized fashion. Some of the tested methods require to transform previously the data to be truly functional. However, our discrete proposal can also deal with discretized data. In addition, when the data is irregular, some information could be lost when transforming the data to functional. This complicates the comparison between the different methods. Therefore, for this kind of discretized data sets we measure two different types of errors.

- Discrete errors: The error is measured using the original discretized data. The discrete version of the proposal (the one that uses covariance vectors) is tested. The predictions returned by the methods that use fully functional data are evaluated on the same grid given by the data.
- Functional errors: The data are transformed to functions using a Bsplines basis before applying the methods. We have found that using Bsplines is more suitable in this setting, since the standard Fourier basis introduces periodicity in the data. For the displayed results 10 functions of the basis are used, but different numbers have been tested without significant changes. The functional implementations of the RKHS proposal, which estimate the whole lagged-covariance functions in a purely functional way, are tested in this case.

As we see in the following subsection, one of the simulated sets is purely functional. In this case only the functional errors are measured. Two different norms are used to measure the error: the standard $L^2[0,1]$ norm and the supremum norm of $C[0,1]$ that has been used along this section. Each of these norms measure different characteristics of the predictions. We also measure two different relative errors,

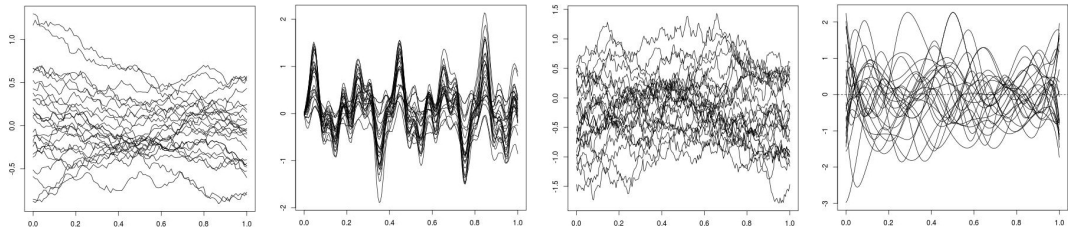
$$\varepsilon_1 = \sum_{i=1}^{n} \frac{\|x_i - \widehat{x}_i\|}{\|x_i\|}, \qquad \varepsilon_2 = \frac{\sum_{i=1}^{n} \|x_i - \widehat{x}_i\|}{\sum_{i=1}^{n} \|x_i\|}. \tag{3.59}$$

The first one gives the same importance to all curves regardless of their norm, while the second one place more importance to the errors in the curves of biggest norms, since it is just a scaling of the absolute error.

Simulated data sets

We test the different methods using simulated sets that fulfill the sparsity assumption of Equation (3.13) as well as some which not. Most of them are inspired by other data sets used in the literature. Some realizations of these processes can be found in Figure 3.7.

- Two data sets satisfying the sparsity assumption with standard Brownian innovations. The true points are $T^* = (0.3, 0.5, 0.9)$ with two different sets of functions α_j. The first ones are logarithms, $\ln((1+s)j^{-1})$ for $j = 1, 2, 3$ and $s \in [0,1]$, similar to the function used for the simulated data of Section 3.7.4. The second set of



(a) Sparse with logs. (b) Sparse with sins. (c) O.U. (d) FAR

Figure 3.7: 25 trajectories of each of the simulated data sets.

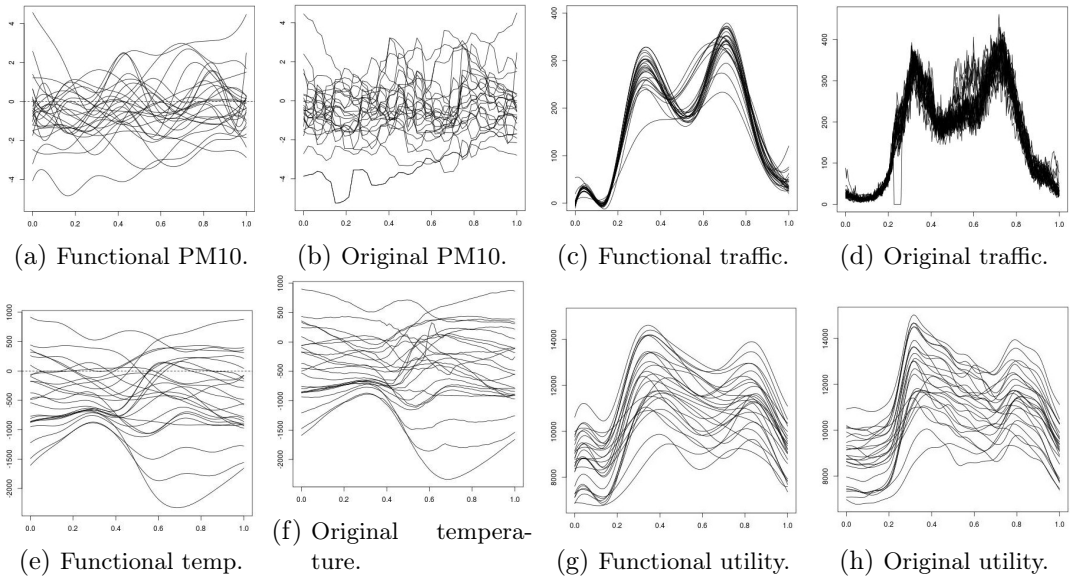
functions is $\sin(30\pi j^{-1}s)$, since we also want to incorporate a data set with high variation. When transforming this last data set to purely functional, 30 Bspline functions are used instead of 10, to be able to capture most of the variation.

- Ornstein-Uhlenbeck process introduced in Example 3.22. This is the only simulated set for which $X_n(s) = Z(s+n)$, so that we can use the model FCAR-sparse(3).
- FAR process with linearly decaying eigenvalues of the covariance operator ($s = 1, \ldots, 15$). Following the simulation example used in Aue et al. (2015), this set consists of spanning a D-dimensional space by the first D Fourier basis functions, and then generate random $D \times D$ parameter matrices and a D-dimensional noise process, where the construction ensures a linear decay of the eigenvalues of the covariance operator. The slow decay of these eigenvalues makes sure that problems with PCA based methods due to non-invertibility of the covariance operator are avoided. In this example only the functional errors are measured since it is purely functional by construction.

Real data sets

We analyze also some real data sets, a couple of them already used in other recent papers.

- Particulate matter concentrations (PM10). This data set is used, for instance, in Aue et al. (2015) and consists on 175 samples. It contains the $\mu g m^{-1}$ concentration in air of a particular substance with aerodynamic diameter less than 10 μm. The measures were taken each half hour from October 1, 2010 to March 31, 2011 in Austria. The data are preprocessed in the same way as suggested in Aue et al. (2015). For the five windows we take blocks of 115 observations, 100 for training and 15 for test.
- Vehicle traffic data (Traffic) presented in Aue and Klepsch (2017). The original data set was provided by the Autobahndirektion Südbayern. It contains the



(a) Functional PM10. (b) Original PM10. (c) Functional traffic. (d) Original traffic.

(e) Functional temp. (f) Original temperature. (g) Functional utility. (h) Original utility.

Figure 3.8: 25 trajectories of the real data sets, both discrete and functional.

amount of vehicles traveling each five minutes on the highway A92 in Southern Bavaria, Germany, from January 1 to June 30, 2014. Retaining only working days, we work with 119 samples divided into 5 windows of size 99; 94 for train and 5 for test.

- Indoor temperature of a "solar house" (Temp). This data set consist in temperature measures each 15 minutes during 42 days in the living room of a SMLsystem solar house. The whole data set (which contains other different attributes) is studied in Zamora-Martínez et al. (2014) and it is available on `http://archive.ics.uci.edu/ml/datasets/SML2010`. This is the smallest set, so it is divided into 5 windows of size 34, from which just 2 curves are used for test. For this set we were forced to use at most 9 PCA components for the fFPE method, in order to avoid computational errors.

- Utility demand data (Utility) which appears in the book Hyndman et al. (2008) and is available in the R package "expsmooth" (Hyndman (2018)). The original set is made of 126 curves of hourly utility demand from a company of the the Midwestern United States, starting on January 2003. Since this work is focused on variable selection for data sampled on a fine grid, the curves have been subsampled to simulate observations each 15 minutes. The five windows into which the curves are split consist in 100 samples for train and 5 for test.

A few curves of these data sets are included in Figure 3.8.

Forecasting experiments

The main goal for which our proposal is designed is the prediction of time series. Accordingly, the greatest part of the experiments is devoted to forecasting.

Table 3.7 summarizes the measurements for the simulated data sets of the two types of errors ε_1 and ε_2 (Equation (3.59)). Regarding our two proposals, there is not a method that uniformly outperform the other one. That is, both cluster and cross-validation perform well when it comes to select the number of points. In general, our proposals are mainly the winners, closely followed by the fPCA approach with the fFPE criterion. These are the expected results, since three out of the four data sets fulfill the sparse model of Equation (3.52). In any case, our proposal also slightly outperforms the others for the FAR data, where this sparsity assumption is far from being satisfied.

				RKHS+cl	RKHS+CV	fFPE	KR	Exact	Naive
Sparse with log.	ε_1	Disc.	L2	**0.64**	0.71	0.65	0.66	0.55	3.24
			sup	**0.68**	0.71	**0.68**	0.82	0.65	1.64
		Funct.	L2	0.65	**0.64**	0.67	0.67	0.56	3.32
			sup	**0.65**	**0.65**	**0.65**	0.83	0.62	1.67
	ε_2	Disc.	L2	**0.16**	0.18	0.17	0.18	0.14	1.32
			sup	**0.30**	0.32	**0.30**	0.39	0.29	0.80
		Funct.	L2	**0.32**	**0.32**	0.33	0.34	0.28	2.63
			sup	**0.55**	**0.55**	0.56	0.77	0.53	1.57
Sparse with sins	ε_1	Disc.	L2	**0.72**	0.74	0.83	0.94	0.60	2.42
			sup	**0.71**	0.73	0.84	0.95	0.67	1.50
		Funct.	L2	**0.78**	0.81	0.81	0.94	0.65	2.60
			sup	**0.77**	**0.77**	**0.77**	0.95	0.69	1.53
	ε_2	Disc.	L2	**0.38**	**0.38**	0.42	0.48	0.33	1.10
			sup	**0.36**	0.37	0.42	0.49	0.34	0.75
		Funct.	L2	**0.78**	0.79	0.80	0.91	0.69	2.19
			sup	**0.75**	**0.75**	**0.75**	0.94	0.68	1.47
O.U.	ε_1	Disc.	L2	1.07	1.01	**1.00**	1.15	0.83	2.33
			sup	**0.88**	**0.88**	0.93	0.98	0.88	1.35
		Funct.	L2	**1.00**	**1.00**	1.05	1.20	0.85	2.49
			sup	**0.91**	**0.91**	0.92	0.98	0.86	1.34
	ε_2	Disc.	L2	**0.31**	**0.31**	0.35	0.42	0.30	0.65
			sup	**0.44**	**0.44**	0.47	0.50	0.44	0.65
		Funct.	L2	**0.66**	**0.66**	0.68	0.83	0.59	1.26
			sup	**0.85**	**0.85**	0.86	0.94	0.81	1.21
FAR	ε_1		L2	**1.00**	1.01	1.01	1.13	0.85	2.20
			sup	**1.00**	**1.00**	**1.00**	1.04	0.91	1.45
	ε_2		L2	**0.96**	**0.96**	0.97	1.08	0.81	1.96
			sup	**0.98**	**0.98**	**0.98**	1.02	0.89	1.39

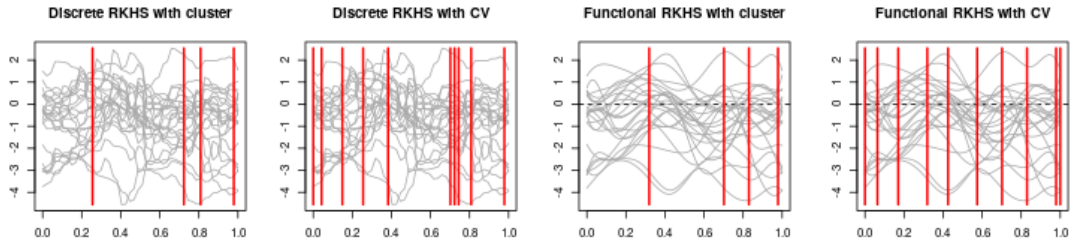
Table 3.7: Errors for the simulated data sets (ε_1 and ε_2 errors of Eq. (3.59))

In Table 3.8 we summarize the different error measurements for the four real data sets tested. Taking these results into account, it is even less clear which implementation of

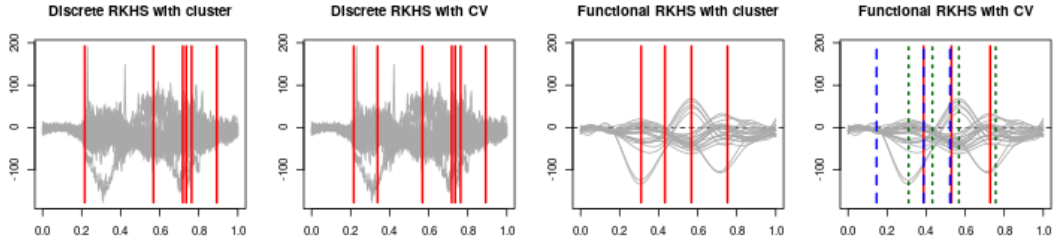
our proposal, the cross-validation one or the cluster one, is the best choice. For the two first data sets it seems that the fPCA approach with fFPE slightly outperforms the other methods. However, the differences between it and our proposals are in general small, even achieving the same error, or improving it, in about half of the measures. By comparison, our proposal is the winner for the last two data sets. It is particularly noteworthy the differences obtained for the temperature data set, which is the smallest one with only 32 curves for training in each window. The error measurements of our proposals for this set fall in the interval $[0.24, 0.82]$, while the measurements for fFPE are in $[1.95, 5.3]$. This could be due to the simplicity of our proposal, which just relies on the computation of the covariance matrix of (at most) 10 real random variables. This simplicity is also reflected in the execution time presented later.

In addition, for these real data sets we have also obtained the selected points, which are shown in Figure 3.9 (these curves are centered versions of the ones in Figure 3.8). It is difficult to reach meaningful conclusions for the four implementations altogether, but we can make a couple of interesting observations. For instance, the points selected for the discrete data sets are more "precise" (in some sense) than the ones for the functional version, which look more equispaced. This could lead to think that we are "dispersing" the dependence of the data when representing them on a functional basis. In addition, the points selected with cluster and cross-validation for the discrete sets are similar, although it seems that the cross-validation implementation selects more points than needed (in view of the prediction performance).

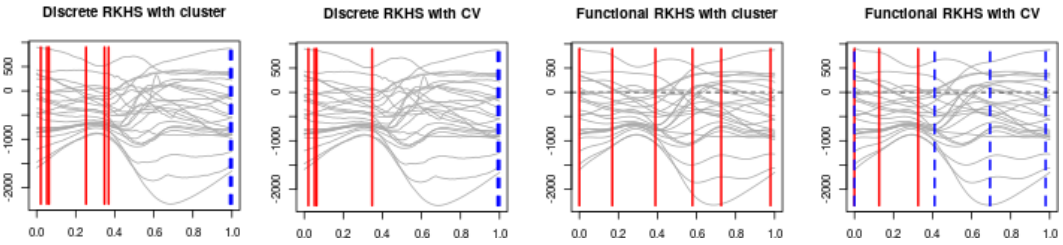
We can also analyze the points selected for each data set separately. We mainly focus on the points selected for the discrete versions of the data. For the pollution data set, it seems that the last few hours of the day are the most informative when predicting the pollution of the following day, which seems reasonable. But it seems also important to measure the pollution early in the morning (since all the methods select at least one point in the interval $[0.2, 0.4]$, which would correspond to between 5:00 and 9:00). With regard to the traffic data, we can identify the most relevant time interval around 17:00, which coincide with one of the moments of greatest traffic volume. All the methods select one point around 13:00 as well, which correspond to the local minimum in the original curves. For the temperature it looks like almost the only relevant hours for prediction purposes are between 0:00 and 2:00 of the previous day (since the blue points correspond to X_{n-2}), along with around 8:00 in the morning. We find this result remarkable, since it is not completely intuitive. Finally, for the utility data set the most noteworthy fact is that the cluster implementation for the discrete version does not select any point for X_{n-1}, which would mean that the dependence lies more backward in time.



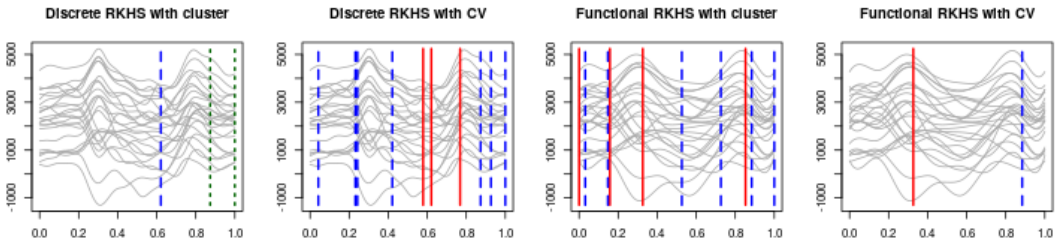
(a) Selected points for PM10.



(b) Selected points for traffic.



(c) Selected points for temperature.



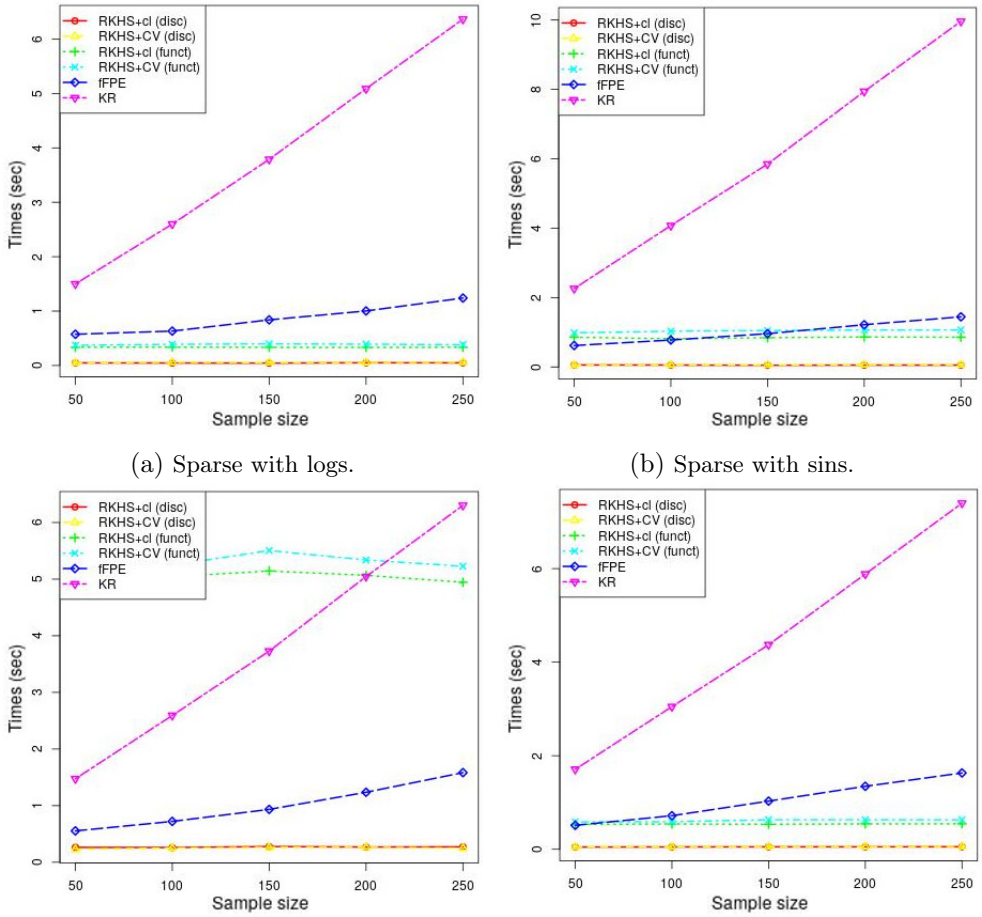
(d) Selected points for utility.

Figure 3.9: Selected points in X_{n-1} (solid red), X_{n-2} (dashed blue) and X_{n-3} (dotted green).

Execution time results

We measured the execution times of all the previous forecasting experiments. Table 3.9 shows the mean execution times of each of the methods for the real data sets. Both the functional (funct) and the discrete (disc) implementations of our proposal are measured. It seems that working with the transformed functional data is slower in general, and that our two discrete implementations are considerably faster than the other methods. The traffic data set is the only one for which our proposal is not the fastest one. This is due to the larger size of the grid, since the curves are sampled every five minutes. Since our procedure checks almost all the points of the grid at each step, the grid size notably affects the execution time.

We also measured how the sample size affects the execution time, increasing it from 50 to 250 observations for the four simulated data sets. The obtained results are available in Table 3.10 and they are summarized in Figure 3.10. We see that the effect of increasing the sample size in our two discrete implementations is almost negligible in comparison with the effect in the remaining methods. The execution times for the functional implementations are also almost constant with the sample size, although we can see that for the O.U. the execution times are quite high. This is due to the use of the model FCAR-sparse(3) for this data set instead of FCAR-sparse(1). Therefore, we analyzed also the impact of the order of the model on the execution time. We use the values $q = 1, \ldots, 5$ for this same data set. The results are summarized in Table 3.11. We can see that the value of this parameter significantly affects the execution time of the functional implementations.



(a) Sparse with logs.

(b) Sparse with sins.

(c) O.U.

(d) FAR.

Figure 3.10: Execution times for the simulated data sets when increasing the sample size.

				RKHS+cl	RKHS+CV	fFPE	KR	Naive
PM10	ε_1 error	Disc.	L2	0.97	**0.74**	0.82	1.48	1.65
			sup	0.92	**0.86**	**0.86**	1.06	1.15
		Func.	L2	**0.70**	0.72	0.82	1.59	1.68
			sup	**0.84**	**0.84**	0.85	1.08	1.11
	ε_2 error	Disc.	L2	0.56	**0.47**	0.50	0.86	0.80
			sup	0.85	0.81	**0.80**	0.98	1.02
		Func.	L2	**0.48**	**0.48**	**0.48**	0.85	0.76
			sup	0.79	0.79	**0.78**	0.99	0.97
Traffic	ε_1 error	Disc.	L2	**1.00**	1.02	**1.00**	1.04	67.48
			sup	1.02	**1.01**	**1.01**	**1.01**	4.70
		Func.	L2	1.37	**1.21**	1.49	1.42	240.57
			sup	1.04	**0.98**	1.05	1.11	10.22
	ε_2 error	Disc.	L2	0.90	0.89	**0.83**	0.95	40.57
			sup	0.99	**0.98**	**0.98**	0.99	4.26
		Func.	L2	0.83	0.80	**0.75**	0.92	62.07
			sup	0.91	**0.89**	0.90	1.00	6.82
Temp	ε_1 error	Disc.	L2	**0.45**	0.53	5.28	1.11	37.78
			sup	**0.66**	0.67	2.10	1.08	3.55
		Func.	L2	**0.63**	0.82	5.30	1.12	38.20
			sup	**0.66**	0.72	2.11	1.07	3.54
	ε_2 error	Disc.	L2	0.25	**0.24**	2.87	1.03	14.25
			sup	0.59	**0.54**	1.95	1.05	2.98
		Func.	L2	**0.37**	**0.37**	2.85	1.03	14.23
			sup	**0.58**	0.60	1.96	1.04	2.98
Utility	ε_1 error	Disc.	L2	0.11	**0.09**	0.10	1.16	18.72
			sup	0.35	**0.34**	**0.34**	1.02	3.33
		Func.	L2	0.09	**0.08**	0.09	1.17	18.94
			sup	**0.29**	0.30	0.31	1.01	3.37
	ε_2 error	Disc.	L2	0.08	**0.07**	**0.07**	0.93	15.08
			sup	0.33	**0.32**	**0.32**	0.98	3.19
		Func.	L2	**0.06**	**0.06**	**0.06**	0.92	15.16
			sup	**0.28**	**0.28**	0.30	0.98	3.24

Table 3.8: Errors for real data sets (ε_1 and ε_2 of Eq. (3.59))

	RKHS+cl (disc)	RKHS+CV (disc)	RKHS+cl (funct)	RKHS+CV (funct)	fFPE	KR
PM10	**0.09**	**0.09**	0.89	1.08	0.71	2.62
Traffic	2.13	2.14	32.66	32.06	**0.52**	1.04
Temp	0.27	**0.20**	3.49	3.80	0.37	0.51
Utility	**0.23**	**0.23**	4.38	4.05	0.56	1.02

Table 3.9: Execution times (secs) for the real data sets.

	n	RKHS+cl (disc)	RKHS+CV (disc)	RKHS+cl (funct)	RKHS+CV (funct)	fFPE	KR
Sparse with log.	50	**0.05**	**0.05**	0.34	0.37	0.57	1.50
	100	**0.05**	0.06	0.34	0.39	0.63	2.60
	150	**0.04**	0.05	0.33	0.40	0.84	3.79
	200	**0.05**	**0.05**	0.33	0.39	1.00	5.09
	250	**0.05**	0.06	0.34	0.38	1.24	6.37
Sparse with sins	50	0.06	**0.05**	0.85	0.98	0.62	2.26
	100	**0.05**	0.06	0.81	1.03	0.78	4.07
	150	**0.05**	0.06	0.84	1.05	0.96	5.85
	200	**0.05**	0.06	0.86	1.06	1.22	7.94
	250	**0.05**	0.06	0.86	1.07	1.44	9.96
O.U	50	0.26	**0.23**	5.07	5.36	0.55	1.47
	100	0.26	**0.25**	5.04	5.25	0.72	2.59
	150	0.28	**0.26**	5.14	5.51	0.93	3.73
	200	**0.27**	**0.27**	5.07	5.34	1.24	5.05
	250	0.27	**0.25**	4.94	5.23	1.58	6.30
FAR	50	**0.05**	**0.05**	0.53	0.59	0.51	1.71
	100	**0.05**	**0.05**	0.54	0.59	0.72	3.05
	150	**0.05**	**0.05**	0.53	0.63	1.03	4.37
	200	**0.05**	**0.05**	0.54	0.63	1.35	5.88
	250	**0.05**	0.06	0.54	0.63	1.63	7.40

Table 3.10: Execution times (secs) for the simulated data sets.

q	RKHS+cl (funct)	RKHS+CV (funct)
1	0.36	0.38
2	1.71	1.82
3	5.04	5.16
4	8.86	9.07
5	13.92	14.25

Table 3.11: Execution times (secs) for FCAR-sparse(q) with functional implementation.

Chapter 4

Functional logistic regression

Throughout this chapter we consider the problem of defining a suitable extension of the classical logistic regression model. The idea behind logistic regression already appeared at the end of nineteenth century (a complete historical overview can be found in Cramer (2003, Ch. 9)) and became quite popular since then. In spite of the name "regression", this technique is often used for binary classification problems. An main advantage of the logistic regression model in comparison with other standard classifiers is that it provides estimations of the probabilities of belonging to each class. Hilbe (2009) is a rather complete book about this technique.

The logistic model is a particular case of the wider family of generalized linear models (we refer to McCullagh and Nelder (1989) for details) which presents some interesting characteristics. According to Hosmer et al. (2013, p. 52), one of its most appealing features is that the coefficients of the model are easily interpretable in terms of the values of the predictors. This technique stems from the attempt to apply well-known linear regression procedures to problems with categorical responses, like binary classification, or non-Gaussian distributions. There is no point in imposing that the categorical response is linear in the predictors x, but there are no objections with assuming that $\log(p(x)/(1-p(x))$ is linear in x (where $p(x)$ is the probability of class 1 given x). Different link functions can be used instead of the logarithm of this quotient. However, an important aspect of this particular model is that it holds whenever the predictors under both classes are Gaussian random variables with common covariance matrix.

This finite-dimensional model has been widely studied. Apart from the already mentioned references, Efron (1975) provides a comparison between logistic predictors and Fisher discriminant analysis under Gaussianity of the predictors. In addition, Munsiwamy and Wakweya (2011) gives a quite user-friendly overview of asymptotic results of the estimators (firstly proved in Fahrmeir and Kaufmann (1985) and Fahrmeir and Kaufmann (1986)).

The motivations for extending logistic regression to functional data are quite obvious. An historical overview of several approaches to functional logistic regression can by

found in Mousavi and Sørensen (2018). We start by establishing the framework of the problem in this functional context. The goal is to explore the relationship between a dichotomous response variable $Y = \{0, 1\}$ and a functional predictor in $L^2[0, 1]$. These functions are trajectories drawn from a L^2-stochastic process X with covariance function K. The random variable Y conditioned to the realizations x of the process follows a Bernoulli distribution with parameter $p(x)$ and the prior probability of class 1 is denoted by $p = \mathbb{P}(Y = 1)$. In this setting, the common functional logistic regression (FLR) model is

$$\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + \exp\{-\beta_0 - \langle \beta, x\rangle_2\}}, \tag{4.1}$$

where $\beta_0 \in \mathbb{R}$, $\beta \in L^2[0, 1]$ and $\langle \cdot, \cdot \rangle_2$ denotes the inner product in $L^2[0, 1]$. This model is the direct extension of the d-dimensional one, where the product in $\mathbb{R}^d$ is replaced by its functional counterpart.

The standard approach to this problem is to reduce the dimension of the curves using PCA. That is, the curves are projected into the first d eigenfunctions of the covariance operator $\mathcal{K}$ (defined in Equation (1.4)) associated with the covariance function K of the process. Then standard d-dimensional logistic regression is applied to the coefficients of these projections. Among others, this strategy has been explored by Escabias et al. (2004) and James (2002) from an applied perspective, where the latter deals with generalized linear models (and not only with logistic regression). These more general models are also studied by Müller and Stadtmüller (2005), but with a more mathematical focus.

Here we propose a novel model based on ideas borrowed from the theory of RKHS's. To be more specific, our proposal is to study the following model instead of (4.1),

$$\mathbb{P}(Y = 1 \,|\, X = x) \;=\; \frac{1}{1 + \exp\{-\beta_0 - \langle \beta, x\rangle_K\}}, \tag{4.2}$$

where the inner product stands for $\Psi_x^{-1}(\beta)$, the inverse of Loève's isometry defined in Equation (1.3). Throughout this chapter we motivate this model and study its main properties. Similarly to the finite-dimensional case, our model holds when the conditional distributions of the process given the two possible values of Y are Gaussian with the same covariance structure. Another interesting property of this new model is that, for some particular choices of the slope function, the model amounts to a finite-dimensional logistic regression model for which the regressors are a finite number of projections of the trajectories of the process. Thus, the impact-point model studied by Lindquist and McKeague (2009) can be seen as a particular case of the RKHS-based model. In general, this provides a mathematical ground to variable selection in logistic regression, which will be the main aim of this chapter. Finally, the model is a real generalization of the finite-dimensional one in the sense that (4.2) coincides with it when the RKHS is that corresponding to the covariance matrix of the regressors.

After defining the model we analyze an interesting behavior of both functional logistic models (the one in Equation (4.1) and the RKHS one), which does not occur in the finite-dimensional case. We give conditions under which the maximum likelihood estimators of the slope function do not exist with probability one. The family of processes for which this happens includes some interesting cases, like the Brownian motion and other related processes. To sort out this difficulty, we propose two sequential maximum likelihood approaches, based on Firth's estimator (e.g. Firth (1993)). The first version is a greedy iterative algorithm inspired by the greedy EM approach of Verbeek et al. (2003), proposed to deal with high-dimensional parameters. The second is merely a simplification of this algorithm. However, we also prove that the dimension of these sequential approximations should be restricted to a finite value. Otherwise, the estimator might not exist asymptotically. This is not really an issue since we are mainly interested in variable selection. Then, it would be unreasonable to allow the number of selected variables to unrestrictedly increase.

In order to asses the performance of this method we compare it with some proposals already existing in the literature for binary classification problems. We use both simulated and real examples in this comparison.

Contents of the chapter

In Section 4.1 we present the RKHS-based model and the maximum likelihood function of the slope parameter. The existence of the maximum likelihood estimator for functional logistic models is carefully analyzed in Section 4.2. The particular proposals implemented in practice are analyzed in Section 4.3. Section 4.4 includes the empirical results.

4.1 RKHS-based functional logistic model

In this section we motivate the reasons why model (4.2) is meaningful. With Theorem 4.1 we prove that the very natural hypothesis that both $X|Y = 0$ and $X|Y = 1$ are Gaussian implies (4.2). We also analyze under which conditions model (4.1) is implied and we clarify the difference between both approaches.

4.1.1 Conditional Gaussian distributions and functional logistic regression

In this functional setting, for $i = 0, 1$, we assume that $\{X(t) : t \in [0, 1]\}$ given $Y = i$ is a Gaussian process with continuous trajectories, continuous mean function m_i and continuous covariance function K (equal for both classes). Let P_0 and P_1 be the probability measures on $C[0, 1]$ induced by the process X under $Y = 0$ and $Y = 1$ respectively. Recall that when m_0 and m_1 both belong to the RKHS, $\mathcal{H}(K)$, corresponding to the

common covariance function K, by Theorem 5A of Parzen (1961a) we have that P_0 and P_1 are mutually absolutely continuous.

Theorem 4.1. *Let P_0, P_1 be as in the previous lines, then*

(a) if $m_0, m_1 \in \mathcal{H}(K)$, then P_0 and P_1 are mutually absolutely continuous and this Gaussian setting entails model (4.2)*,*

$$\mathbb{P}(Y = 1 \,|\, X = x) \;=\; \frac{1}{1 + \exp\{-\beta_0 - \langle \beta, x \rangle_K\}} \;\equiv\; \frac{1}{1 + \exp\left\{-\beta_0 - \Psi_x^{-1}(\beta)\right\}},$$

where Ψ_x is Loève's isometry (Eq. (1.3)*), $\beta := m_1 - m_0$ and $\beta_0 := (\|m_0\|_K^2 - \|m_1\|_K^2)/2 - \log((1-p)/p)$ (with $p = \mathbb{P}(Y = 1)$).*

(b) if $m_1 - m_0 \in \mathcal{K}(L^2) = \{\mathcal{K}(f) : f \in L^2[0,1]\}$, then P_0 and P_1 are mutually absolutely continuous and model (4.1) *holds.*

(c) if $m_1 - m_0 \notin \mathcal{K}(L^2)$ model (4.1) *is never recovered, but different situations are possible. In particular if $m_0 = 0$, $m_1 \in \mathcal{H}(K)$ recovers scenario (a), but if $m_1 \notin \mathcal{H}(K)$, P_0 and P_1 are mutually singular.*

Proof. (a) The conditional probability of the first class can be expressed in terms of the Radon-Nikodym derivative of P_1 with respect to P_0 (see Baíllo et al. (2011, Th.1)) by

$$\mathbb{P}(Y = 1 \,|\, X) \;=\; \frac{p \frac{\mathrm{d}P_1}{\mathrm{d}P_0}(X)}{p \frac{\mathrm{d}P_1}{\mathrm{d}P_0}(X) + (1-p)} \;=\; \left(1 + \frac{1-p}{p} \frac{\mathrm{d}P_0}{\mathrm{d}P_1}(X)\right)^{-1}. \tag{4.3}$$

Now, let P_G be the measure induced by a Gaussian process with covariance function K but zero mean function, $m \equiv 0$. According to Theorem 7A in Parzen (1961b) (or again Theorem 5A of Parzen (1961a)), these Radon-Nikodym derivatives can be expressed as in Equation (1.9). That is, in this case

$$\frac{\mathrm{d}P_i}{\mathrm{d}P_G}(X) = \exp\left\{\langle X, m_i \rangle_K - \frac{1}{2}\|m_i\|_K^2\right\} \quad i = 0, 1,$$

where $\langle X, m_i \rangle_K$ stands for the inverse of the Loève's isometry $\Psi_X^{-1}(m_i)$. From the last two displayed equations (and using the chain rule for Radon-Nikodym densities), one can rewrite

$$\mathbb{P}(Y = 1 \,|\, X) = \left(1 + \frac{1-p}{p} \exp\left\{\langle X, m_0 - m_1 \rangle_K - \frac{\|m_0\|_K^2 - \|m_1\|_K^2}{2}\right\}\right)^{-1}.$$

Then, rewriting this probability we obtain the logistic model of Equation (4.2).

(b) In this setting, Theorem 6.1 in Rao and Varadarajan (1963) gives the following expression:

$$\log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(x)\right) = \langle x - m_0,\, \mathcal{K}^{-1}(m_1 - m_0) \rangle_2 - \frac{1}{2} \langle m_1 - m_0,\, \mathcal{K}^{-1}(m_1 - m_0) \rangle_2, \tag{4.4}$$

for $x \in L^2[0,1]$. Using Equation (4.3), it is easy to see that the L_2 functional logistic regression model holds.

(c) Also as a consequence of Theorem 6.1 in Rao and Varadarajan (1963), if $m_1 - m_0 \notin \mathcal{K}(L^2)$ it is not possible to express the Radon-Nikodym derivative in terms of inner products in L_2 or, equivalently, there is not a continuous linear functional $L(x)$ and $c \in \mathbb{R}$ such that $\log(\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(x)) = L(x) + c$. The last sentence of the statement is a consequence of Theorem 5A of Parzen (1961a). □

Remark. Part (b) of this theorem has been recently observed by Petrovich et al. (2018), see Theorem 1. Incidentally in that paper, the last sentence of the theorem is wrong. It is not true that P_1 and P_0 must necessarily be orthogonal if $m_1 - m_0 \notin \mathcal{K}(L^2)$, as the discussion of Section 1.2.2 shows.

From part (c) of the theorem follows that RKHS functional logistic regression can be seen as a generalization of the usual L_2 functional logistic regression, in the sense that the usual L_2 is recovered when a higher degree of smoothness on the mean functions is imposed (recall that $\mathcal{H}(K) = \{\mathcal{K}^{1/2}(f) : f \in L^2[0,1]\}$, Eq. (1.5), so $\mathcal{K}(L^2) \subset \mathcal{H}(K)$). Indeed, the functions in $\mathcal{K}(L^2)$ are convolutions of the functions in $L^2[0,1]$ with the covariance function of the process. The discussion of the next section makes clear that this difference is of key importance in practice and not merely a technicality.

It is a well-known fact that in the finite dimensional case, although the logistic model is recovered under Gaussianity, it is more general. Clearly it is also the case for the model in (4.2). Besides, when $\mathcal{H}(K)$ is the RKHS corresponding to a covariance matrix, this model coincides with the usual finite-dimensional logistic model. That is, it seems a natural extension to functional data of the model. This connection between the functional model and the finite-dimensional one is even deeper, as we see in the following section.

Remark. Let put ourselves in the functional classification setting introduced in Section 1.2.2; m_0 equals zero and λ_j, e_j are the eigenvalues and eigenfunctions of $\mathcal{K}$. If we denote $x_j = \langle x, e_j \rangle_2$ and $\mu_j = \langle m_1, e_j \rangle_2$, then (4.4) reduces to

$$\log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_0}(x)\right) \;=\; \langle x, \mathcal{K}^{-1} m_1 \rangle_2 - \frac{1}{2}\langle m_1, \mathcal{K}^{-1} m_1 \rangle_2 \;=\; \sum_{j=1}^{\infty} \frac{x_j \mu_j}{\lambda_j} - \frac{1}{2}\sum_{j=1}^{\infty} \frac{\mu_j^2}{\lambda_j},$$

which corresponds to the centroid optimal classifier derived by Delaigle et al. (2012) if x is classified to class 1 when $(\mathrm{d}P_1/\mathrm{d}P_0)(x) > (1-p)p^{-1}$. Besides, the Bayes error of the problem would be the one of Equation (1.10), presented in Section 1.2.2. Whenever both classes have the same prior probability, this optimal error coincides with the value $\Phi(-\sqrt{-\beta_0/2})$, where Φ is the cumulative distribution function of a standard Gaussian variable.

4.1.2 Finite RKHS model and variable selection

Dimension reduction in the functional logistic regression model may be often appropriate in terms of interpretability of the model and classification accuracy. This reduction must be done losing as little information as possible. We propose to perform variable selection, as we did in Chapter 3. By variable selection we mean to replace each curve x_i by the finite-dimensional vector $(x_i(t_1), \ldots, x_i(t_p))$, for some $t_1, \ldots, t_p$ chosen in an optimal way. In this section we analyze under which conditions it is possible to perform functional variable selection, which is only feasible under the RKHS-model. In the following section we asses how to do it: integrating the points $t_1, \ldots, t_p$ to the estimation procedure as additional parameters (in particular to the modified Maximum Likelihood estimator we propose).

Whenever the slope function β has the form

$$\beta(\cdot) = \sum_{j=1}^{p} \beta_j K(t_j, \cdot), \tag{4.5}$$

the model in (4.2) is reduced to the finite-dimensional one,

$$\mathbb{P}(Y = 1|X) = \left(1 + \exp\Big\{ - \beta_0 - \sum_{j=1}^{p} \beta_j X(t_j)\Big\}\right)^{-1}. \tag{4.6}$$

The main difference between the standard finite-dimensional model and this one is that now the proper choice of the points $T = (t_1, \ldots, t_p) \in [0,1]^p$ is part of the estimation procedure. This fact leads to a critical difference between the functional and multivariate problems, as we will see in Section 4.2. Then, our aim is to approximate the general model described by Equation (4.2) with finite-dimensional models as those of Equation (4.6). This amounts to get an approximation of the slope function in terms of a finite linear combination of kernel evaluations $K(t_j, \cdot)$. This model, for $p = 1$ and a particular type of Gaussian processes X, is analyzed in Lindquist and McKeague (2009).

From the discussion above, it is clear that the differences between the RKHS model and the L_2 one are not minor technical questions. The functions of type $\beta(\cdot) = K(\cdot, t)$ belong to $\mathcal{H}(K)$ but do not belong to $\mathcal{K}(L^2)$. This fact implies that within the setting of the RKHS model it is possible to regress Y on any finite dimensional projection of X, whereas this does not make sense if we consider the L_2 model. This feature is clearly relevant if one wishes to analyze properties of variable selection methods.

4.1.3 Maximum Likelihood estimation

The most common way to estimate the slope function in logistic models is to use the maximum likelihood estimator (MLE). In order to apply this technique, we need to

derive the likelihood function. Let assume that $\{X(s), s \in [0,1]\}$ follows the RKHS logistic model described in Equation (4.2). That is,

$$\beta_0 + \Psi_X^{-1}(\beta) \;\equiv\; \beta_0 + \langle X, \beta \rangle_K \;=\; \log\Big(\frac{p_{\beta,\beta_0}(X)}{1 - p_{\beta,\beta_0}(X)}\Big),$$

where $p_{\beta,\beta_0}(X) = \mathbb{P}(Y = 1|X, \beta, \beta_0)$, $\beta_0 \in \mathbb{R}$ and $\beta \in \mathcal{H}(K)$. The random element $(X(\cdot), Y)$ takes values in the space $Z = L^2[0,1] \times \{0,1\}$, which is a measurable space with measure $\mathrm{d}z = P_X \times \mu$, where P_X is the distribution induced by the process X and μ is the counting measure on $\{0,1\}$. We can define in Z the measure $P_{(X,Y);\beta,\beta_0}$, the joint probability induced by $(X(\cdot), Y)$ for a given slope function β and an intercept β_0. Then we define,

$$\begin{aligned} f_{\beta,\beta_0}(x,y) &= \frac{\mathrm{d}P_{(X,Y);\beta,\beta_0}}{\mathrm{d}z}(x,y) \;=\; \frac{\mathrm{d}(P_{(Y|X);\beta,\beta_0}P_X)}{\mathrm{d}(\mu \times P_X)}(x,y) \\ &= \frac{\mathrm{d}\big(P_{(Y|X);\beta,\beta_0}(x,y)P_X(x)\big)}{\mathrm{d}(\mu(y) \times P_X(x))} \;=\; \frac{\mathrm{d}P_{(Y|X);\beta,\beta_0}(x,y)}{\mathrm{d}\mu(y)}\,\frac{\mathrm{d}P_X}{\mathrm{d}P_X}(x) \\ &= f_{\beta,\beta_0}(y|x) \;=\; \Big(\frac{1}{1+e^{-\beta_0-\langle\beta,x\rangle_K}}\Big)^y \Big(\frac{e^{-\beta_0-\langle\beta,x\rangle_K}}{1+e^{-\beta_0-\langle\beta,x\rangle_K}}\Big)^{1-y}. \end{aligned}$$

Given this density function, the log-likelihood function for a given sample $(x_1^i, y_1^i), \ldots, (x_{n_i}^i, y_{n_i}^i)$ in $L^2[0,1] \times \{0,1\}$, $i = 0, 1$, is

$$L_n(\beta,\beta_0) \;=\; \frac{1}{n_0}\sum_{i=1}^{n_0} \log\left(\frac{e^{-\beta_0-\langle\beta,x_i^0\rangle_K}}{1+e^{-\beta_0-\langle\beta,x_i^0\rangle_K}}\right) + \frac{1}{n_1}\sum_{i=1}^{n_1} \log\left(\frac{1}{1+e^{-\beta_0-\langle\beta,x_i^1\rangle_K}}\right). \quad (4.7)$$

The maximum log-likelihood estimator is the pair $(\widehat{\beta}, \widehat{\beta}_0)$ that maximizes this function L_n. In order to study the asymptotic properties of this estimator, one needs to define the expected log-likelihood function,

$$L(\beta,\beta_0) \;=\; \mathbb{E}_Z\left[\log f(X,Y,\beta,\beta_0)\right] \;=\; \mathbb{E}_Z\left[\log\Big(p_{\beta,\beta_0}(X)^Y\,(1-p_{\beta,\beta_0}(X))^{1-Y}\Big)\right], \quad (4.8)$$

where $\mathbb{E}_Z[\cdot]$ denotes the expectation with respect to the measure $\mathrm{d}z$ and $p_{\beta,\beta_0}(X)$ stands for $(1+\exp(-\beta_0-\Psi_X(\beta)))^{-1}$. When one uses maximum likelihood (ML), it is standard to prove that the true parameters that define the model are a maximum of this expected likelihood function. We prove that this is also the case in this setting.

Proposition 4.2. *The parameters $\beta^* \in \mathcal{H}(K)$ and $\beta_0^* \in \mathbb{R}$ that define the probability of class one are the unique maximum in $\mathcal{H}(K) \times \mathbb{R}$ of the expected log-likelihood function $L(\beta,\beta_0)$ of Eq. (4.8).*

Proof. If $\beta^* \in \mathcal{H}(K)$ and $\beta_0^* \in \mathbb{R}$ are the true slope function and intercept of the model, one can rewrite the likelihood function, evaluated in another $\beta \in \mathcal{H}(K), \beta_0$, as

$$\begin{aligned} L(\beta, \beta_0) &= \mathbb{E}_X\left[\mathbb{E}_Y[\log f(X, Y, \beta, \beta_0) \mid X, \beta, \beta_0]\right] \\ &= \mathbb{E}_X\left[p_{\beta^*,\beta_0^*}(X) \log\left(p_{\beta,\beta_0}(X)\right) + \left(1 - p_{\beta^*,\beta_0^*}(X)\right) \log\left(1 - p_{\beta,\beta_0}(X)\right)\right], \end{aligned}$$

where $\mathbb{E}_X[\cdot]$ and $\mathbb{E}_Y[\cdot]$ are the expectations with respect to P_X and μ respectively.

Now the fact that β^* is a maximum of $L(\beta, \beta_0)$ is straightforward, just following the same reasoning as for the multiple logistic regression to check that $L(\beta, \beta_0) - L(\beta^*, \beta_0^*)$ is always less of equal zero. If there is another β, β_0 that maximizes this function,

$$L(\beta^*, \beta_0^*) - L(\beta, \beta_0) = \mathbb{E}_X\left[p_{\beta^*,\beta_0^*}(X) \log \frac{p_{\beta^*,\beta_0^*}(X)}{p_{\beta,\beta_0}(X)} + (1 - p_{\beta^*,\beta_0^*}(X)) \log \frac{1 - p_{\beta^*,\beta_0^*}(X)}{1 - p_{\beta,\beta_0}(X)}\right]$$

equals zero. Given $0 < x, y < 1$ real numbers, the function of the integrand $x \log(x/y) + (1-x) \log((1-x)/(1-y))$ is always less or equal zero, and the inequality is strict unless $x = y$. Therefore $p_{\beta^*,\beta_0^*}(X) = p_{\beta,\beta_0}(X)$ with probability one (if not, the expectation would be positive over the set of positive measure where $p_{\beta^*,\beta_0^*}(X) \neq p_{\beta,\beta_0}(X)$). Since the logistic function is injective, $\beta_0^* + \Psi_X(\beta^*) = \beta_0 + \Psi_X(\beta)$. Both $\Psi_X(\beta^*)$ and $\Psi_X(\beta)$ are random variables with zero mean, so β_0 must coincide with β_0^*. Therefore, β agrees with β^* in $\mathcal{H}(K)$, since Loève's isometry is also injective. □

4.2 On the non-existence of MLE in logistic models

In the finite-dimensional setting, it is well-known that the use of ML is not suitable when there exists an hyperplane separating the observations of the two classes. This fact, which is presented in detail next, becomes dramatically worst for functional data:

- For a wide class of process (including the Brownian motion), the MLE never exists.
- Under less restrictive conditions, but still in the Gaussian case, the probability of non-existence of the MLE tends to one.

4.2.1 A brief overview of the finite dimensional case

Despite the fact that the maximum likelihood estimation of the slope function for multiple logistic regression is widely used, it has an important issue that is sometimes overlooked. Given a sample $x_i^0 \in \mathbb{R}^d$ for $i = 1, \ldots, n_0$ drawn from population zero and

another sample $x_i^1 \in \mathbb{R}^d$ for $i = 1, \ldots, n_1$ drawn from population one, the classical MLE in logistic regression is the vector $(b_0, b) \in \mathbb{R} \times \mathbb{R}^d$ that maximizes the log-likelihood

$$L_n(b, b_0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \log\Big(\frac{e^{-b_0 - b' x_i^0}}{1 + e^{-b_0 - b' x_i^0}}\Big) + \frac{1}{n_1} \sum_{i=1}^{n_1} \log\Big(\frac{1}{1 + e^{-b_0 - b' x_i^1}}\Big).$$

The existence and uniqueness of such a maximum was carefully studied by Albert and Anderson (1984) (and previously by Silvapulle (1981) and Gourieroux and Monfort (1981)). As stated in Theorem 1 of Albert and Anderson (1984), the latter expression can be made arbitrarily close to zero (note that the log-likelihood is always negative) whenever the samples of the two populations are linearly separable. In that case the maximum can not be attained and then the MLE does not exist (the idea behind the proof is similar to the one of Theorem 4.4 below). There is another scenario where this estimator does not exist; the samples are linearly separable except for some points of both populations that fall into the separation hyperplane (named "quasicomplete separation"). In this case the supremum of the log-likelihood function is strictly less than zero, but it is anyway unattainable.

4.2.2 Non-existence of the MLE in functional settings

When moving from the finite-dimensional model to the functional one, the problem of the non-existence of the MLE is drastically worsened. We will show that, under some conditions, the maximum likelihood estimator for the slope function in the functional logistic regression model does not exist with probability one. We confine ourselves to the RKHS-based model, although the result can be easily extended, with a completely similar method of proof, for the standard L^2 based model of Equation (4.1). This result can be added to the list of conceptual differences between Functional Data Analysis and finite-dimensional statistics.

Recall that, given a sample $(x_i, y_i) \in L^2[0,1] \times \{0,1\}$ of size n, the log-likelihood function is, for $\beta \in \mathcal{H}(K), \beta_0 \in \mathbb{R}$,

$$L_n(\beta, \beta_0) \;=\; \frac{1}{n} \sum_{i=1}^{n} \log\big(p_{\beta,\beta_0}(x_i)^{y_i}\,(1 - p_{\beta,\beta_0}(x_i))^{1-y_i}\big).$$

One of the ways in which the linear separability condition mentioned above can be translated to functional data is presented hereunder.

Assumption 4.3 (SC). The multivariate process $Z(t) = (X_1(t), \ldots, X_n(t))$, $t \in [0,1]$ satisfies the "Sign Choice" (SC) property when for all possible choice of signs $(s_1, \ldots, s_n)$, where s_j is either $+$ or $-$, we have that, with probability one, there exists some $t_0 \in [0,1]$ such that $\text{sign}(X_1(t_0)) = s_1,\ \ldots,\ \text{sign}(X_n(t_0)) = s_n$.

Now, the main result is as follows:

Theorem 4.4. *Let $X(s)$, $s \in [0,1]$, be a L^2 stochastic process with $\mathbb{E}[X(s)] = 0$. Denote by K the corresponding covariance function. Consider a logistic model (4.2) based on $X(s)$. Let $X_1, \ldots, X_n$ be independent copies of X. Assume that the n-dimensional process $Z_n(s) = (X_1(s), \ldots, X_n(s))$ fulfills the SC property. Then, with probability one, the MLE estimator of β (Eq. (4.7)) **does not exist** for any sample size n.*

Proof. Let $x_1(s) \ldots, x_n(s)$ be a random sample drawn from $X(s)$. From the SC assumption there is (with probability 1) one point t_0 such that $x_i(t_0) > 0$ for all i such that $y_i = 1$ (n_1 in total) and $x_i(t_0) < 0$ for those (n_0) indices i with $y_i = 0$. Recall that the sample log-likelihood function given in Equation (4.7) is

$$L_n(\beta, \beta_0) \; = \frac{1}{n_1} \sum_{\{i:\, y_i = 1\}} \log\Big(\frac{1}{1 + e^{-\beta_0 - \langle \beta, x_i \rangle_K}}\Big) + \frac{1}{n_0} \sum_{\{i:\, y_i = 0\}} \log\Big(\frac{e^{-\beta_0 - \langle \beta, x_i \rangle_K}}{1 + e^{-\beta_0 - \langle \beta, x_i \rangle_K}}\Big).$$

Note also that $L_n(\beta, \beta_0) \leq 0$ for all β. Now, take a numerical sequence $c_m \uparrow \infty$ and define

$$\beta_m(\cdot) \; = \; c_m K(t_0, \cdot).$$

Then, since we are identifying $\langle \beta, x_i \rangle_K$ with the inverse of Loève's isometry, for every j such that $y_j = 1$, we have

$$\langle \beta_m, x_j \rangle_K \; = \; c_m x_j(t_0) \; \to \; \infty, \quad \text{as } m \to \infty,$$

since we have taken t_0 such that $x_j(t_0) > 0$ for those indices i with $y_j = 1$. Likewise, $\langle \beta_m, x_j \rangle_K$ goes to $-\infty$ whenever $y_i = 0$ since we have chosen t_0 such that $x_i(t_0) < 0$ for those indices. As a consequence $L_n(\beta_m, 0) \to 0$ as $m \to \infty$. Therefore the likelihood function can be made arbitrarily large so that the MLE does not exist. □

Remark. A non-existence result for the MLE estimator, analogous to that of Theorem 4.4, can be also obtained with a very similar reasoning for the L^2-based logistic model of Equation (4.1). The main difference in the proof would be the construction of β_m which, in the L^2 case, should be obtained as an approximation to the identity (that is, a linear "quasi Dirac delta") centered at the point t_0.

Although it could seem a somewhat restrictive assumption, the following proposition shows that the SC property applies to some important and non-trivial situations.

Proposition 4.5. *(a) The n-dimensional Brownian motion fulfills the SC property.*

(b) The same holds for any other n-dimensional process in $[0,1]$ whose independent marginals have a distribution absolutely continuous with respect to that of the Brownian motion.

Proof. (a) Given the n dimensional Brownian motion $\mathcal{B}_n = (B_1, \ldots, B_n)$, where the B_j are independent copies of the standard Brownian motion $B(t)$, $t \in [0,1]$, take a sequence of signs $(s_1, ..., s_n)$ and define the event

$$A = \{\text{for any given } t \text{ there exists } 0 < t_0 < t \text{ s.t. } \operatorname{sign}(B_j(t_0)) = s_j,\ j = 1, \ldots, n\} \tag{4.9}$$

We may express this event by

$$A = \bigcap_{t \in (0,1] \cap \mathbb{Q}} A_t, \tag{4.10}$$

where, for each $t \in (0,1] \cap \mathbb{Q}$,

$$A_t = \{\text{there exists } t_0 < t \text{ such that } \operatorname{sign}(B_j(t_0)) = s_j,\ j = 1, \ldots, n\}$$

Now, the result follows directly from Blumenthal's 0-1 Law for n-dimensional Brownian processes (see, e.g., Mörters and Peres (2010, p. 38)). Such result establishes that for any event $A \in \mathcal{F}^+(0)$ we have either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. Here $\mathcal{F}^+(0)$ denotes the *germ σ-algebra* of events depending only on the values of $\mathcal{B}_n(t)$ where t is in an arbitrarily small interval on the right of 0. More precisely,

$$\mathcal{F}^+(0) = \bigcap_{t>0} \mathcal{F}^0(t), \text{ where } \mathcal{F}^0(t) = \sigma(\mathcal{B}_n(s), 0 \leq s \leq t).$$

From (4.9) and (4.10) it is clear that the above defined event A belongs to the germ σ-algebra $\mathcal{F}^+(0)$. However, we cannot have $\mathbb{P}(A) = 0$ since (from the symmetry of the Brownian motion) for any given t_0 the probability of $\operatorname{sign}(B_j(t_0)) = s_j,\ j = 1, \ldots, n$ is $1/2^n$. So, we conclude $\mathbb{P}(A) = 1$ as desired.

(b) If $X(t)$ is another process whose distribution is absolutely continuous with respect to that of the n-dimensional Brownian motion $\mathcal{B}_n$, then the set A, defined by (4.9) and (4.10) in terms of $\mathcal{B}_n$ has also probability one when it is defined in terms of the process $X(t)$. Recall that, from the definition of absolute continuity, $\mathbb{P}(\mathcal{B}_n \in A^c) = 0$ implies $\mathbb{P}(X \in A^c) = 0$ and therefore $\mathbb{P}(X \in A) = 1$. □

Remark. Following the comment in Mörters and Peres (2010) about processes with strong Markov property, this result based on RKHS theory may be extended for Lévy processes whenever the covariance function was continuous (like Poisson process in the real line). However, apart from the Brownian motion, this type of processes have discontinuous trajectories, and this situation is not considered in this work.

This property would be the functional counterpart of having a finite-dimensional problem where the supports of both classes (0 and 1) are linearly separable. However, in a similar sense as with the "near perfect classification" phenomenon introduced in Section 1.2.2, this separability issue does not only appear in degenerate problems in the functional setting.

In practice this problem would rarely be encountered, since the curves are usually provided in a discretized fashion. Nevertheless, in the next section we suggest a couple of techniques that completely avoid the problem. From a theoretical perspective and in view of Theorem 4.4, it is clear that there is no hope of obtaining a general convergence result of the standard MLE defined by the maximization of function in (4.7). That is, one should define a different estimator or impose some conditions on the process X to avoid the SC property. For instance, Lindquist and McKeague (2009) prove consistency results of the model with a single impact point $\theta \in [0,1]$ for processes $X(t) = Z + B_\theta(t)$, where B_θ is a two-sided Brownian motion centered in θ (i.e. two independent Brownian motions starting at θ and running in opposite directions) and Z is a real random variable independent of B_θ. Then, due to the independence assumption, it is clear that accumulation points (like 0 for the Brownian motion) are avoided.

4.2.3 Asymptotic non-existence for Gaussian processes

In the previous section we have seen that the problem of non-existence of the MLE is dramatically aggravated for functional data, where the regressors $X(s)$ are not fixed in advance. But this is not the only issue with MLE in functional logistic regression. In this section we see that the probability that the MLE does not exist goes to one as the sample size increases, for any Gaussian process satisfying very mild assumptions.

We use the following notation: for $T = \{t_1, \dots, t_p\} \subset [0,1]$ and $f \in L_2[0,1]$, let $f(T) := (f(t_1), \dots, f(t_p))'$ and let Σ_T be the $p \times p$ matrix whose (i,j) entry is given by $K(t_i, t_j)$.

Theorem 4.6. *Let $(x_1, y_1), \dots, (x_n, y_n)$ be a random sample of independent observations satisfying model (4.2). Assume that X is a Gaussian process such that K is continuous and Σ_T is invertible for any finite set $T \subset (0,1)$. It holds*

$$\lim_{n\to\infty} \mathbb{P}(\textit{MLE exists}) = 0.$$

Proof. Let $\beta^* \in \mathcal{H}_K, \beta_0^*$ be the true values of the parameters. Since $\|\beta^*\|_K < \infty$, we have $h(\beta_0^*, \|\beta^*\|_K) < \infty$, where h is the function defined in Equation (6) of Candès and Sur (2018) (see the remark below). Let p_n be an increasing sequence of natural numbers such that $\lim_{n\to\infty} p_n/n = \kappa > h(\beta_0^*, \|\beta^*\|_K)$. Consider the set of equispaced points $0 < t_1 < t_2 < \cdots < t_{p_n} < 1$ and denote $T_n = \{t_1, \dots, t_{p_n}\}$. Define $\alpha_{T_n} = \Sigma_{T_n}^{-1} \beta^*(T_n)$. Now, consider the following sequence of finite-dimensional logistic regression models

$$\mathbb{P}(Y = 1 \,|\, X) = \frac{1}{1 + \exp\left\{-\beta_0^* - \alpha_{T_n}' X(T_n)\right\}},$$

and the following sequence of events

$$E_n = \{\text{There exists } \alpha \in \mathbb{R}^{p_n} : \alpha' x_i(T_n) \geq 0, \text{ if } y_i = 1; \;\; \alpha' x_i(T_n) \leq 0, \text{ if } y_i = 0\}.$$

Recall that the event E_n amounts to non-existence of MLE for finite-dimensional logistic regression models (see Albert and Anderson (1984)).

Now let us prove the validity of condition (3) in Candès and Sur (2018), which is required for the validity of Theorem 1 in that paper. In our case, such condition amounts to

$$\lim_{n\to\infty} \text{var}\big(\alpha'_{T_n} X(T_n)\big) = \lim_{n\to\infty} \alpha'_{T_n} \Sigma_{T_n} \alpha_{T_n} = \|\beta^*\|_K^2,$$

but this directly follows from Theorem 6E of Parzen (1959). Since $\lim_{n\to\infty} p_n/n = \kappa > h(\beta_0^*, \|\beta^*\|_K^2)$ we apply Theorem 1 in Candès and Sur (2018) to deduce $\lim_{n\to\infty} \mathrm{P}(E_n) = 1$.

Now we define the auxiliary sequence of events

$$\widetilde{E}_n = \{\text{There exists } \alpha \in \mathbb{R}^{p_n} : \alpha' x_i(T_n) > 0, \text{ if } y_i = 1; \quad \alpha' x_i(T_n) < 0, \text{ if } y_i = 0\},$$

with strict inequalities. Assume that $\widetilde{E}_n$ happens so that there exists a separating hyperplane defined by $\alpha \in \mathbb{R}^{p_n}$. Then, in the same spirit as in the proof of Theorem 4.4, it is possible to show that if $\hat{\beta}_{m,n} = m \sum_{j=1}^{p_n} \alpha_j K(\cdot, t_j) \in \mathcal{H}_K$, then $\lim_{m\to\infty} L(\hat{\beta}_{m,n}, 0) = 0$, where $L(\beta, \beta_0)$ is the log-likelihood function. As a consequence, for all n, if $\widetilde{E}_n$ happens, then the MLE for the RKHS functional logistic regression model does not exist. The result follows from the fact that $\mathbb{P}(E_n) = \mathbb{P}(\widetilde{E}_n)$ and the events $\alpha' x_i(T_n) = 0$ have probability zero. Because we are assuming that the process does not have degenerate marginals. □

Remark. Theorem 1 in Candès and Sur (2018) is a remarkable result. It applies to logistic finite-dimensional regression models with a number p of covariables, which is assumed to grow to infinity with the sample size n, in such a way that $p/n \to \kappa$. Of course, the sample is given by data (x_i, y_i), $i = 1, \ldots, n$. Essentially the result establishes that there is a critical value such that, if κ is smaller than such critical value, one has $\lim_{n,p\to\infty} \mathbb{P}(\text{MLE exists}) = 1$; otherwise we have $\lim_{n,p\to\infty} \mathbb{P}(\text{MLE exists}) = 0$. Such critical value is given in terms of a function h (which is mentioned in the proof of the previous result) whose definition is as follows. Let us use the notation $(\widetilde{Y}, V) \sim F_{\beta_0,\gamma_0}$ whenever $(\widetilde{Y}, V) \stackrel{d}{=} (Y, YX)$, for $\widetilde{Y} = 2Y - 1$ (note that, in the notation of Candès and Sur (2018), the model is defined for the case that the variable Y is coded in $\{-1, 1\}$), $\beta_0, \gamma_0 \in \mathbb{R}$, $\gamma_0 \geq 0$ and where $X \sim \mathcal{N}(0, 1)$ and $\mathbb{P}(\widetilde{Y} = 1|X) = (1 + \exp\{-\beta_0 - \gamma_0 X\})^{-1}$. Now, define $h(\beta_0, \gamma_0) = \min_{t_0,t_1\in\mathbb{R}} \mathbb{E}[(t_0\widetilde{Y} + t_1 V - Z)_+^2]$, where $Z \sim \mathcal{N}(0, 1)$ independent of $(\widetilde{Y}, V)$ and $x_+ = \max\{x, 0\}$. Then, Theorem 1 in Candès and Sur (2018) proves that the above mentioned critical value for κ is precisely $h(\beta_0, \gamma_0)$.

4.3 The estimation of β in practice

The problem of non-existence of the MLE can be circumvented if the goal is variable selection. The main idea behind the proof of Theorem 4.6 is that one can approximate

the functional model with finite approximations as those in (4.6) with p increasing as fast as desired. Therefore, if we constrain p to be less than a finite fixed value, Theorem 4.6 does not apply.

In order to sort out the non-existence problem for a given sample (due to the SC property), it would be enough to use a finite-dimensional estimator that is always defined, even for linearly separable samples. As mentioned, an extensive study of existence and uniqueness conditions of the MLE for multiple logistic regression can be found in the paper of Albert and Anderson (1984). We suggest to use *Firth's estimator*, firstly proposed by Firth (1993), which is always finite and unique.

The author initially proposed this approach to reduce the bias of the standard MLE, but afterwards it was used to obtain ML estimators for linearly separable samples (see e.g. Heinze and Schemper (2002), where the technique is presented in a quite accessible manner). Besides, the reduction of the bias leads to better results in practice. The general idea of Firth's procedure is not to use the original sample responses $(y_1, \ldots, y_n)$ to compute the usual score equations one must solve to compute the MLE, but a modification of them. Each response y_i is split into two new responses $(1 + h_i/2)y_i$ and $(h_i/2)(1 - y_i)$, where the coefficients h_i go to zero with the sample size. The idea behind this duplication is to avoid the problem of the linear separability of the sample (previously discussed), which leads to the non-existence of the MLE. It is easy to see that in the new modified sample with duplicated observations there is always overlapping between the two classes, so ML can be safely applied. The specific coefficients h_i are the diagonal elements of the matrix $W^{1/2}x_T(x_T'Wx_T)^{-1}W^{1/2}$, W being the diagonal matrix with elements $p_{\beta,\beta_0}(x_i)(1 - p_{\beta,\beta_0}(x_i))$ and x_T the matrix whose rows are $(1, x_i(t_1), \ldots, x_i(t_p))$. This estimator is implemented in the "brglm" function of the `brglm` R-package (see Kosmidis (2017)).

With this procedure we obtain estimators $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$. An interesting observation is that the value of the independent term $\widehat{\beta}_0$ can be used to estimate the Bayes error of any homoscedastic Gaussian problem with equiprobable classes ($p = 1/2$), as we discussed at the end of Section 4.1.1.

4.3.1 Greedy "max-max" algorithm

In view of the previous discussion, the objective is to find the points $T = (t_1, \ldots, t_p) \in [0, 1]^p$ and the coefficients $(\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$, for a fixed p, that maximize the log-likelihood associated with model of Equation (4.6).

We propose an iterative algorithm in which the points and the coefficients are updated alternatively. This approach is reminiscent of the well-known Expectation-Maximization (EM) technique, which is typically applied to estimate mixtures in supervised classification (see e.g. Bishop (2016, Chapter 9)). In general, the EM algorithm is used to

compute ML estimators for models with non-observable data. In our setting, these non-observable parameters would correspond to the points $t_1, \ldots, t_p$. Estimating the β_j's given a set of t_j's is straightforward (via Firth's approach), and once the parameters β_j are known, the points t_j can be obtained maximizing the log-likelihood over $[0,1]^p$. The algorithm is as follows: first the coefficients β_j are randomly initialized, and then we iterate the following steps until convergence.

- **(Maximization 1)** Compute the set of points $T \in [0,1]^p$ that maximizes the log-likelihood function for the current set of coefficients $\beta_0, \ldots, \beta_p$.

- **(Maximization 2)** Then, use the just obtained points T to compute the MLE of the model (via Firth's estimator).

It is also suitable to start initializing the points t_j, and then start the iterations in the second step. Besides, in practice the maximization over the continuous set $T \in [0,1]^p$ is unfeasible, so it should be made using some grid. This is not usually an issue, since the sample curves are typically provided in a discretized fashion. Then one could directly search the points in the grid provided by the sample.

However, the proposed algorithm only ensures the convergence to a local maximum, and this maximum strongly depends on the initial (random) points of the algorithm. Besides, the number of local maxima of the likelihood function likely increases with the dimension of the search space. That is, the accuracy of the results might deteriorate when p increases. A possible solution is to replicate the execution several times for the same sample and keep the parameters that give a maximum value of the likelihood. However, this could be computationally expensive depending on the dimension. Then, we suggest to adapt the greedy EM methodology proposed in Verbeek et al. (2003) for Gaussian mixtures. The idea is to start with a model of dimension one ($p = 1$ in Equation (4.6)) to estimate $\widehat{t}_1$. Then, once this first point is fixed, add a second random point $\widetilde{t}_2$ and start again the maximization iterations, now with $p = 2$, using as starting points $(\widehat{t}_1, \widetilde{t}_2)$. The algorithm continues adding points until it reaches the desired dimension p. This approach should work better than simply initializing all the points t_j at random, since only one random point is added in each step. In addition, for small values of p it is more likely to obtain meaningful points, since the likelihood function should have less local maxima.

In practical problems, it is also important to determine how many points p one should retain. The common approach is to fix this value $\widehat{p}$ by cross-validation, whenever it is possible. Another reasonable approach is to stop when the difference between the likelihoods obtained with $p-1$ and p points is less than a threshold ϵ. This was the method used in Chapter 3, where some techniques to determine the value ϵ were addressed.

Once that the parameters are estimated for a given sample, one can also approximate

the β function in $\mathcal{H}(K)$ as

$$\widehat{\beta}(\cdot) = \sum_{j=1}^{\widehat{p}} \widehat{\beta}_j K(\widehat{t}_j, \cdot).$$

When the function K is not known, a feasible alternative is to replace $K(\widehat{t}_j, \cdot)$ with the empirical covariance function $\widehat{K}(\widehat{t}_j, \cdot)$. However, in this case the estimation $\widehat{\beta}$ would not be in $\mathcal{H}(K)$ with probability 1 (since the sample curves X_i used to compute $\widehat{K}$ do not belong to $\mathcal{H}(K)$). A possible solution is to regularize the trajectories before computing $\widehat{K}$, to force them to belong to $\mathcal{H}(K)$ (in the same spirit as in Chapter 2).

4.3.2 Sequential maximum likelihood

The greedy approach we have described above directly suggests another greedy algorithm to compute both the points t_j and the coefficients β_j. The idea is to exchange the execution of the iterative algorithm by the direct maximization of the likelihood function. As in the previous case, we will need also a grid over $[0, 1]$ to search the points t_j. The procedure would be as follows:

1. For each t on the grid, we fit the logistic model of Equation (4.6) with $p = 1$. The log-likelihood achieved for this t at the ML estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is stored in $\ell_1(t)$. Then, the first point $\widehat{t}_1$ is fixed as the point at which $\ell_1(t)$ achieves its maximum value.

2. Once $\widehat{t}_1$ has been selected, for each t in the grid we fit the model

$$\mathbb{P}(Y = 1|X) = \left(1 + \exp\left\{\beta_0 + \beta_1 X(\widehat{t}_1) + \beta_2 X(t_2)\right\}\right)^{-1}.$$

   As in the previous step, $\ell_2(t)$ would be be the likelihood achieved at $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ and $\widehat{t}_2$ the maximum of $\ell_2(t)$.

3. We proceed in the same way until a suitable number of points p has been selected.

The number of points to select and the complete estimation of the β function can be obtained as in the previous proposal.

With this greedy approach the dependence on the random initial points is erased. However, it may be more computationally expensive, specially if the size of the grid is large.

4.4 Simulation study

4.4.1 Binary classification

The typical application of logistic regression is binary classification, where each curve is classified to the class with the highest probability. In order to check the classification performance of our proposals, for a sample $(y_1, x_1), \ldots, (y_n, x_n)$ with $y_i \in \{0, 1\}$, we measure the misclassification rate

$$\frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i|,$$

where $\widehat{y}_i$, $i = 1, \ldots, n$, are the predicted labels. We have measured also the execution times, which are shown at the end of this section.

Methods

The two implementations of our proposal are compared with other recent methods of the literature. Some of these methods are different approaches to the typical functional logistic regression model (4.1), but we include also other general classifiers which have shown good performances lately. The names used for each method in the result tables are shown in bold in brackets. We compare the following procedures:

- The two approaches to the RKHS-based functional logistic model (4.2) presented in the previous section: the one based on greedy maximization-maximization procedure (**RKHS-mm**) and the sequential search (**RKHS-sq**). For the Firth's estimator we use the R function "brglm" of the package `brglm` (Kosmidis (2017)). The number of points included in the model is selected by 5-fold cross-validation.
- The first method presented in Mousavi and Sørensen (2017) (**wav**), which is an adaptation to functional classification of the method proposed in Zhao et al. (2012). The curves are represented in a wavelet basis and the number of coefficients in the representation is shrunk with LASSO. Mousavi and Sørensen (2018) compare three functional logistic proposals and this wavelet based one, presented in Section 2.1 of the paper, seems to be the most competitive out of the three. As suggested in this last paper, we use the modified least asymmetric version of Daubechies wavelets, via the R function "hardThresholding" of the package `RFgroove` (Gregorutti (2016)). The detail level of the basis ("s2" to "s12" of the parameter "wavFilter") is selected by 5-fold cross-validation.
- Functional logistic model (**PCA**) that is a particular case of the "generalized linear model" of Müller and Stadtmüller (2005). The idea is to represent the curves in the base of functional principal components of the process and to apply then a finite logistic model to the coefficients of the curves in this base (this

method can be also found in Section 2.2 of Mousavi and Sørensen (2018)). The number of principal components included in the FLR model is fixed by 5-fold cross-validation. In order to obtain the functional principal components we use the R function "fdata2pc" of the package `fda.usc` (see Febrero-Bande and de la Fuente (2012)). The number of coefficients retained is fixed by 5-fold cross-validation.

- For this method we identify each curve with its coefficients in the principal components base, as before. But instead of applying multiple logistic regression to these coefficients, we use k-nearest neighbors with $k = 5$ (**PCA-knn**). To compute the classifier we use the R function "knn" of the package `class` (Ripley and Venables (2015)).

- The non-parametric regression method proposed in Ferraty and Vieu (2006) adapted to perform binary classification (**nonP**), which is based on the model $Y = r(X) + \varepsilon$ with r unknown. This method uses a Nadaraya–Watson kernel estimator of the conditional expectation to estimate $\widehat{r}$, and then a curve x is classified to class 1 if $\widehat{r}(x) > 0.5$ and 0 otherwise. As suggested by Delaigle and Hall (2012), we use the function "funopare.knn.gcv" of the R-code provided together with the book of Ferraty and Vieu (2006). The specific estimator $\widehat{r}(x)$ for a sample $(y_1, x_1), \ldots, (y_n, x_n)$ is $\sum_{i=1}^{n} y_i \phi(h_5^{-1} d(x_i, x)) / \sum_{i=1}^{n} \phi(h_5^{-1} d(x_i, x))$. The kernel ϕ we use is quadratic, the semimetric d among curves is of PLS type (as described in Section 3.4.2 of the book) and the bandwidth h_5 is selected by 5-nearest neighbors in the sense that $\#\{i : d(x_i, x) < h_5\} = 5$. The number of factors for the PLS-semimetric is fixed by 5-fold cross-validation.

- RK-VS method for variable selection proposed in Berrendero et al. (2017) with two different classifiers. The Gaussian setting introduced at the beginning of Section 4.1.1 can be defined equivalently for regressors in $\mathbb{R}^d$, where $m_0 = (m_0^1, \ldots, m_0^d)$ and $m_1 = (m_1^1, \ldots, m_1^d)$ are the mean vectors of both populations and Σ is the common covariance matrix. As stated in the above mentioned paper (see also Izenman (2008)), the optimal Bayes classifier in this setting is a decreasing function of $(m_1 - m_0)' \Sigma^{-1} (m_1 - m_0)$ (which coincides with the Mahalanobis distance between the mean vectors). Then, the authors propose to select the p points $(t_1, \ldots, t_p)$ that maximize the sample version of this last expression with $m_0 = (m_0(t_1), \ldots, m_0(t_p))$ and $m_1 = (m_1(t_1), \ldots, m_1(t_p))$. Since this variable selection method is independent of any classification technique, we perform two multivariate classification procedures on the selected variables $X(t_1), \ldots, X(t_p)$. We try Linear Discriminant Analysis (**RK**) and k-nearest neighbors with $k = 5$ (**RK-knn**). We use our own R translation of the original MATLAB code provided by Berrendero et al. (2017).

- The Mahalanobis-type classifier described in Chapter 2 (**Mah**), where the smoothing parameter α is selected by 5-fold cross-validation among 20 equispaced values between 10^{-4} and 0.1.

- Functional k-nearest neighbors with $k = 5$ (**knn5**), using the function "classif.knn" of the `fda.usc` R package. In spite of its simplicity, the performance of this classifier is usually good, although it is not very efficient in terms of execution time for large sample sizes.

For the methods (RKHS-sq, RKHS-mm, RK and RK-knn) that perform variable selection directly on the curves, the number of t_i points is limited to a maximum of ten. In the tested examples we have seen that usually a small number of points (less than ten) is selected, so there is no real penalty to pay with this restriction and the execution time is not unnecessarily increased. For the two methods based on PCA, at most 30 basis elements are considered.

Simulated data sets

Hereunder we present the different models under study. We aim at presenting a miscellaneous selection, with some models satisfying the RKHS logistic model as well as other that do not. We also include both Gaussian and non-Gaussian processes, and processes with smooth and rough trajectories. The trend and slope functions have been selected in order to define non-trivial problems and to simultaneously ensure enough dependence between the curves and the response. The logarithm and trigonometric functions already appeared in Chapter 3.

- The first data set follows the Gaussian setting described at the beginning of Section 4.1.1 (**Bm fin**), where P_0 is a standard centered Brownian motion and P_1 is a standard Brownian motion plus the trend $m_1(s) = 2\min(0.2, s) - 3\min(0.5, s) + \min(0.7, s)$. This function belongs to the RKHS of the problem since $K(s,t) = \min(s,t)$ is the covariance function of the standard Brownian motion.

- In this case the population P_0 is defined as in the previous point and for P_1 we add the mean function $m_1(s) = \log(s+1)$ (**Bm logs**). This function has already appeared in the previous chapter. It also belongs to the RKHS of the Brownian motion, which consists of all the absolutely continuous functions $f \in L^2[0,1]$ such that $f(0) = 0$ and $f' \in L^2[0,1]$. However it is clear that, in this case, m_1 has not a finite representation of type (4.5).

- The regressors X are generated from an integrated Brownian motion $X(s) = \int_0^s B(t)\mathrm{d}t$, with B a standard Brownian motion (**iBm**). The trajectories of this process are rather smooth. The responses Y are drawn from a Bernoulli random variable whose parameter is given by the functional logistic regression model presented in Theorem 4.1. The intercept β_0 is equal to zero and the slope function used is $\beta(s) = 2K(0.2, s) - 4K(0.5, s) - K(0.7, s)$, K being the covariance function of X. Note that it is not necessary to know the explicit expression of K since in this case we recover the finite dimensional model (4.6) with $(\beta_1, \beta_2, \beta_3) = (2, -4, -1)$ and $(t_1, t_2, t_3) = (0.2, 0.5, 0.7)$.

- In order to include another process with smooth trajectories, the fractional Brownian Motion with Hurst's exponent $H = 0.9$ is used (**fBm**). These processes already appeared in Chapter 3. Although for $H > 0.5$ the trajectories look smoother than the ones of the Brownian motion, they are still not differentiable at every point. The classes are assigned as is the previous point.

- The process X is a mixture of a standard centered Brownian motion $B(s)$ and another independent Brownian motion $\sqrt{2}B'(s)$, being both distributions equiprobable (**MixtSd**). The response Y is generated as in the previous point using the same points t_j but with $(\beta_1, \beta_2, \beta_3) = (2, -3, 1)$.

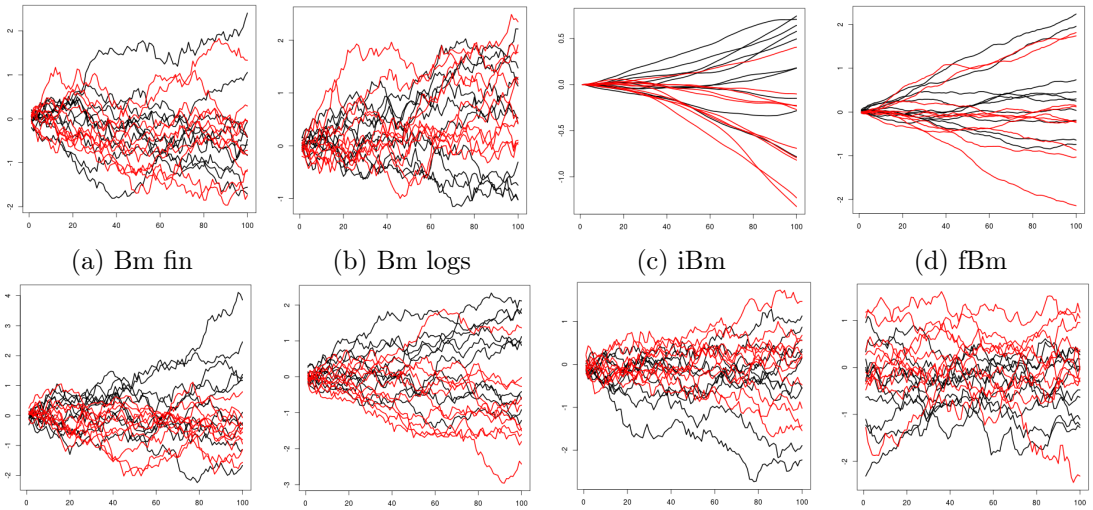
- Similar to the previous one, but now X is an equiprobable mixture of a standard Brownian motion with trend $m(s) = s$ and another standard Brownian motion with trend $m(s) = -s$ (**MixtM**). The responses are generated as in the previous point.

- The covariates X are drawn from a standard centered Brownian motion. As mentioned at the end of Section 3.2.1, in this case the inverse of the Loève's isometry is the stochastic integral $\int_0^1 \beta'(s)\mathrm{d}X(s)$ for $\beta \in \mathcal{H}(K)$ (in the second point of this list we recall that all the functions in this $\mathcal{H}(K)$ are a.s. derivable with respect to Lebesgue measure). Then, the responses Y are realizations of a Bernoulli variable with parameter given by Equation (4.2) with slope function $\beta(s) = \sin(\pi s)$ (**Bm sin**).

- Now the curves are generated from a long-term (stationary) Ornstein-Uhlenbeck process, as explained in Example 3.22 (**OU**). For the responses we would like to use the same procedure as in the previous point, with $\beta(s) = \sin(\pi s)$. However, we do not know the exact expression of the inverse of Loève's isometry for this RKHS. Then, we approximate it as $\Psi_X^{-1}(\beta) \simeq \beta(S)'\Sigma_S^{-1}X(S)$, where $S = \{s_1, \ldots, s_m\}$ is an equispaced grid in $[0, 1]$ and $\beta(S)' = (\beta(s_1), \ldots, \beta(s_m))$. Equivalently for $X(S)$. By Theorem 6D of Parzen (1959) (and Theorem 6E for the convergence of the norms), we know that this expression converges to $\Psi_X^{-1}(\beta)$ when the number of points in the grid increases.

Twenty trajectories of each data set can be found in Figure 4.1. 200 samples with equal prior probability of the classes are used to train the classifiers, and 50 for test. Each experiment is replicated 100 times, in order to obtain a good approximation of the mean error.

The estimated misclassification rates are available in Table 4.1, where the standard deviation of these rates are in brackets. The two best results for each data set appear in bold (in the case of a tie, it is marked the set with less variance). We can see that our sequential proposal and RK-VS are mainly the winners. It is worth mention that RK-VS is proved to be optimal for the first data set, and our proposal obtains almost the same error without assuming Gaussianity of the data. Regarding our two proposals,



(a) Bm fin (b) Bm logs (c) iBm (d) fBm

(e) MixtM (f) MixtSd (g) Bm sin (h) OU

Figure 4.1: Simulated data sets, 10 trajectories of each class (0 black and 1 red).

it seems that RKHS-sq approach outperforms RKHS-mm for most of these data sets. However, this difference is not so clear for data sets with smooth trajectories. It might be because the likelihood function had less local maxima in this case.

Real data sets

The proposals have been also tested for four real data sets, which is a more neutral scenario. The presented sets are commonly used in the literature of functional classifications and are all freely available.

- The first set consists of log-periodogram curves (**Phoneme**) of the pronunciation of the two phonemes AA and AO. For phoneme AA we have 695 samples (class 0) and for phoneme AO, 1022 (class 1). For each recording, 150 frequencies are kept, sampled over a grid of 256 points. This data set is quite common and it is used, for instance, by Ferraty and Vieu (2006). The complete data set can be found along with the online material of that book. A smaller version with 500 samples can be found as part of the R package `fda.usc`.

- Mitochondrial calcium overload data sets, which originally appeared in Ruiz-Meana et al. (2003). The overload is measured for two groups of mouse cardiac cells, one belonging to a control group and one receiving a treatment. The measures are taken every 10 seconds in an hour. The first three minutes are removed since the curves have then an erratic behavior not relevant for the prediction problem, so the grid has size 341. For technical reasons, the experiment was done

	Bm fin	Bm logs	iBm	fBm
RKHS-sq	**0.309** (0.065)	0.395 (0.074)	0.351 (0.079)	0.211 (0.061)
RKHS-mm	0.314 (0.063)	0.397 (0.071)	**0.332** (0.069)	0.216 (0.063)
wav	0.319 (0.071)	0.411 (0.078)	0.337 (0.078)	**0.205** (0.056)
PCA	0.329 (0.076)	0.399 (0.068)	**0.328** (0.072)	0.210 (0.059)
PCA-knn	0.392 (0.074)	0.411 (0.071)	0.374 (0.064)	0.221 (0.053)
nonP	0.312 (0.067)	**0.388** (0.064)	0.338 (0.077)	**0.206** (0.056)
RK	**0.306** (0.067)	0.388 (0.075)	0.333 (0.071)	0.211 (0.056)
RK-knn	0.348 (0.067)	0.413 (0.075)	0.355 (0.069)	0.222 (0.059)
Mah	0.340 (0.070)	**0.384** (0.074)	0.335 (0.064)	**0.206** (0.056)
knn5	0.375 (0.071)	0.411 (0.069)	0.375 (0.067)	0.220 (0.058)

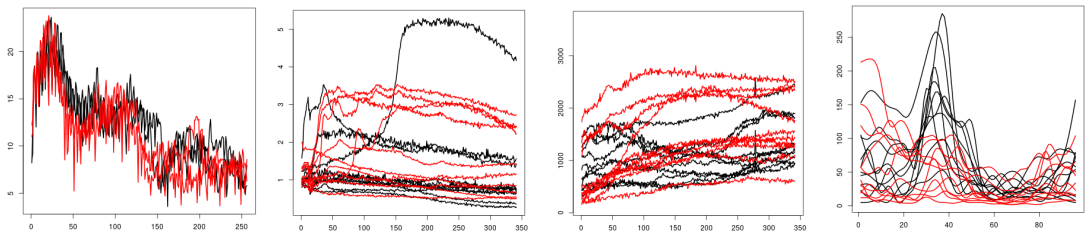
	MixtSd	MixtM	Bm sin	OU
RKHS-sq	**0.301** (0.068)	**0.333** (0.074)	**0.239** (0.058)	**0.235** (0.063)
RKHS-mm	0.313 (0.076)	0.336 (0.072)	0.243 (0.062)	0.236 (0.066)
wav	0.333 (0.071)	0.347 (0.078)	0.247 (0.063)	0.246 (0.063)
PCA	0.311 (0.069)	**0.331** (0.076)	0.248 (0.060)	0.245 (0.063)
PCA-knn	0.398 (0.074)	0.410 (0.077)	0.297 (0.067)	0.323 (0.068)
nonP	0.321 (0.081)	0.335 (0.066)	0.247 (0.058)	0.240 (0.064)
RK	**0.310** (0.068)	0.334 (0.071)	**0.240** (0.055)	**0.233** (0.063)
RK-knn	0.354 (0.066)	0.371 (0.077)	0.275 (0.064)	0.294 (0.065)
Mah	0.358 (0.069)	0.362 (0.088)	0.258 (0.061)	0.277 (0.063)
knn5	0.393 (0.078)	0.407 (0.068)	0.292 (0.068)	0.314 (0.066)

Table 4.1: Misclassification rates for the simulated data sets.

with both the original intact cells (**MCO-I**) and "permeabilized" cells (**MCO-P**). The class label indicates the membership to the control group (class 0, 45 samples for both sets) or the treatment group (class 1, 45 samples for permeabilized cells and 44 for intact ones). The complete data set is available in `fda.usc` package.

- Levels of air pollution measured in Poblenou in Barcelona (Spain), where each curve represents the daily nitrogen oxide measurements, NO_x (**Poblenou**). The original curves were recorded hourly, but we have sub-sampled them to obtain measures every 15 minutes. In order to do this, we have represented the curves in a Bspline base of 50 elements and then evaluated them in a thinner grid of size 115. The weekends and festive days belong to class 1 (39 curves), while the working days are labeled as class 0 (76 curves).

Ten trajectories of each class of these data sets are presented in Figure 4.2. For the phoneme data set only two trajectories of each class are plotted, since these curves are rather rough. However, we decided not to pre-process and smooth the trajectories



(a) Phoneme (b) MCO-P (c) MCO-I (d) Poblenou

Figure 4.2: Real data sets, at most 10 trajectories of each class (0 black and 1 red).

of the sets. This determination is founded on the conclusions of Carroll et al. (2013), where the authors found that under-smoothing is in general desirable for functional classification problems.

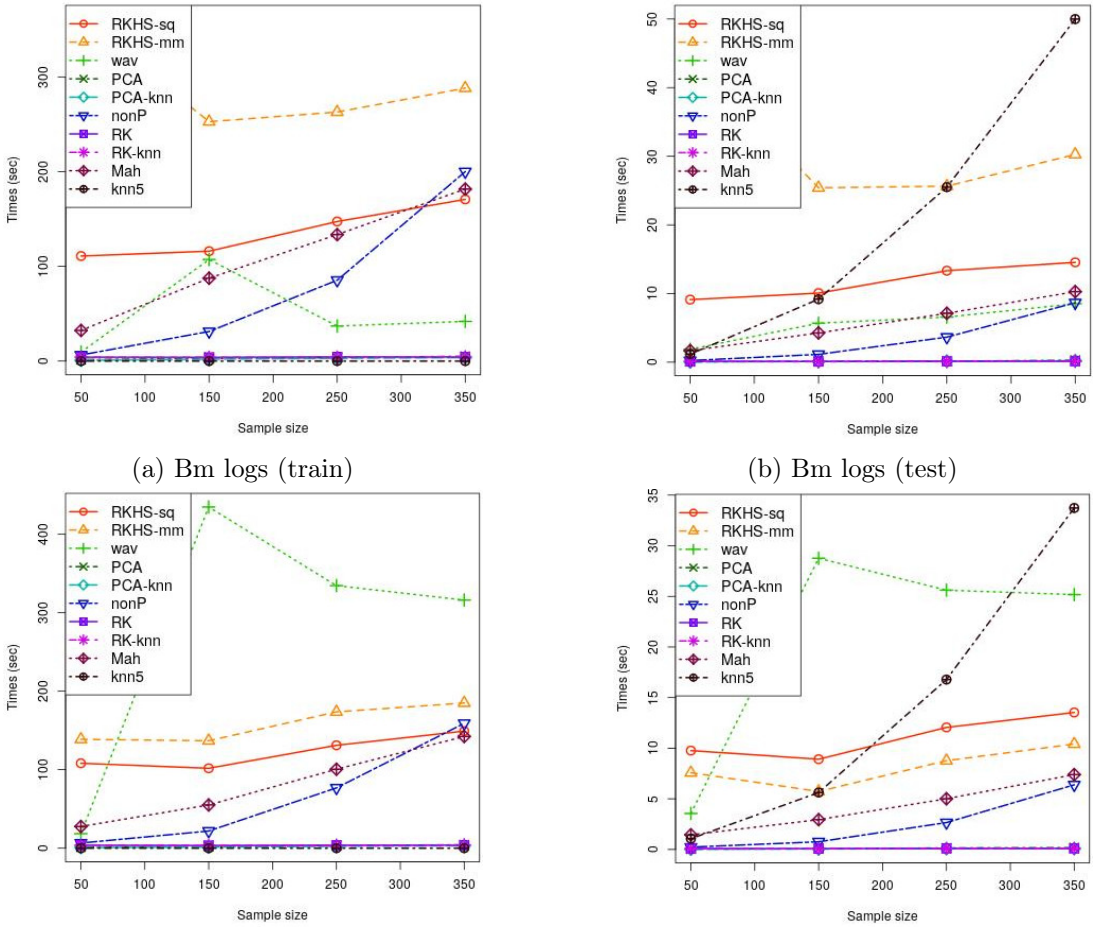
In order to better approximate the misclassification rate, we use 5-fold cross validation. The resulting mean rates and their standard deviations (in brackets) can be found in Table 4.2. The non-parametric method seems to outperform the others in general. Our proposals are competitive and are among the best options for the phoneme set, which is the largest one. The non-parametric method also performs well for this set but, as we will see down below, this method is rather slow for large data sets. Regarding variable selection, both of our proposals select 8.5 points on average for the real data sets.

	Phoneme	MCO-P	MCP-I	Poblenou
RKHS-sq	**0.181** (0.022)	0.256 (0.150)	0.170 (0.108)	0.113 (0.050)
RKHS-mm	0.197 (0.015)	0.344 (0.115)	0.113 (0.070)	0.113 (0.058)
wav	0.181 (0.027)	**0.233** (0.046)	**0.068** (0.048)	0.096 (0.036)
PCA	0.248 (0.043)	0.411 (0.128)	0.180 (0.174)	0.174 (0.097)
PCA-knn	0.319 (0.040)	0.278 (0.104)	0.258 (0.149)	0.209 (0.089)
nonP	**0.170** (0.027)	0.267 (0.099)	**0.045** (0.025)	**0.078** (0.048)
RK	0.187 (0.028)	0.322 (0.133)	0.091 (0.066)	**0.052** (0.036)
RK-knn	0.220 (0.017)	0.344 (0.061)	0.201 (0.100)	0.087 (0.043)
Mah	0.209 (0.013)	0.389 (0.056)	0.112 (0.056)	0.252 (0.048)
knn5	0.233 (0.031)	**0.244** (0.084)	0.190 (0.133)	0.113 (0.073)

Table 4.2: Misclassification rates for the real data sets.

Execution times

We have measured the execution times of all the methods for both the real and simulated data sets. In the latter case we have analyzed the impact of the sample size in the efficiency of the methods, using sample sizes $n = 50, 150, 250$ and 350. We have distinguished between train and test execution time and measured them separately. For



(a) Bm logs (train)

(b) Bm logs (test)

(c) iBm (train)

(d) iBm (test)

Figure 4.3: Train and test execution times when increasing the sample size for "Bm logs" and "iBm" data sets.

instance, k-nearest neighbors needs no train time, but it is rather slow in testing. The measures have been repeated 100 times for each sample size. Instead of present all the possible tables (which would be quite daunting and uninformative), we present just the average times for the different data sets in Table 4.3. The behavior of the method is almost identical for all the data sets, except for the integrated Brownian motion, whose trajectories are rather smooth. It can be seen graphically (for the Brownian motion with logarithms and the integrated Brownian motion sets) in Figure 4.3.

The execution times for the real data sets are in Table 4.4. As mentioned before, the non-parametric method for the phoneme set, which is the best regarding classification error, is rather slow in this case.

Train

	50	150	250	350
RKHS-sq	142.988 (34.073)	136.464 (26.741)	169.873 (32.535)	199.868 (38.020)
RKHS-mm	385.097 (94.214)	266.770 (57.423)	284.623 (57.384)	318.108 (64.460)
wav	10.779 (3.781)	210.592 (66.123)	129.558 (48.619)	112.721 (34.186)
PCA	1.729 (0.700)	2.714 (1.041)	4.346 (1.070)	5.867 (0.989)
PCA-knn	0.921 (0.381)	2.006 (0.772)	3.045 (0.874)	3.905 (0.747)
nonP	7.007 (1.977)	31.521 (7.737)	88.500 (18.403)	195.991 (36.461)
RK	4.294 (1.135)	4.146 (1.082)	4.468 (1.035)	4.679 (1.029)
RK-knn	4.229 (1.111)	4.132 (1.116)	4.347 (1.044)	4.670 (1.093)
Mah	34.605 (9.635)	86.517 (21.158)	140.661 (31.676)	188.376 (37.497)
knn5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)

Test

	50	150	250	350
RKHS-sq	11.497 (9.554)	12.769 (8.840)	16.890 (10.868)	21.094 (12.666)
RKHS-mm	42.846 (33.520)	30.715 (19.248)	32.150 (18.186)	35.612 (18.563)
wav	2.147 (0.995)	18.852 (12.826)	12.038 (5.481)	13.229 (5.211)
PCA	0.064 (0.070)	0.189 (0.172)	0.331 (0.222)	0.484 (0.270)
PCA-knn	0.043 (0.053)	0.137 (0.139)	0.226 (0.199)	0.301 (0.250)
nonP	0.256 (0.086)	1.128 (0.424)	3.424 (1.153)	8.385 (2.282)
RK	0.142 (0.175)	0.158 (0.162)	0.185 (0.188)	0.223 (0.215)
RK-knn	0.130 (0.173)	0.142 (0.158)	0.167 (0.163)	0.206 (0.205)
Mah	1.760 (0.688)	4.582 (1.826)	7.477 (2.910)	10.629 (3.870)
knn5	1.281 (0.497)	9.335 (3.307)	27.110 (8.165)	53.178 (15.185)

Table 4.3: Mean train and test execution times in seconds for the simulated data sets.

Although our approaches are the slowest for the tested data sets, it seems that they scale better with the sample size than others. For instance, our sequential proposal RKHS-sq shows almost no deterioration when the sample size increases. This is in contrast to the performance we observe for training for the non-parametric and Mahalanobis-type methods and knn for testing. Concerning our two proposals, RKHS-sq seems to perform better in general, although this claim is not so clear for smooth data sets. The RKHS-mm approach is slower than the sequential version, despite the latter performs an exhaustive search, since RKHS-mm optimizes a function in $\mathbb{R}^{10}$. However, RKHS-mm could outperform RKHS-sq in a thinner grid. For the integrated Brownian motion (with smooth trajectories) the wavelet-based method shows a somewhat erratic behavior, which is not observed in the remaining models and methods. Then, taken into account all the different results of this section, it seems that our proposals are more appropriate for large data sets.

	Phoneme	MCO-P	MCO-I	Poblenou
RKHS	515.782 (16.052)	236.248 (11.396)	490.844 (21.084)	161.134 (12.205)
RKHSem	661.897 (48.661)	401.165 (90.112)	413.640 (74.679)	320.167 (41.019)
wav	107.774 (4.919)	3.682 (0.236)	7.682 (0.694)	6.181 (0.527)
PCA	30.359 (1.542)	0.817 (0.091)	1.817 (0.280)	1.507 (0.202)
PCA-knn	30.215 (0.387)	0.307 (0.014)	0.658 (0.139)	0.582 (0.111)
nonP	2105.373 (102.223)	1.300 (0.043)	2.858 (0.276)	3.261 (0.379)
rkc	2.741 (0.232)	1.211 (0.090)	2.425 (0.281)	0.759 (0.112)
rkc-knn	2.523 (0.254)	1.179 (0.131)	2.330 (0.307)	0.699 (0.013)
mah	1262.118 (17.513)	77.178 (1.897)	151.884 (21.226)	22.371 (2.421)
knn5	105.009 (10.501)	0.175 (0.008)	0.364 (0.170)	0.333 (0.095)

Table 4.4: Execution times (train plus test) in seconds for the real data sets.

4.4.2 What if we increase the number p of selected points?

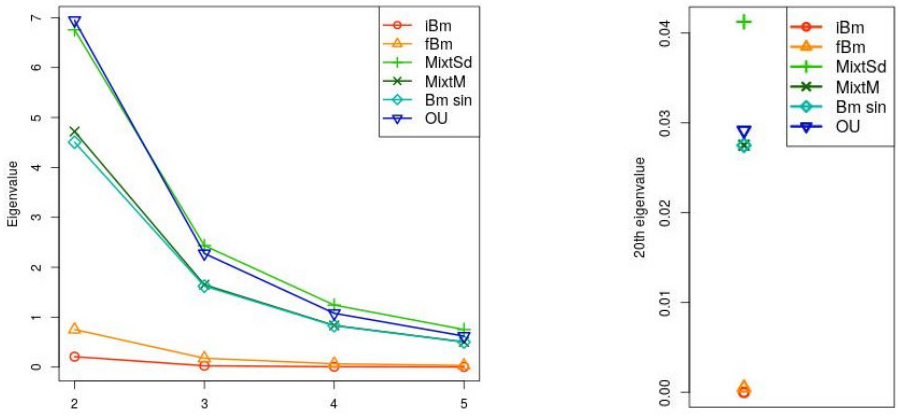
In order to analyze whether finite models (4.6) approximate model (4.2) when p increases, we have measured the distances in the reproducing kernel Hilbert spaces between the true and the estimated β functions. We measure it for the six last simulated data sets, for which model (4.2) is fulfilled.

We are interested in the asymptotic behavior when increasing the number of points p in (4.6). Then, we adjust the finite logistic model for p randomly generated points uniformly in $[0, 1]$, for $p = 1, 5, 10, 15, 20$. For most RKHS's we do not know the explicit expression of the norm. Then, for a grid $S = \{s_1, \ldots, s_m\}$, a kernel function K and $f \in \mathcal{H}(K)$, we estimate the squared norm $\|f\|_K^2$ as

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f(s_i)K(s_i, s_j)f(s_j).$$

By Theorem 6E of Parzen (1959), we know that this expression converges to $\|f\|_K^2$ when $m \to \infty$. The covariance and β functions for each model are:

- iBm: $K(s,t) = \frac{1}{6}\big(3\max(s,t)\min(s,t)^2 - \min(s,t)^3\big)$ and $\beta(s) = 2K(s, 0.2) - 4K(s, 0.5) - K(s, 0.7)$.
- fBm: $K(s,t) = 0.5(s^{1.8} + t^{1.8} - |s-t|^{1.8})$ and $\beta(s) = 2K(s, 0.2) - 4K(s, 0.5) - K(s, 0.7)$.
- MixtSd: $K(s,t) = \frac{3}{2}\min(s,t)$ and $\beta(s) = 2K(s, 0.2) - 3K(s, 0.5) + K(s, 0.7)$.
- MixtM: $K(s,t) = \min(s,t) + st$ and $\beta(s) = 2K(s, 0.2) - 3K(s, 0.5) + K(s, 0.7)$.
- Bm sin: $K(s,t) = \min(s,t)$ and $\beta(s) = \sin(\pi s)$.
- OU: $K(s,t) = \exp(-|t-s|)$ and $\beta(s) = \sin(\pi s)$.



(a) 2^{nd} to 5^{th} eigenvalues. (b) 20^{th} eigenvalues.

Figure 4.4: Eigenvalues of the covariance operator for the different data sets.

Each distance has been measured for 100 independent samples of size 1000. The results are available in Table 4.5, where the standard deviation of the measures are in brackets. We can see that for the data sets with rough trajectories the norm decreases when increasing p. However, for the two smoothest sets, the norm starts to increase again for p greater than 10.

p	1	5	10	15	20
iBm	0.230 (0.211)	0.097 (0.028)	*0.116* (0.032)	*0.139* (0.040)	*0.159* (0.046)
fBm	5.774 (0.979)	*6.005* (1.220)	*6.070* (1.219)	*6.115* (1.259)	*6.155* (1.277)
MixtSd	1.759 (0.295)	1.076 (0.360)	0.762 (0.253)	0.649 (0.201)	0.569 (0.160)
MixtM	1.234 (0.225)	0.734 (0.239)	0.539 (0.174)	0.452 (0.138)	0.418 (0.120)
Bm sin	3.953 (0.689)	1.972 (0.768)	1.293 (0.533)	1.104 (0.426)	1.026 (0.348)
OU	2.249 (0.305)	1.367 (0.263)	0.954 (0.294)	0.789 (0.253)	0.674 (0.204)

Table 4.5: Approximations to $\|\widehat{\beta} - \beta\|_K^2$.

This increasing of the norms has a theoretical explanation. There are in the literature several results on the asymptotic consistency of the MLE when the number of predictors increases (for instance, Portnoy (1988)). However, as pointed out by Cardot and Sarda (2005), *"the main point in the above works is to suppose that the covariance matrix is bounded below. That is not the case for functional data since the covariance operator is compact"*. This boundedness is closely related with the limit of the eigenvalues, which should be strictly greater than zero in order to obtain consistency. In fact, we can see in Figure 4.4 that the eigenvalues of the covariance operator of the two smoothest data sets converge to zero much faster than the others.

Therefore, there is no hope to obtain a fully general consistent estimator with the proposed methodology. But one could think about using different techniques, like the

penalized ML proposed in Cardot and Sarda (2005). After all, we know that our target exists and it is the only global maximum of the expected log-likelihood function.

Chapter 5

Conclusions

The interest of the scientific community in functional data analysis has undergone an enormous increase in the last decades. Then, a considerable effort is being dedicated to develop statistical techniques involving functional data. However, in our opinion, there is still a lack of a methodological and theoretical basis on some of the algorithmic proposals.

With this thesis we have tried to delve further into some statistical problems with functional data from a mathematical point of view. In particular, we have made use of RKHS theory. RKHS's have proven to be very useful to establish well-founded connections between functional and multivariate problems. This helps us to better understand the differences between these two settings, which are often crucial. These new insights allow us to design statistical techniques specially conceived for functional data, which are usually more powerful that the direct applications of multivariate procedures. From a general, methodological point of view, this work represents an additional example of the surprising usefulness of reproducing kernels in statistics. Additional examples, apart from the already mentioned throughout this thesis, can be found in Berrendero et al. (2017); Berlinet and Thomas-Agnan (2004); Hsing and Eubank (2015); Yuan and Cai (2010) or Kadri et al. (2016).

Each chapter of this thesis is devoted to a different statistical problem. Then we discuss each of them separately.

Mahalanobis-type distance

There are in the literature a couple of interesting proposals to extend the Mahalanobis distance to functional data. Up to now efforts have focused on forcing the convergence of the series in Equation (2.7) when measuring the distance between trajectories of the process. However, the underlying problem was not explicitly considered. By means of RKHS theory we are able to tackle the real cause of the divergence of the series. In view of the discussion of Chapter 2, the proposed methodology seems rather natural.

Besides, the proposed distance is truly a metric and shares some interesting properties with the classical definition.

Mahalanobis distance has found many applications for multivariate problems. We have checked the functional proposal in three of them: exploratory analysis (including functional boxplots and outliers detection), functional binary classification and inference on the mean. In view of the obtained results for these problems, the proposal seems quite competitive.

Functional linear regression

The RKHS approach we have introduced in Chapter 3 for functional linear regression provides a natural framework for a formal unified theory of variable selection. The "sparse" models (those where the variable selection techniques are fully justified) appear as particular cases in this setup. As a consequence, it is possible to derive asymptotic consistency results as those we have obtained. Likewise, it is also possible to consider the problem of estimating the "true" number of relevant variables in a consistent way, as we do in Section 3.4. This is in contrast with other standard proposals for which the number of variables is previously fixed as an input, or it is determined using cross validation or other computationally expensive methods. Then, our proposal is more firmly founded in theory and, at the same time, provides a much faster method in practice, which is important when dealing with large data sets.

For the problem with scalar response, the empirical results we have obtained are encouraging. In short, according to our experiments, the RKHS-based method works better than other variable selection methods in those sparse models that fulfill the ideal theoretical conditions for the method. In the non sparse model considered in the simulations, the RKHS method is slightly outperformed by other proposals (but still behaves reasonably). Finally, in the "neutral" field of real data examples the performance looks also satisfactory and competitive.

Afterwards we have extended the theory developed for regression with scalar response to functional response and we have adapted it to the setting of prediction of functional time series, whose dependence is modeled using an autoregressive structure. Our variable selection approach helps to overcome some of the usual problems coming from the use of other dimension reduction techniques for AR processes.

When compared with other prediction methods of the literature, our proposal is quite competitive. The results obtained for the real data sets tested are encouraging and specially relevant for the smallest data set (which include only 32 curves). This might be due to the simplicity of our estimator, which does not require large sample sizes to obtain good approximations. In addition, the execution times of our implementation are smaller than the competitors. That is, our proposals, particularly the discrete

approaches, are also more suitable for large data sets. Furthermore, the proposed estimators can be directly adapted to discrete or fully functional data sets.

Functional logistic regression

In Chapter 4 a variable selection methodology for functional logistic regression is presented. Variable selection in this setting is fully mathematically grounded, similarly to the methodology developed in Chapter 3. We present an RKHS-based model which follows from Gaussian conditional distributions, but it is more general. The classical functional logistic model based on the inner product in $L^2[0,1]$ can also be obtained on this Gaussian setting. However, our proposal can be seen as a generalization of the L^2 model, in the sense that we require minimal conditions on the mean functions of the data.

The finite dimensional model of Equation (4.6) is a particular case of the proposed RKHS-model. We propose to estimate the coefficients of the slope functions with the maximum likelihood estimator (MLE). It is well-known that MLE may not exists for multiple logistic models. We have then carefully analyzed the existence of the MLE for functional data. We proved that the non-existence problems are drastically worsened. Briefly, for some important processes, the MLE does not exists with probability one for any sample size and, even if it exists for a given sample size, it does not exist asymptotically with probability one. This represents another crucial difference between finite and infinite settings.

Due to the compactness of the covariance operator, the proposed estimator is not consistent for non-sparse slope functions. However, the proposal is conceptually simple and performs good in practice. In fact, it is quite competitive for binary classification and the execution time scales really well with the sample size. Then, our proposal would be more suitable for large data sets. Indeed, the sample size has only a marginal effect on the execution time of our proposal, which is more affected by the grid size.

Open problems

We summarize now some interesting topics that remain open for future research:

- Equation (1.8) clearly resembles a kernel density estimator. Although this type of embeddings has been used to estimate density functions, it was from a rather different perspective. Then, given an empirical probability distribution F_n, one could think about directly estimating the density function as μ_{F_n}. Then, since $\mu_{F_n} \in \mathcal{H}(K)$, it would be interesting to analyze the properties of such an estimator from an RKHS perspective.

- In Section 1.2.2 we introduce the functional binary classification problem for two populations P_0 and P_1. Whenever P_0, P_1 are Gaussian distributions sharing covariance structure with mean functions zero and $m \in \mathcal{H}(K)$ respectively, the optimal classifier assigns a realization x to class one if $2\langle x, m\rangle_K > \|m\|_K^2$ (where the "inner product" is interpreted as Loève's isometry). Besides that, the natural classifier using Mahalanobis-type distance classifies x to class one whenever $\|x_\alpha\|_K > \|x_\alpha - m_\alpha\|_K$, which equals $2\langle x_\alpha, m_\alpha\rangle_K > \|m_\alpha\|_K^2$. Both classifiers look very similar and, in fact, the Mahalanobis one equals the optimal for $\alpha = 0$ and m sparse (i.e. $m \in \mathcal{H}_0(K)$). Then, it would be interesting to analyze this relationship in depth. For instance, we conjecture that the new classifier may converge to the optimal one for a general mean function if $\alpha_n \to 0$ at an adequate rate when $n \to \infty$.

- The focus of Chapter 2 is mainly theoretical and we did not have any concrete application in mind. Thus, we put less efforts in the concrete implementations of the applications. This is specially significant on the selection of the smoothing parameter α when cross-validation is not suitable. As mentioned in that chapter, it would be interesting to deepen in different ways to automatically select α, in a similar sense as it is made in the adjusted outliergram of Arribas-Gil and Romo (2014).

- Throughout Chapter 3 we restrict ourselves to variable selection problems with a small fixed number of points p. Using some results of Parzen (1959) and Cambanis (1985) one can derive asymptotic results like Proposition 3.19. However, the consistency of the response estimators when $p \to \infty$ is not clear. In this case one should handle p and n going both to infinity. Some version of the Representer Theorem, in the same spirit as it is used in Preda (2007), might be useful to prove consistency of these estimators. We propose to minimize $n^{-1}\sum_i (y_i - \Psi_{x_i}^{-1}(\beta))^2$, so one could understand $\Psi_{x_i}^{-1}(\beta) = f_\beta(x_i)$, where $f_\beta : L^2[0,1] \to \mathbb{R}$. Then the corresponding penalization term would be $\omega(\|f_\beta\|_R)$, with $R : L^2[0,1] \times L^2[0,1] \to \mathbb{R}$ a reproducing kernel. It would be interesting to study the solutions of these penalization problems and if this type of penalization is related to the natural penalization $\|\beta\|_K^2$.

- In order to fix the number of variables p to select in the sparse representation, we propose an approach based on 2-means, but different techniques could be used. For instance, it would be interesting to analyze the following idea, suggested by Prof. Juan Cuesta-Albertos. If one compares the points T_p selected for $p < p^*$ with different bootstrap samples, they should not significantly vary. However, if $p > p^*$ is used, the variance of the selected points would be larger than in the previous case.

- When we extend the results on variable selection for regression with scalar response to functional response, we merely made a translation of all the results. This introduces some "artificial" restrictions on the slope function when proving

the convergence of the cross-covariance function in Lemma 3.14. Although this result is indispensable to obtain a consistent estimator of the selected variables, it may not be so important if one is only interested in prediction accuracy. That is, one may obtain consistent estimators of the response with a similar technique but forgetting about the particular involved points. As before, this also suggests also to analyze the behavior of the estimator when the number of points p goes to infinity.

- In connection with the previous point, when variable selection is applied to functional time series, the existence of a unique stationary solution is only proved for sparse kernels. As mentioned there, the proof we obtained for a general kernel involves the assumption that the limit in $\mathcal{L}$ (the space of bounded linear operators on $C[0,1]$) of finite rank operators as in (3.55) is bounded. This is a rather artificial condition and we do not have any intuition about its fulfillment. Then, it would be interesting to think about more natural conditions on the process that would imply this boundedness.

- In this last part of Chapter 3 about functional time series, we need the process to be stationary in order to have $\mathcal{H}(K)$ unaffected by n. We are currently starting a collaboration work with Florian Heinrichs, from Ruhr-Universität Bochum, to adapt the proposed RKHS techniques to define a test to detect stationarity in this context.

- Chapter 4 is focused on variable selection for logistic regression. Unlike in Chapter 3, with the discussion in section 4.4.2 we positively know that we would not be able to obtain a general consistent estimator with the proposed technique. Thus, it would be interesting to think about different estimators for this problem. For instance, one possibility would be to analyze the proposal of Cardot and Sarda (2005) about penalized Maximum Likelihood. If we are able to obtain consistent estimators of the coefficients of the finite logistic model when the number of regressors increase, we would be able to prove consistency of the general RKHS model by means of the results in Parzen (1959).

- In view of the remark after Proposition 4.5, it would be interesting to analyze the properties of functional logistic models for Lévy processes, even though most of them have discontinuous trajectories.

- Regarding the two particular implementations of our proposal in Chapter 4, it seems that the size of the grid might strongly affect the execution time of the sequential approach. Besides, the "max-max" algorithm seems to perform better for smooth data sets. It would be interesting to conduct a more detailed simulation study to better understand the behavior of these implementations.

Capítulo 6

Conclusiones

El interés de la comunidad científica por el análisis de datos funcionales ha experimentado un enorme crecimiento en las últimas décadas. Por lo tanto, se ha dedicado un esfuerzo considerable al desarrollo de técnicas estadísticas para datos funcionales. Sin embargo, en nuestra opinión, todavía hay una carencia teórica y metodológica en alguna de las propuestas algorítmicas actuales.

Con esta tesis hemos tratado de profundizar en algunos problemas estadísticos con datos funcionales desde un punto de vista matemático. En particular, hemos hecho uso de la teoría de RKHS's. Estos espacios resultan ser de gran utilidad para establecer conexiones fundadas entre problemas funcionales y multivariantes. Esto nos ayuda a entender mejor las diferencias entre estos dos contextos, que suelen ser de vital importancia. Este nuevo punto de vista nos permite diseñar técnicas estadísticas especialmente concebidas para datos funcionales, que suelen ser más potentes que las meras aplicaciones de procedimientos multivariantes. Desde un punto de vista metodológico más general, este trabajo representa un ejemplo adicional de la sorprendente utilidad de los núcleos reproductores en estadística. Otros ejemplos, además de los que ya han ido apareciendo a lo largo de la tesis, son Berrendero et al. (2017); Berlinet and Thomas-Agnan (2004); Hsing and Eubank (2015); Yuan and Cai (2010) o Kadri et al. (2016).

Cada capítulo de esta tesis está dedicado a un problema estadístico diferente. Por lo tanto, vamos a discutir cada uno de ellos por separado.

Distancia de Mahalanobis funcional

Existen en la literatura un par de propuestas interesantes en las que se extiende la distancia de Mahalanobis a datos funcionales. Hasta el momento los esfuerzos se han centrado en forzar la convergencia de la serie presente en la ecuación (2.7) cuando se mide la distancia entre trayectorias del proceso. Sin embargo, en dichos trabajos no se considera explícitamente el problema de fondo. Por medio de la teoría de RKHS's somos capaces de explicar la verdadera causa de la divergencia de esta serie. En vista de

la discusión del capítulo 2, la metodología propuesta parece bastante natural. Además, la distancia propuesta es una verdadera métrica que comparte algunas propiedades interesantes con la definición clásica.

La distancia de Mahalanobis tiene muchas y diversas aplicaciones en problemas multivariantes. Hemos probado la metodología propuesta en tres de ellas: el análisis exploratorio de datos (incluyendo los gráficos de cajas funcionales y la detección de atípicos), la clasificación binaria funcional y la inferencia para la media. En vista de los resultados obtenidos para estos problemas, la distancia propuesta parece bastante competitiva.

Regresión lineal funcional

El enfoque RKHS introducido en el capítulo 3 para regresión lineal funcional nos proporciona un marco natural para desarrollar una teoría formal y unificada para la selección de variables. Los modelos finitos (aquellos para los cuales las técnicas de selección de variables están plenamente justificadas) aparecen como un caso particular en este contexto. Como consecuencia, es posible obtener resultados asintóticos de consistencia como los presentados en dicho capítulo. Del mismo modo, también es posible considerar el problema de estimar consistentemente el "verdadero" número de variables relevantes, como hacemos en la sección 3.4. A diferencia de otras propuestas clásicas para las cuales el número de variables está fijado de antemano, o se determina usando validación cruzada y otros métodos computacionalmente costosos. Por lo tanto, nuestra propuesta está más firmemente sustentada desde un punto de vista teórico y, al mismo tiempo, el procedimiento obtenido es mucho más rápido en la práctica, lo que es importante cuando se trata con conjuntos grandes de datos.

Para el problema con respuesta escalar, los resultados empíricos obtenidos son alentadores. En resumen, de acuerdo con nuestros experimentos, el método basado en RKHS's funciona mejor que otras técnicas de selección de variables en aquellos modelos finitos que satisfacen las condiciones teóricas ideales para el método propuesto. En los modelos no finitos considerados en las simulaciones, el modelo RKHS es superado ligeramente por otras propuestas (pero su comportamiento es razonable). Finalmente, en el campo "neutral" de los datos reales, el rendimiento también parece satisfactorio y competitivo.

A continuación hemos extendido la teoría desarrollada para respuesta escalar a respuesta funcional, y la hemos adaptado a la predicción de series temporales funcionales, cuya dependencia es modelada por medio de una estructura autorregresiva. Nuestra propuesta de selección de variables ayuda a resolver alguno de los problemas usuales asociados al uso de otras técnicas de reducción de dimensión para procesos AR.

Cuando se compara con otras técnicas de predicción disponibles en la literatura, nuestra propuesta es bastante competitiva. Los resultados obtenidos para los conjuntos de

datos reales son alentadores, siendo especialmente relevantes las diferencias para el conjunto de datos más pequeño (que incluye únicamente 32 curvas para entrenamiento). Esto podría deberse a la simplicidad de nuestro estimador, que no requiere de grandes muestras para obtener buenas aproximaciones. Además, los tiempos de ejecución de nuestras implementaciones son menores que los de los competidores probados. Es decir, nuestra propuesta, particularmente las implementaciones discretas, sería también más adecuada para conjuntos de datos grandes. Por otro lado, el estimador propuesto puede ser adaptado directamente para datos discretizados o completamente funcionales.

Regresión logística funcional

En el capítulo 4 se presenta una metodología para regresión logística funcional. La selección de variables en este contexto está completamente fundada, de forma similar a lo que pasaba en el capítulo 3. Presentamos un modelo basado en los RKHS's, el cual se sigue del modelo con distribuciones condicionales Gaussianas, pero siendo más general. El modelo clásico de regresión logística funcional que involucra el producto escalar en $L^2[0,1]$ también puede obtenerse en este contexto Gaussiano. Sin embargo, nuestra propuesta puede verse como una generalización del modelo L^2, en el sentido que imponemos las mínimas condiciones necesarias a las funciones de medias de las clases.

El modelo finito-dimensional de la ecuación (4.6) es un caso particular del modelo RKHS presentado. Proponemos estimar los coeficientes de la función de pendientes de regresión ("slope function") a través del estimador de máxima verosimilitud (EMV). Es bien sabido que el EMV puede no estar definido para modelos de regresión logística múltiple. En nuestro caso probamos que el problema de no-existencia empeora drásticamente para datos funcionales. En resumen, para algunos procesos importantes, incluyendo el movimiento Browniano, el EMV no existe con probabilidad uno para cualquier muestra, e incluso si existe para una muestra dada, la probabilidad de no existencia aumenta asintóticamente. Esto representa otra diferencial crucial entre los contextos finito e infinito.

Debido a la compacidad del operador de covarianza, el estimador propuesto no es consistente para funciones de pendientes de regresión no dispersas. Sin embargo, la propuesta es competitiva y conceptualmente sencilla. De hecho, el rendimiento es bastante bueno para clasificación binaria y el tiempo de ejecución escala realmente bien con el tamaño muestral. Por tanto, nuestra propuesta sería más recomendable para conjuntos de datos grandes. En realidad el tamaño muestral tiene un efecto mínimo en el tiempo de ejecución de nuestro método, que se ve más afectado por el tamaño de la malla.

Problemas abiertos

A continuación resumimos algunas cuestiones interesantes que han quedado abiertas para trabajo futuro:

- La ecuación (1.8) recuerda claramente a un estimador núcleo de la densidad. Aunque este tipo de inmersiones han sido usadas para estimar densidades, ha sido desde una perspectiva diferente. De hecho, dada una distribución de probabilidad empírica F_n, se podría pensar en estimar directamente la función de densidad como μ_{F_n}. Por tanto, como $\mu_{F_n} \in \mathcal{H}(K)$, sería interesante analizar las propiedades de este estimador desde la perspectiva de la teoría RKHS.

- En la sección 1.2.2 introducimos el problema de clasificación binaria funcional para dos poblaciones P_0 y P_1. Siempre que P_0, P_1 son distribuciones Gaussianas con una estructura de covarianza común y funciones de medias cero y $m \in \mathcal{H}(K)$ respectivamente, el clasificador óptimo asigna la realización x a la clase uno si $2\langle x, m\rangle_K > \|m\|_K^2$ (donde el "producto escalar" debe interpretarse como la isometría de Loève). Por otro lado, el clasificador natural basado en la distancia de Mahalanobis propuesta asigna x a la clase uno si $\|x_\alpha\|_K > \|x_\alpha - m_\alpha\|_K$, que equivale a $2\langle x_\alpha, m_\alpha\rangle_K > \|m_\alpha\|_K^2$. Ambos clasificadores tienen expresiones similares y, de hecho, el clasificador basado en la distancia de Mahalanobis coincide con el óptimo para $\alpha = 0$ y m dispersa (i.e. $m \in \mathcal{H}_0(K)$). Por tanto, sería interesante analizar esta relación más en profundidad. Por ejemplo, conjeturamos que el nuevo clasificador convergerá al óptimo para cualquier función de medias si $\alpha_n \to 0$ a un ritmo adecuado cuando $n \to \infty$.

- El enfoque del capítulo 2 es principalmente teórico y se desarrolló sin ninguna aplicación concreta en mente. Por lo tanto, el esfuerzo dedicado a las implementaciones concretas de las aplicaciones fue menor. Esto se nota especialmente en la selección del parámetro de suavizado α cuando no es posible aplicar validación cruzada. Como se menciona en el capítulo, sería interesante profundizar en el análisis de diferentes técnicas para seleccionar α automáticamente, de forma similar a como se hace para el "adjusted outliergram" de Arribas-Gil and Romo (2014).

- A lo largo del capítulo 3 nos limitamos a problemas de selección de variables para un número pequeño de puntos p. Usando algunos resultados de Parzen (1959) y Cambanis (1985) se pueden derivar resultados asintóticos como en la proposición 3.19. Sin embargo, la consistencia de las respuestas estimadas cuando $p \to \infty$ no está clara. En este caso se debería manejar al mismo tiempo la tendencia de p y n a infinito. Para intentar probar la consistencia de estos estimadores podría ser de utilidad alguna versión del "Representer Theorem", de forma similar a como se usa en Preda (2007). Concretamente, en el capítulo proponemos minimizar $n^{-1}\sum_i (y_i - \Psi_{x_i}^{-1}(\beta))^2$, así que se podría interpretar $\Psi_{x_i}^{-1}(\beta) = f_\beta(x_i)$, donde

$f_\beta : L^2[0,1] \to \mathbb{R}$. Por tanto, el correspondiente término de penalización sería $\omega(\|f_\beta\|_R)$, con $R : L^2[0,1] \times L^2[0,1] \to \mathbb{R}$ un núcleo reproductor. Sería interesante estudiar las soluciones de este tipo de problemas de penalización y ver si existe alguna relación con la penalización natural $\|\beta\|_K^2$.

- Para fijar el número de variables p que se seleccionan, proponemos una aproximación basada en el algoritmo 2-medias, pero podrían usarse diferentes técnicas. Por ejemplo, sería interesante estudiar la siguiente idea, sugerida por el profesor Juan Cuesta-Albertos. Si se comparan los puntos T_p seleccionados para $p < p^*$ con diferentes muestras bootstrap, no deberían variar demasiado. Sin embargo, si se usa $p > p^*$, la variación de los puntos seleccionados será mayor que en el caso anterior.

- Cuando extendemos los resultados de selección de variables con respuesta escalar a respuesta funcional, hacemos simplemente una traducción de todos los resultados. Esto introduce algunas restricciones un poco "artificiales" en la función de pendientes de regresión al probar la convergencia de la función de covarianza cruzada en el lema 3.14. Aunque este resultado es indispensable para obtener un estimador consistente de las variables seleccionadas, podría no ser tan relevante si uno está únicamente interesado en la precisión de la predicción. Es decir, quizás se podría obtener un estimador consistente de la respuesta con una técnica similar pero olvidándose de los puntos concretos que se seleccionan. Como antes, esto también sugiere analizar el comportamiento del estimador cuando p tiende a infinito.

- En relación con el punto anterior, cuando se aplica selección de variables a series temporales funcionales, la existencia de una única solución estacionaria se prueba únicamente para núcleos finitos. Como se menciona allí, la demostración que obtuvimos para un núcleo general requería imponer la condición de que el límite en $\mathcal{L}$ (el espacio de operadores lineales acotados en $C[0,1]$) de operadores de rango finito como en (3.55) es acotado. Esta condición es bastante artificial y no tenemos ninguna intuición sobre en qué casos se satisface. Por tanto, sería interesante pensar en condiciones más naturales sobre el proceso que implicasen la acotación del operador.

- En esa última parte del capítulo 3 sobre series temporales, necesitamos que el proceso sea estacionario para evitar que $\mathcal{H}(K)$ se vea afectado por n. Actualmente estamos comenzando una colaboración con Florian Heinrichs, de la Ruhr-Universität Bochum, para adaptar las técnicas RKHS propuestas a la definición de un test para detectar estacionaridad en este contexto funcional.

- El capítulo 4 está dedicado a la selección de variables para regresión logística. A diferencia del capítulo 3, con la discusión de la sección 4.4.2 sabemos que no vamos a ser capaces de obtener un estimador consistente con la técnica propuesta. Por tanto, sería interesante pensar en diferentes estimadores para este problema.

Por ejemplo, una posibilidad sería analizar la propuesta de Cardot and Sarda (2005) sobre Máxima Verosimilitud penalizada. Si fuésemos capaces de obtener estimadores consistentes para los coeficientes de un modelo de regresión logística finito cuando el número de regresores aumenta, podríamos probar la consistencia del modelo RKHS general por medio de los resultados de Parzen (1959).

- En vista de la observación tras la proposición 4.5, sería interesante analizar las propiedades de los modelos de regresión logística para procesos de Lévy, aunque la mayoría de ellos no tengan trayectorias continuas.
- Respecto a las dos implementaciones propuestas en el capítulo 4, parece que el tamaño de la malla podría afectar considerablemente el tiempo de ejecución de la implementación secuencial. Además, el algoritmo "max-max" parece funcionar mejor para conjuntos de datos suaves. Sería interesante llevar a cabo un estudio de simulación más detallado para entender mejor el comportamiento de ambas implementaciones.

Bibliography

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984. doi: 10.1093/biomet/71.1.1.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. ISSN 1935-8237.

J. Álvarez-Liébana. A review and comparative study on functional time series techniques (under review). *arXiv preprint, arXiv:1706.06288*, 2017.

A. A. Amini and M. J. Wainwright. Approximation properties of certain operator-induced norms on Hilbert spaces. *Journal of Approximation Theory*, 164(2):320–345, 2012.

G. Aneiros and P. Vieu. Variable selection in infinite-dimensional problems. *Statistics and Probability Letters*, 94(C):12–20, 2014.

G. Aneiros and P. Vieu. Sparse nonparametric model for regression with functional covariate. *Journal of Nonparametric Statistics*, 28(4):839–859, 2016.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

A. Arribas-Gil and J. Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 2014.

R. B. Ash and M. F. Gardner. *Topics in Stochastic Processes: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Academic Press, 2014.

A. Aue and J. Klepsch. Estimating functional time series by moving average model fitting. *arXiv preprint, arXiv:1701.00770[ME]*, 2017.

A. Aue, D. Dubart Norinho, and S. Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392, 2015.

A. Baíllo, A. Cuevas, and J. A. Cuesta-Albertos. Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics*, 38(3):480–498, 2011.

J. H. Beder. A sieve estimator for the mean of a Gaussian process. *The Annals of Statistics*, 15(1):59–78, 1987.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, 2004. ISBN 1-4020-7679-7. URL `http://opac.inria.fr/record=b1128469`.

P. Bernard. Analyse de signaux physiologiques. *Mémoire Université Catholique Angers*, 1997.

J. R. Berrendero, A. Cuevas, and J. L. Torrecilla. Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*, 26(2):619–638, 2016.

J. R. Berrendero, A. Cuevas, and J. L. Torrecilla. On the use of reproducing kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association (to appear)*, 2017. doi: 10.1080/01621459.2017.1320287. URL `http://dx.doi.org/10.1080/01621459.2017.1320287`.

J. R. Berrendero, B. Bueno-Larraz, and A. Cuevas. An RKHS model for variable selection in functional linear regression. *Journal of Multivariate Analysis (to appear)*, 2018a. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2018.04.008. URL `http://www.sciencedirect.com/science/article/pii/S0047259X17305596`.

J. R. Berrendero, B. Bueno-Larraz, and A. Cuevas. On Mahalanobis distance in functional settings. *arXiv preprint, arXiv:1803.06550*, 2018b.

P. Besse and H. Cardot. Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics*, 24(4):467–487, 1996.

G. Biau, B. Cadre, and Q. Paris. Cox process functional learning. *Statistical Inference for Stochastic Processes*, 18(3):257–277, 2015.

C. M. Bishop. *Patterns Recognition and Machine Learning*. Springer-Verlag New York, 2016.

E. G. Bongiorno and A. Goia. Some insights about the small ball probability factorization for Hilbert random elements. *Statistica Sinica*, 27(4):1949–1965, 2017.

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, New York, 2000.

D. Bosq and D. Blanke. *Inference and Prediction in Large Dimensions*. Wiley-Dunod, Chichester, 2007.

B. Bueno-Larraz and J. Klepsch. Variable selection for the prediction of C[0, 1]-valued autoregressive processes using reproducig kernel Hilbert spaces. *Technometrics (to appear)*, 2018. doi: https://doi.org/10.1080/00401706.2018.1505660.

T. T. Cai, P. Hall, et al. Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179, 2006.

S. Cambanis. Sampling designs for time series. *Handbook of Statistics*, 5:337–362, 1985.

E. J. Candès and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint, arXiv:1804.09753*, 2018.

A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9(Jul):1615–1646, 2008.

H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41, 2005.

H. Cardot and P. Sarda. Functional linear regression. In F. Ferraty and Y. Romain, editors, *Handbook of Functional Data Analysis*, pages 21–46. Oxford University Press, Oxford, 2010.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

R. J. Carroll, A. Delaigle, and P. Hall. Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier. *The Annals of Statistics*, 41(6):2739–2767, 2013.

J. B. Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer Science & Business Media, 2 edition, 1990.

J. S. Cramer. *Logit Models from Economics and Other Fields*. Cambridge University Press, 2003.

F. Cucker and S. Smale. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2001.

F. Cucker and D. X. Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

A. Cuevas, M. Febrero, and R. Fraiman. Linear functional regression: The case of fixed design and functional response. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(2):285–300, 2002. ISSN 03195724. URL `http://www.jstor.org/stable/3315952`.

X. Dai, H.-G. Müller, and F. Yao. Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 104(3):545–560, 2017.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, 1982.

A. Delaigle and P. Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286, 2012.

A. Delaigle, P. Hall, and N. Bathia. Componentwise classification and clustering of functional data. *Biometrika*, 99(2):299–313, 2012.

D. Didericksen, P. Kokoszka, and X. Zhang. Empirical properties of forecasts with the functional autoregressive model. *Computational Statistics*, 27(2):285–298, 2012.

B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

M. Escabias, A. M. Aguilera, and M. J. Valderrama. Principal component estimation of functional logistic regression: discussion of two different approaches. *Journal of Nonparametric Statistics*, 16(3-4):365–384, 2004.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1): 342–368, 1985.

L. Fahrmeir and H. Kaufmann. Asymptotic inference in discrete response models. *Statistische Hefte*, 27(1):179–205, 1986.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

M. Febrero-Bande and M. de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software, Articles*, 51(4):1–28, 2012. doi: 10.18637/jss.v051.i04. URL `https://www.jstatsoft.org/v051/i04`.

J. Feldman. Equivalence and perpendicularity of Gaussian processes. *Pacific Journal of Mathematics*, 8(4):699–708, 1958.

F. Ferraty and Y. Romain. *The Oxford Handbook of Functional Data Analaysis*. Oxford University Press, 2011.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media, 2006.

F. Ferraty, P. Hall, and P. Vieu. Most-predictive design points for functional data predictors. *Biometrika*, 97(4):807–824, 2010.

F. Ferraty, I. Van Keilegom, and P. Vieu. Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109:10–28, 2012.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

R. Fraiman, Y. Gimenez, and M. Svarc. Feature selection for functional data. *Journal of Multivariate Analysis*, 146:191–208, 2016.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/v033/i01`.

P. Galeano, E. Joseph, and R. E. Lillo. The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291, 2015.

A. Ghiglietti and A. Paganoni. Statistical inference for functional data based on a generalization of Mahalanobis distance. Technical report, Department of Mathematics, Politecnico di Milano, 2014. MOX–Report No. 39/2014.

A. Ghiglietti and A. M. Paganoni. Exact tests for the means of Gaussian stochastic processes. *Statistics & Probability Letters*, 131:102–107, 2017.

A. Ghiglietti, F. Ieva, and A. M. Paganoni. Statistical inference for stochastic processes: two-sample hypothesis tests. *Journal of Statistical Planning and Inference*, 180:49–68, 2017.

I. Gohberg and S. Goldberg. *Basic Operator Theory*. Birkhäuser, 2013.

C. Gourieroux and A. Monfort. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics*, 17(1):83–97, 1981.

S. Graves, G. Hooker, and J. O. Ramsay. *Functional Data Analysis with R and MATLAB*. Springer, 2009.

B. Gregorutti. *RFgroove: Importance Measure and Selection for Groups of Variables with Random Forests*, 2016. R package version 1.1.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520, 2007.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, 2008.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, pages 673–681, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

G. Heinze and M. Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.

J. M. Hilbe. *Logistic Regression Models*. CRC Press, 2009.

S. Hörmann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010.

S. Hörmann, L. Kidziński, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 77(2):319–348, 2015.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer Science & Business Media, 2012.

D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 3 edition, 2013.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

T. Hsing and H. Ren. An RKHS formulation of the inverse regression dimension-reduction problem. *Annals of Statistics*, 37(2):726–755, 04 2009. doi: 10.1214/07-AOS589. URL `http://dx.doi.org/10.1214/07-AOS589`.

R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with Exponential Smoothing: the State Space Approach*. Springer Science & Business Media, 2008.

R. J. Hyndman. *expsmooth: Data sets from "Exponential smoothing: a state space approach" by Hyndman, Koehler, Ord and Snyder (Springer, 2008)*, 2018. URL `http://pkg.robjhyndman.com/expsmooth`. R package version 2.4.

A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. Springer, 2008.

G. M. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, 2002.

S. Janson. *Gaussian Hilbert Spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge university press, 1997.

H. Ji and H.-G. Müller. Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):859–876, 2017.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In D. Precup and Y. W. Teh, editors, *ICML*, volume 70, 2017. URL `http://proceedings.mlr.press/v70/jitkrittum17a.html`.

H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016. URL `http://jmlr.org/papers/v17/11-315.html`.

V. Kargin and A. Onatski. Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99:2508–2526, 2008.

T. Kato. *Perturbation Theory for Linear Operators*. Springer Science & Business Media, 2013.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

J. Klepsch and C. Klüppelberg. An innovations algorithm for the prediction of functional linear processes. *Journal of Multivariate Analysis*, 155:252–271, 2017.

J. Klepsch, C. Klüppelberg, and T. Wei. Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics*, 1:128–149, 2017.

A. Kneip, D. Poss, and P. Sarda. Funcional linear regression with points of impact. *Annals of Statistics*, 44(1):1–30, 2016.

P. Kokoszka and M. Reimherr. Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34(1):116–129, 2013.

P. Kokoszka, I. Maslova, J. Sojka, and L. Zhu. Testing for lack of dependence in the functional linear model. *Canadian Journal of Statistics*, 36(2):207–222, 2008.

I. Kosmidis. *brglm: Bias Reduction in Binary-Response Generalized Linear Models*, 2017. URL `http://www.ucl.ac.uk/~ucakiko/software.html`. R package version 0.6.1.

R. G. Laha and V. K. Rohatgi. *Probability Theory*. John Wiley & Sons, New York-Chichester-Brisbane, Wiley Series in Probability and Mathematical Statistics, 1979.

C. Lam and Q. Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

M. A. Lindquist and I. W. McKeague. Logistic regression with Brownian-like predictors. *Journal of the American Statistical Association*, 104(488):1575–1585, 2009.

Z. Z. Liu. *The Doubly Adaptive LASSO Methods for Time Series Analysis*. PhD dissertation, University of Western Ontario, 2014.

M. N. Lukić and J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10): 3945–3969, 2001. ISSN 00029947. URL `http://www.jstor.org/stable/2693779`.

P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

K. V. Mardia. Assessment of multinormality and the robustness of Hotelling's t 2 test. *Applied Statistics*, 24(2):163–171, 1975.

P. McCullagh and J. Nelder. *Generalized Linear Models*. CRC Press, 1989.

I. W. McKeague and B. Sen. Fractals with point impact in functional linear regression. *Annals of Statistics*, 38(4):2559–2586, 08 2010. doi: 10.1214/10-AOS791. URL `http://dx.doi.org/10.1214/10-AOS791`.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

A. J. Miller. *Subset Selection in Regression*, volume 95 of *Monographs on statistics and applied probability*. Chapman & Hall/CRC, Boca Raton, 2002. ISBN 1-58488-171-2. URL `http://opac.inria.fr/record=b1131181`.

F. Mokhtari and T. Mourid. Prediction of continuous time autoregressive processes via the reproducing kernel spaces. *Statistical Inference for Stochastic Processes*, 6(3): 247–266, 2003.

J. S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2: 321–359, 2015.

P. Mörters and Y. Peres. *Brownian Motion*, volume 30 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2010.

S. N. Mousavi and H. Sørensen. Multinomial functional regression with wavelets and LASSO penalization. *Econometrics and Statistics*, 1:150–166, 2017.

S. N. Mousavi and H. Sørensen. Functional logistic regression: a comparison of three methods. *Journal of Statistical Computation and Simulation*, 88(2):250–268, 2018.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25*, pages 10–18, 2012.

H.-G. Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33(2):774–805, 2005.

H.-G. Müller and F. Yao. Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544, 2008.

B. Munsiwamy and S. T. Wakweya. Asymptotic properties of estimates for the parameters in the logistic regression model. *Asian-African Journal of Economics and Econometrics*, 11(1):165–174, 2011.

E. Parzen. Statistical inference on time series by Hilbert space methods, I. Technical Report 23, Stanford University, 1959.

E. Parzen. Regression analysis of continuous parameter time series. In *Proceedings of the Fourth Berkeley Symposion on Mathematical Statistics and Probability*, volume 1, pages 469–489, 1961a.

E. Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, 32(4):951–989, 1961b.

K. I. Penny. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(1):73–81, 1996. ISSN 00359254, 14679876. URL `http://www.jstor.org/stable/2986224`.

S. Peszat and J. Zabczyk. *Stochastic Partial Differential Equations with Lévy Noise: An Evolution Equation Approach*, volume 113 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 2007.

J. Petrovich, M. Reimherr, and C. Daymont. Functional regression models with highly irregular designs. *arXiv preprint, arXiv:1805.08518*, 2018.

N. S. Pillai, Q. Wu, F. Liang, S. Mukherjee, and R. L. Wolpert. Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research*, 8 (Aug):1769–1797, 2007.

S. Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16(1):356–366, 1988.

C. Preda. Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137(3):829–840, 2007.

B. Pumo. Prediction of continuous time processes by $C[0, 1]$–valued autoregressive process. *Statistical Inference for Stochastic Processes*, 1(3):297–309, 1998. doi: 10.1023/A:1009951104780.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

J. Ramsay. When the data are functions. *Psychometrika*, 47(4):379–396, 1982.

J. Ramsay and B. Silverman. *Applied Functional Data Analysis*. Springer Series in Statistics, 2002.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics, 2005.

C. R. Rao and V. Varadarajan. Discrimination of Gaussian processes. *Sankhyā: The Indian Journal of Statistics, Series A*, 25(3):303–330, 1963.

M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Functional Analysis*. Academic New York, 1980.

A. C. Rencher. *Methods of Multivariate Analysis*. John Wiley & Sons, 3 edition, 2012.

F. Riesz and B. Szökefalvi-Nagy. *Functional Analysis*. Dover Publications, 1990.

B. Ripley and W. Venables. *class: Functions for Classification*, 2015. R package version 7.3-14.

P. J. Rousseeuw and B. C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990. doi: 10.1080/01621459.1990.10474920. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474920`.

M. Ruiz-Meana, D. Garcia-Dorado, P. Pina, J. Inserte, L. Agulló, and J. Soler-Soler. Cariporide preserves mitochondrial proton gradient and delays ATP depletion in cardiomyocytes during ischemic conditions. *American Journal of Physiology-Heart and Circulatory Physiology*, 285(3):H999–H1006, 2003.

M. Ruiz-Medina and J. Álvarez-Liébana. Strong-consistent autoregressive predictors in abstract Banach spaces. *Journal of Multivariate Analysis (to appear)*, 2018.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. J. Smola. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017, 1999.

A. Segall and T. Kailath. Radon-Nikodym derivatives with respect to measures induced by discontinuous independent-increment processes. *The Annals of Probability*, 3(3): 449–464, 1975.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

L. A. Shepp. Radon-Nikodym derivatives of Gaussian measures. *The Annals of Mathematical Statistics*, 37(2):321–354, 1966.

J. Q. Shi and T. Choi. *Gaussian Process Regression Analysis for Functional Data*. CRC Press, 2011.

M. J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(3):310–313, 1981.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer Berlin Heidelberg, 2007.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statististics*, 35(6):2769–2794, 12 2007. doi: 10.1214/009053607000000505. URL `http://dx.doi.org/10.1214/009053607000000505`.

H. Tran, N. Muttil, and B. Perera. Selection of significant input variables for time series forecasting. *Environmental Modelling & Software*, 64:156–163, 2015.

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2000.

D. Varberg. On equivalence of Gaussian measures. *Pacific Journal of Mathematics*, 11 (2):751–762, 1961.

D. E. Varberg. On Gaussian measures equivalent to Wiener measure. *Transactions of the American Mathematical Society*, 113(2):262–273, 1964.

J. J. Verbeek, N. Vlassis, and B. Kröse. Efficient greedy learning of Gaussian mixture models. *Neural Computation*, 15(2):469–485, 2003.

F. Yao, H.-G. Müller, J.-L. Wang, et al. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

C. D. Yenigün and M. L. Rizzo. Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation*, 85(8):1692–1705, 2015.

M. Yuan and T. T. Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

F. Zamora-Martínez, P. Romeu, P. Botella-Rocamora, and J. Pardo. On-line learning of indoor temperature forecasting models towards energy efficiency. *Energy and Buildings*, 83:162–172, 2014. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2014.04.034.

X. Zhao, J. Marron, and M. T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14(3):789–808, 2004.

Y. Zhao, R. T. Ogden, and P. T. Reiss. Wavelet-based LASSO in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617, 2012.

Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.