Journal of Machine Learning Research 17 (2016) 1-54 Submitted 10/11; Revised 8/15; Published 4/16

Operator-valued Kernels for Learning from Functional Response Data

Hachem Kadri HACHEM.KADRI@LIF.UNIV-MRS.FR
Aix-Marseille Université, LIF (UMR CNRS 7279)
F-13288 Marseille Cedex 9, France

Emmanuel Duflos EMMANUEL.DUFLOS@EC-LILLE.FR
Ecole Centrale de Lille, CRIStAL (UMR CNRS 9189)
59650 Villeneuve d'Ascq, France

Philippe Preux PHILIPPE.PREUX@UNIV-LILLE3.FR
Université de Lille, CRIStAL (UMR CNRS 9189)
59650 Villeneuve d'Ascq, France

Stéphane Canu SCANU@INSA-ROUEN.FR
INSA de Rouen, LITIS (EA 4108)
76801, St Etienne du Rouvray, France

Alain Rakotomamonjy ALAIN.RAKOTOMAMONJY@INSA-ROUEN.FR
Université de Rouen, LITIS (EA 4108)
76801, St Etienne du Rouvray, France

Julien Audiffren JULIEN.AUDIFFREN@CMLA.ENS-CACHAN.FR
ENS Cachan, CMLA (UMR CNRS 8536)
94235 Cachan Cedex, France

Editor: John Shawe-Taylor

Abstract

In this paper[1] we consider the problems of supervised classification and regression in the case where attributes and labels are functions: a data is represented by a set of functions, and the label is also a function. We focus on the use of reproducing kernel Hilbert space theory to learn from such functional data. Basic concepts and properties of kernel-based learning are extended to include the estimation of function-valued functions. In this setting, the representer theorem is restated, a set of rigorously defined infinite-dimensional operator-valued kernels that can be valuably applied when the data are functions is described, and a learning algorithm for nonlinear functional data analysis is introduced. The methodology is illustrated through speech and audio signal processing experiments.

Keywords: nonlinear functional data analysis, operator-valued kernels, function-valued reproducing kernel Hilbert spaces, audio signal processing

1. Introduction

In this paper, we consider the supervised learning problem in a functional setting: each attribute of a data is a function, and the label of each data is also a function. For the sake

1. This is a combined and expanded version of previous conference papers (Kadri et al., 2010, 2011c).

of simplicity, one may imagine real functions, though the work presented here is much more general; one may also think about those functions as being defined over time, or space, though again, our work is not tied to such assumptions and is much more general. To this end, we extend the traditional scalar-valued attribute setting to a function-valued attribute setting.

This shift from scalars to functions is required by the simple fact that in many applications, attributes are functions: functions may be one dimensional such as economic curves (variation of the price of "actions"), load curve of a server, a sound, etc., or two or higher dimensional (hyperspectral images, etc.). Due to the nature of signal acquisition, one may consider that in the end, a signal is always acquired in a discrete fashion, thus providing a real vector. However, with the resolution getting finer and finer in many sensors, the amount of discrete data is getting huge, and one may reasonably wonder whether a functional point of view may not be better than a vector point of view. Now, if we keep aside the application point of view, the study of functional attributes may simply come as an intellectual question which is interesting for its own sake.

From a mathematical point of view, the shift from scalar attributes to function attributes will come as a generalization from scalar-valued functions to function-valued functions, a.k.a."operators". Reproducing Kernel Hilbert Spaces (RKHS) has become a widespread tool to deal with the problem of learning a function mapping the set $\mathbb{R}^p$ to the set of real numbers $\mathbb{R}$. Here, we have to deal with RKHS of operators, that are functions that map a function, belonging to a certain space of functions, to a function belonging to an other space of functions. This shift in terminology is accompanied with a dramatic shift in concepts, and technical difficulties that have to be properly handled.

This *functional regression problem*, or *functional supervised learning*, is a challenging research problem, from statistics to machine learning. Most previous work has focused on the discrete case: the multiple-response (finite and discrete) function estimation problem. In the machine learning literature, this problem is better known under the name of vector-valued function learning (Micchelli and Pontil, 2005a), while in the field of statistics, researchers prefer to use the term multiple output regression (Breiman and Friedman, 1997). One possible solution is to approach the problem from a univariate point of view, that is, assuming only a single response variable output from the same set of explanatory variables. However it would be more efficient to take advantage of correlation between the response variables by considering all responses simultaneously. For further discussion of this point, we refer the reader to Hastie et al. (2001) and references therein. More recently, relevant works in this context concern regularized regression with a sequence of ordered response variables. Many variable selection and shrinkage methods for single response regression are extended to the multiple response data case and several algorithms following the corresponding solution paths are proposed (Turlach et al., 2005; Simila and Tikka, 2007; Hesterberg et al., 2008).

Learning from multiple responses is closely related to the problem of multi-task learning where the goal is to improve generalization performance by learning multiple tasks simultaneously. There is a large literature on this subject, in particular Evgeniou and Pontil (2004); Jebara (2004); Ando and Zhang (2005); Maurer (2006); Ben-david and Schuller-Borbely (2008); Argyriou et al. (2008) and references therein. One paper that has come to our attention is that of Evgeniou et al. (2005) who showed how Hilbert spaces of vector-valued

functions (Micchelli and Pontil, 2005a) and matrix-valued reproducing kernels (Micchelli and Pontil, 2005b; Reisert and Burkhardt, 2007) can be used as a theoretical framework to develop nonlinear multi-task learning methods.

A primary motivation for this paper is to build on these previous studies and provide a similar framework for addressing the general case where the output space is infinite dimensional. In this setting, the output space is a space of functions and elements of this space are called functional response data. Functional responses are frequently encountered in the analysis of time-varying data when repeated measurements of a continuous response variable are taken over a small period of time (Faraway, 1997; Yao et al., 2005). The relationships among the response data are difficult to explore when the number of responses is large, and hence one might be inclined to think that it could be helpful and more natural to consider the response as a smooth real function. Moreover, with the rapid development of accurate and sensitive instruments and thanks to the currently available large storage resources, data are now often collected in the form of curves or images. The statistical framework underlying the analysis of these data as a single function observation rather than a collection of individual observations is called functional data analysis (FDA) and was first introduced by Ramsay and Dalzell (1991).

It should be pointed out that in earlier studies a similar but less explicit statement of the functional approach was addressed in Dauxois et al. (1982), while the first discussion of what is meant by "functional data" appears to be by Ramsay (1982). Functional data analysis deals with the statistical description and modeling of random functions. For a wide range of statistical tools, ranging from exploratory and descriptive data analysis to linear models and multivariate techniques, a functional version has been recently developed. Reviews of theoretical concepts and prospective applications of functional data can be found in the two monographs by Ramsay and Silverman (2005, 2002). One of the most crucial questions related to this field is "What is the correct way to handle large data? Multivariate or Functional?" Answering this question requires better understanding of complex data structures and relationship among variables. Until now, arguments for and against the use of a functional data approach have been based on methodological considerations or experimental investigations (Ferraty and Vieu, 2003; Rice, 2004). However, we believe that without further improvements in theoretical issues and in algorithm design of functional approaches, exhaustive comparative studies will remain hard to conduct.

This motivates the general framework we develop in this paper. To the best of our knowledge, nonlinear methods for functional data is a topic that has not been sufficiently addressed in the FDA literature. Unlike previous studies on nonlinear supervised classification or real response regression of functional data (Rossi and Villa, 2006; Ferraty and Vieu, 2004; Preda, 2007), this paper addresses the problem of learning tasks where the output variables are functions. From a machine learning point of view, the problem can be viewed as that of learning a function-valued function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ where $\mathcal{X}$ is the input space and $\mathcal{Y}$ the (possibly infinite-dimensional) Hilbert space of the functional output data. Various situations can be distinguished according to the nature of input data attributes (scalars or/and functions). We focus in this work on the case where input attributes are functions, too, but it should be noted that the framework developed here can also be applied when the input data are either discrete, or continuous. Lots of practical applications involve a blend of both functional and non functional attributes, but we do not mix non functional

attributes with functional attributes in this paper. This point has been discussed in (Kadri et al., 2011b). To deal with non-linearity, we adopt a kernel-based approach and we design operator-valued kernels that perform the mapping between the two spaces of functions. Our main results demonstrate how basic concepts and properties of kernel-based learning known in the case of multivariate data can be restated for functional data.

Extending learning methods from multivariate to functional response data may lead to further progress in several practical problems of machine learning and applied statistics. To compare the proposed nonlinear functional approach with other multivariate or functional methods and to apply it in a real world setting, we are interested in the problems of speech inversion and sound recognition, which have attracted increasing attention in the speech processing community in the recent years (Mitra et al., 2010; Rabaoui et al., 2008). These problems can be cast as a supervised learning problem which include some components (predictors or responses) that may be viewed as random curves. In this context, though some concepts on the use of RKHS for functional data similar to those presented in this work can be found in Lian (2007), the present paper provides a much more complete view of learning from functional data using kernel methods, with extended theoretical analysis and several additional experimental results.

This paper is a combined and expanded version of our previous conference papers (Kadri et al., 2010, 2011c). It gives the full justification, additional insights as well as new and comprehensive experiments that strengthen the results of these preliminary conference papers. The outline of the paper is as follows. In Section 2, we discuss the connection between the two fields Functional Data Analysis and Machine Learning, and outline our main contributions. Section 3 defines the notation used throughout the paper. Section 4, describes the theory of reproducing kernel Hilbert spaces of function-valued functions and shows how vector-valued RKHS concepts can be extended to infinite-dimensional output spaces. In Section 5, we exhibit a class of operator-valued kernels that perform the mapping between two spaces of functions and discuss some ideas for understanding their associated feature maps. In Section 6, we provide a function-valued function estimation procedure based on inverting block operator kernel matrices, propose a learning algorithm that can handle functional data, and analyze theoretically its generalization properties. Finally in Section 7, we illustrate the performance of our approach through speech and audio processing experiments.

2. The Interplay of FDA and ML Research

To put our work in context, we begin by discussing the interaction between functional data analysis (FDA) and machine learning (ML). Then, we give an overview of our contributions.

Starting from the fact that "new types of data require new tools for analysis", FDA emerges as a well-defined and suitable concept to further improve classical multivariate statistical methods when data are functions (Levitin et al., 2007). This research field is currently very active, and considerable progress has been made in recent years in designing statistical tools for infinite-dimensional data that can be represented by real-valued functions rather than by discrete, finite dimensional vectors (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Shi and Choi, 2011; Horváth and Kokoszka, 2012). While the FDA viewpoint is conventionally adopted in the mathematical statistics community to deal

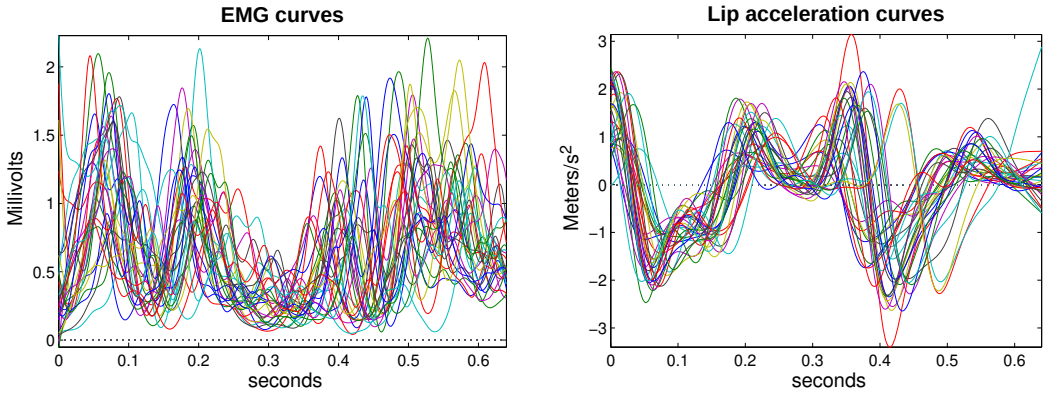

Figure 1: Electromyography (EMG) and lip acceleration curves. The left panel displays EMG recordings from a facial muscle that depresses the lower lip, the depressor labii inferior. The right panel shows the accelerations of the center of the lower lip of a speaker pronouncing the syllable "bob", embedded in the phrase "Say bob again", for 32 replications (Ramsay and Silverman, 2002, Chapter 10).

with data in infinite-dimensional spaces, it does not appear to be commonplace for machine learners. One possible reason for this lack of success is that the formal use of infinite dimensional spaces for practical ML applications may seem unjustified; because in practice traditional measurement devices are limited in providing discrete and not functional data, and a machine learning algorithm can process only finitely represented objects. We believe that for applied machine learners it should be vital to know the full range of applicability of functional data analysis and infinite-dimensional data representations. But due to limitation of space we shall say only few words about the occurrence of functional data in real applications and about the real learning task lying behind this kind of approach. The reader is referred to Ramsay and Silverman (2002) for more details and references. Areas of application discussed and cited there include medical diagnosis, economics, meteorology, biomechanics, and education. For almost all these applications, the high-sampling rate of today's acquisition devices makes it natural to directly handle functions/curves instead of discretized data. Classical multivariate statistical methods may be applied to such data, but they cannot take advantage of the additional information implied by the smoothness of the underlying functions. FDA methods can have beneficial effects in this direction by extracting additional information contained in the functions and their derivatives, not normally available through traditional methods (Levitin et al., 2007).

To get a better idea about the natural occurrence of functional data in ML tasks, Figure 1 depicts a functional data set introduced by Ramsay and Silverman (2002). The data set consists of 32 records of the movement of the center of the lower lip when a subject was repeatedly required to say the syllable "bob", embedded in the sentence, "Say bob again" and the corresponding EMG activities of the primary muscle depressing the lower lip, the depressor labii inferior (DLI)[2]. The goal here is to study the dependence of the acceleration

2. The data set is available at `http://www.stats.ox.ac.uk/~silverma/fdacasebook/lipemg.html`. More information about the data collection process can be found in Ramsay and Silverman (2002, Chapter 10).

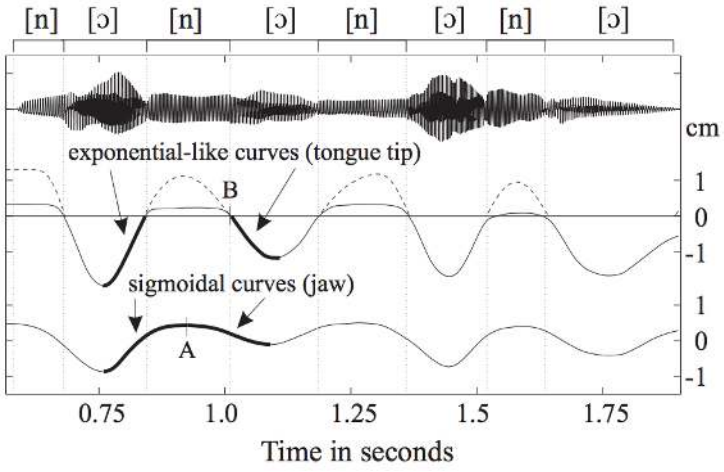

Figure 2: "Audio signal (top), tongue tip trajectory (middle), and jaw trajectory (bottom) for the utterance [ˌnɔnɔˈnɔnɔ]. The trajectories were measured by electromagnetic articulography (EMA) for coils on the tongue tip and the lower incisors. Each trajectory shows the displacement along the first principal component of the original two-dimensional trajectory in the midsagittal plane. The dashed curves show hypothetical continuations of the tongue tip trajectory towards and away from virtual targets during the closure intervals." (Birkholz et al., 2010).

of the lower lip in speech on neural activity. EMG and lip accelerations curves can be well modeled by continuous functions of time that allow to capture functional dependencies and interactions between samples (feature values). Thus, we face a regression problem where both input and output data are functions. In much the same way, Figure 2 also shows a "natural" representation of data in terms of functions [3]. It represents a speech signal used for acoustic-articulatory speech inversion and produced by a subject pronouncing a sequence of [ˌCVCVˈCVCV] (C=consonant, V=vowel) by combining the vowel {/ɔ/} with the consonant {/n/}. The articulatory trajectories are represented by the upper and lower solid curves that show the displacement of fleshpoints on the tongue tip and the jaw along the main movement direction of these points during the repeated opening and closing gestures. This example is from a recent study on the articulatory modeling of speech signals (Birkholz et al., 2010). The concept of articulatory gestures in the context of speech-to-articulatory inversion will be explained in more details in Section 7. As shown in the figure, the observed articulatory trajectories are typically modeled by smooth functions of time with periodicity properties and exponential or sigmoidal shape, and the goal of speech inversion is to predict and recover geometric data of the vocal tract from the speech information.

In both examples given above, response data clearly present a functional behavior that should be taken into account during the learning process. We think that handling these data as what they really are, that is functions, is a promising way to tackle prediction problems and design efficient ML systems for continuous data variables. Moreover, ML methods which

3. This figure is from Birkholz et al. (2010).

can handle functional features can open up plenty of new areas of application, where the flexibility of functional and infinite-dimensional spaces would allow to enable us to achieve significantly better performance while managing huge amounts of training data.

In the light of these observations, there is an interest in overcoming methodological and practical problems that hinder the wide adoption and use of functional methods built for infinite-dimensional data. Regarding the practical issue related to the application and implementation of infinite-dimensional spaces, a standard means of addressing it is to choose a functional space *a priori* with a known predefined set of basis functions in which the data will be mapped. This may include a preprocessing step, which consists in converting the discretized data into functional objects using interpolation or approximation techniques. Following this scheme, parametric FDA methods have emerged as a common approach to extend multivariate statistical analysis in functional and infinite-dimensional situations (Ramsay and Silverman, 2005). More recently, nonparametric FDA methods have received increasing attention because of their ability to avoid fixing a set of basis functions for the functional data beforehand (Ferraty and Vieu, 2006). These methods are based on the concept of semi-metrics for modeling functional data. The reason for using a semi-metric rather than a metric is that the coincidence axiom, namely $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$, may result in curves with very similar shapes being categorized as distant (not similar to each other). To define closeness between functions in terms of shape rather than location semi-metrics can be used. In this spirit, Ferraty and Vieu (2006) provided a semi-metric based methodology for nonparametric functional data analysis and argued that this can be a sufficiently general theoretical framework to tackle infinite-dimensional data without being "too heavy" in terms of computational time and implementation complexity.

Thus, although both parametric and nonparametric functional data analyses deal with infinite-dimensional data, they are computationally feasible and quite practical since the observed functional data are approximated in a basis of the function space with possibly finite number of elements. What we really need is the inner or semi-inner product of the basis elements and the representation of the functions with respect to that basis. We think that Machine Learning research can profit from exploring other representation formalisms that support the expressive power of functional data. Machine learning methods which can accommodate functional data should open up new possibilities for handling practical applications for which the flexibility of infinite-dimensional spaces could be exploited to achieve performance benefits and accuracy gains. On the other hand, in the FDA field, there is clearly a need for further development of computationally efficient and understandable algorithms that can deliver near-optimal solutions for infinite-dimensional problems and that can handle a large number of features. The transition from infinite-dimensional statistics to efficient algorithmic design and implementation is of central importance to FDA methods in order to make them more practical and popular. In this sense, Machine Learning can have a profound impact on FDA research.

In reality, ML and FDA have more in common than it might seem. There are already existing machine learning algorithms that can also be viewed as FDA methods. For example, these include kernel methods which use a certain type of similarity measure (called a kernel) to map observed data in a high dimensional feature space, in which linear methods are used for learning problems (Shawe-Taylor and Cristanini, 2004; Schölkopf and Smola,

2002). Depending on the choice of the kernel function, the feature space can be infinite-dimensional. The kernel trick is used, allowing to work with finite Gram matrix of inner products between the possibly infinite-dimensional features which can be seen as functional data. This connection between kernel and FDA methods is clearer with the concept of kernel embedding of probability distributions, where, instead of (observed) single points, kernel means are used to represent probability distributions (Smola et al., 2007; Sriperumbudur et al., 2010). The kernel mean corresponds to a mapping of a probability distribution in a feature space which is rich enough so that its expectation uniquely identifies the distribution. Thus, rather than relying on large collections of vector data, kernel-based learning can be adapted to probability distributions that are constructed to meaningfully represent the discrete data by the use of kernel means (Muandet et al., 2012). In some sense, this represents a similar design to FDA methods, where data are assumed to lie in a functional space even though they are acquired in a discrete manner. There are also other papers that deal with machine learning problems where covariates are probability distributions and discuss their relation with FDA (Poczos et al., 2012, 2013; Oliva et al., 2013). At that point, however, the connection between ML and FDA is admittedly weak and needs to be bolstered by the delivery of more powerful and flexible learning machines that are able to deal with functional data and infinite-dimensional spaces.

In the FDA field, linear models have been explored extensively. Nonlinear modeling of functional data is, however, a topic that has not been sufficiently investigated, especially when response data are functions. Reproducing kernels provide a powerful tool for solving learning problems with nonlinear models, but to date they have been used more to learn scalar-valued or vector-valued functions than function-valued functions. Consequently, kernels for functional response data and their associated function-valued reproducing kernel Hilbert spaces have remained mostly unknown and poorly studied. In this work, we aim to rectify this situation, and highlight areas of overlap between the two fields FDA and ML, particularly with regards to the applicability and relevance of the FDA paradigm coupled with machine learning techniques. Specifically, we provide a learning methodology for nonlinear FDA based on the theory of reproducing kernels. The main contributions are as follows:

- we introduce a set of rigorously defined operator-valued kernels suitable for functional response data, that can be valuably applied to model dependencies between samples and take into account the functional nature of the data, like the smoothness of the curves underlying the discrete observations,

- we propose an efficient algorithm for learning function-valued functions (operators) based on the spectral decomposition of block operator matrices,

- we study the generalization performance of our learned nonlinear FDA model using the notion of algorithmic stability,

- we show the applicability and suitability of our framework to two problems in audio signal processing, namely speech inversion and sound recognition, where features are functions that are dependent on each other.

3. Notations and Conventions

We start by some standard notations and definitions used all along the paper. Given a Hilbert space $\mathcal{H}$, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ refer to its inner product and norm, respectively. $H^n = \underbrace{\mathcal{H} \times \ldots \times \mathcal{H}}_{n \text{ times}}$, $n \in \mathbb{N}_+$, denotes the topological product of n spaces $\mathcal{H}$. We denote by $\mathcal{X} = \{x : \Omega_x \longrightarrow \mathbb{R}\}$ and $\mathcal{Y} = \{y : \Omega_y \longrightarrow \mathbb{R}\}$ the separable Hilbert spaces of input and output real-valued functions whose domains are Ω_x and Ω_y, respectively. In functional data analysis domain, the space of functions is generally assumed to be the Hilbert space of equivalence classes of square integrable functions, denoted by L^2. Thus, in the rest of the paper, we consider $\mathcal{Y}$ to be the space $L^2(\Omega_y)$, where Ω_y is a compact set. The vector space of functions from $\mathcal{X}$ into $\mathcal{Y}$ is denoted by $\mathcal{Y}^{\mathcal{X}}$ endowed with the topology of uniform convergence on compact subsets of $\mathcal{X}$. We denote by $\mathcal{C}(\mathcal{X}, \mathcal{Y})$ the vector space of continuous functions from $\mathcal{X}$ to $\mathcal{Y}$, by $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ the Hilbert space of function-valued functions $F : \mathcal{X} \longrightarrow \mathcal{Y}$, and by $\mathcal{L}(\mathcal{Y})$ the set of bounded linear operators from $\mathcal{Y}$ to $\mathcal{Y}$.

We now fix the following conventions for bounded linear operators and block operator matrices.

Definition 1 *(adjoint, self-adjoint, and positive operators)*
Let $A \in \mathcal{L}(\mathcal{Y})$. Then:

(i) *A^*, the adjoint operator of A, is the unique operator in $\mathcal{L}(\mathcal{Y})$ that satisfies*

$$\langle Ay, z \rangle_{\mathcal{Y}} = \langle y, A^* z \rangle_{\mathcal{Y}} \ , \ \forall y \in \mathcal{Y}, \forall z \in \mathcal{Y},$$

(ii) *A is self-adjoint if $A = A^*$,*

(iii) *A is positive if it is self-adjoint and $\forall y \in \mathcal{Y}$, $\langle Ay, y \rangle_{\mathcal{Y}} \geq 0$ (we write $A \geq 0$),*

(iv) *A is larger or equal than $B \in \mathcal{L}(\mathcal{Y})$, if $A - B$ is positive, i.e., $\forall y \in \mathcal{Y}$, $\langle Ay, y \rangle_{\mathcal{Y}} \geq \langle By, y \rangle_{\mathcal{Y}}$ (we write $A \geq B$).*

Definition 2 *(block operator matrix)*
Let $n \in \mathbb{N}$, let $\mathcal{Y}^n = \underbrace{\mathcal{Y} \times \ldots \times \mathcal{Y}}_{n \text{ times}}$.

(i) *$\mathbf{A} \in \mathcal{L}(\mathcal{Y}^n)$, given by*

$$\mathbf{A} = \begin{pmatrix} A_{11} & \ldots & A_{1n} \\ \vdots & & \vdots \\ A_{n1} & \ldots & A_{nn} \end{pmatrix}$$

where each $A_{ij} \in \mathcal{L}(\mathcal{Y})$, $i, j = 1, \ldots, n$, is called a block operator matrix,

(ii) *the adjoint (or transpose) of $\mathbf{A}$ is the block operator matrix $\mathbf{A}^* \in \mathcal{L}(\mathcal{Y}^n)$ such that $(A^*)_{ij} = (A_{ji})^*$,*

(iii) *self-adjoint and order relations of block operator matrices are defined in the same way as for bounded operators (see Definition 1).*

real numbers	α, β, γ, ...	Greek characters
integers	i, j, m, n	
vector spaces[4]	$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{H}$, ...	Calligraphic letters
subsets of the real plain	Ω, Λ, Γ, ...	capital Greek characters
functions[5](or vectors)	x, y, f, ...	small Latin characters
vector of functions	$\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, ...	small bold Latin characters
operators (or matrices)	A, B, K, ...	capital Latin characters
block operator matrices	$\mathbf{A}$, $\mathbf{B}$, $\mathbf{K}$, ...	capital bold Latin characters
adjoint operator	$*$	A^* adjoint of operator A
identical equality	$\equiv$	equality of mappings
definition	$\triangleq$	equality by definition

Table 1: Notations used in this paper.

Note that item (ii) in Definition 2 is obtained from the definition of adjoint operator. It is easy to see that $\forall \mathbf{y} \in \mathcal{Y}^n$ and $\forall \mathbf{z} \in \mathcal{Y}^n$; we have: $\langle \mathbf{A}\mathbf{y}, \mathbf{z} \rangle_{\mathcal{Y}^n} = \sum_{i,j} \langle A_{ij} y_j, z_i \rangle_{\mathcal{Y}} = \sum_{i,j} \langle y_j, A_{ij}^* z_i \rangle_{\mathcal{Y}} = \sum_{i,j} \langle y_j, (A^*)_{ji} z_i \rangle_{\mathcal{Y}} = \langle \mathbf{y}, \mathbf{A}^* \mathbf{z} \rangle_{\mathcal{Y}^n}$, where $(A^*)_{ji} = (A_{ij})^*$.

To help the reader, notations frequently used in the paper are summarized in Table 1.

4. Reproducing Kernel Hilbert Spaces of Function-valued Functions

Hilbert spaces of scalar-valued functions with reproducing kernels were introduced and studied in Aronszajn (1950). Due to their crucial role in designing kernel-based learning methods, these spaces have received considerable attention over the last two decades (Shawe-Taylor and Cristanini, 2004; Schölkopf and Smola, 2002). More recently, interest has grown in exploring reproducing Hilbert spaces of vector functions for learning vector-valued functions (Micchelli and Pontil, 2005a; Carmeli et al., 2006; Caponnetto et al., 2008; Carmeli et al., 2010; Zhang et al., 2012), even though the idea of extending the theory of Reproducing Kernel Hilbert Spaces from the scalar-valued case to the vector-valued one is not new and dates back to at least Schwartz (1964). For more details, see the review paper by Álvarez et al. (2012).

In the field of machine learning, Evgeniou et al. (2005) have shown how Hilbert spaces of vector-valued functions and matrix-valued reproducing kernels can be used in the context of multi-task learning, with the goal of learning many related regression or classification tasks simultaneously. Since this seminal work, it has been demonstrated that these kernels and their associated spaces are capable of solving various other learning problems such as multiple output learning (Baldassarre et al., 2012), manifold regularization (Minh and Sindhwani, 2011), structured output prediction (Brouard et al., 2011; Kadri et al., 2013a), multi-view learning (Minh et al., 2013; Kadri et al., 2013b) and network inference (Lim et al., 2013, 2015).

4. We also use the standard notations such as $\mathbb{R}^n$ and L^2.
5. We denote by small Latin characters scalar-valued functions. Operator-valued functions (or kernels) are denoted by capital Latin characters $A(\cdot, \cdot)$, $B(\cdot, \cdot)$, $K(\cdot, \cdot)$, ...

In contrast to most of these previous works, here we are interested in the general case where the output space is a space of vectors with infinite dimension. This may be valuable from a variety of perspectives. Our main motivation is the supervised learning problem when output data are functions that could represent, for example, one-dimensional curves (this was mentioned as future work in Szedmak et al. 2006). One of the simplest ways to handle these data is to treat them as multivariate vectors. However this method does not consider any dependency of different values over subsequent time-points within the same functional datum and suffers when data dimension is very large. Therefore, we adopt a functional data analysis viewpoint (Zhao et al., 2004; Ramsay and Silverman, 2005; Ferraty and Vieu, 2006) in which multiple curves are viewed as functional realizations of a single function. It is important to note that matrix-valued kernels for infinite-dimensional output spaces, commonly known as operator-valued kernels, have been considered in previous studies (Micchelli and Pontil, 2005a; Caponnetto et al., 2008; Carmeli et al., 2006, 2010); however, they have been only studied in a theoretical perspective. Clearly, further investigations are needed to illustrate the practical benefits of the use of operator-valued kernels, which is the main focus of this work.

We now describe how RKHS theory can be extended from real or vector to functional response data. In particular, we focus on reproducing kernel Hilbert spaces whose elements are function-valued functions (or operators) and we demonstrate how basic properties of real-valued RKHS can be restated in the functional case, if appropriate conditions are satisfied. Extension to the functional case is not so obvious and requires tools from functional analysis (Rudin, 1991). Spaces of operators whose range is infinite-dimensional can exhibit unusual behavior, and standard topological properties may not be preserved in the infinite-dimensional case because of functional analysis subtleties. So, additional restrictions imposed on these spaces are needed for extending the theory of RKHS towards infinite-dimensional output spaces. Following Carmeli et al. (2010), we mainly focus on separable Hilbert spaces with reproducing operator-valued kernels whose elements are continuous functions. This is a sufficient condition to avoid topological and measurability problems encountered with this extension. For more details about vector or function-valued RKHS of measurable and continuous functions, see Carmeli et al. (2006, Sections 3 and 5). Note that the framework developed in this section should be valid for any type of input data (vectors, functions, or structures). In this paper, however, we consider the case where both input and output data are functions.

Definition 3 *(Operator-valued kernel)*
An $\mathcal{L}(\mathcal{Y})$-valued kernel K on $\mathcal{X}^2$ is a function $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \longrightarrow \mathcal{L}(\mathcal{Y})$;

(i) *K is Hermitian if $\forall x, z \in \mathcal{X}$, $K(w,z) = K(z,w)^*$, (where the superscript * denotes the adjoint operator),*

(ii) *K is nonnegative on $\mathcal{X}$ if it is Hermitian and for every natural number r and all $\{(w_i, u_i)_{i=1,\ldots,r}\} \in \mathcal{X} \times \mathcal{Y}$, the matrix with ij-th entry $\langle K(w_i, w_j)u_i, u_j\rangle_{\mathcal{Y}}$ is nonnegative (positive-definite).*

Definition 4 *(Block operator kernel matrix)*
Given a set $\{w_i\} \in \mathcal{X}$, $i = 1, \ldots, n$ with $n \in \mathbb{N}_+$, and an operator-valued kernel K, the

corresponding block operator kernel matrix is the matrix $\mathbf{K} \in \mathcal{L}(\mathcal{Y}^n)$ *with entries*

$$\mathbf{K}_{ij} = K(w_i, w_j).$$

The block operator kernel matrix is simply the kernel matrix associated to an operator-valued kernel. Since the kernel outputs an operator, the kernel matrix is in this case a block matrix where each block is an operator in $\mathcal{L}(\mathcal{Y})$. It is easy to see that an operator-valued kernel K is nonnegative if and only if the associated block operator kernel matrix $\mathbf{K}$ is positive.

Definition 5 *(Function-valued RKHS)*
A Hilbert space $\mathcal{F}$ *of functions from* $\mathcal{X}$ *to* $\mathcal{Y}$ *is called a reproducing kernel Hilbert space if there is a nonnegative* $\mathcal{L}(\mathcal{Y})$*-valued kernel* K *on* $\mathcal{X}^2$ *such that:*

(i) *the function* $z \longmapsto K(w, z)g$ *belongs to* $\mathcal{F}, \forall z, w \in \mathcal{X}$ *and* $g \in \mathcal{Y}$,

(ii) *for every* $F \in \mathcal{F}, w \in \mathcal{X}$ *and* $g \in \mathcal{Y}$, $\langle F, K(w, \cdot)g\rangle_{\mathcal{F}} = \langle F(w), g\rangle_{\mathcal{Y}}$.

On account of (ii), the kernel is called the reproducing kernel of $\mathcal{F}$. In Carmeli et al. (2006, Section 5), the authors provided a characterization of RKHS with operator-valued kernels whose functions are continuous and proved that $\mathcal{F}$ is a subspace of $\mathcal{C}(\mathcal{X}, \mathcal{Y})$, the vector space of continuous functions from $\mathcal{X}$ to $\mathcal{Y}$, if and only if the reproducing kernel K is locally bounded and separately continuous. Such a kernel is qualified as Mercer (Carmeli et al., 2010). In the following, we will only consider separable RKHS $\mathcal{F} \subset \mathcal{C}(\mathcal{X}, \mathcal{Y})$.

Theorem 1 *(Uniqueness of the reproducing operator-valued kernel)*
If a Hilbert space $\mathcal{F}$ *of functions from* $\mathcal{X}$ *to* $\mathcal{Y}$ *admits a reproducing kernel, then the reproducing kernel* K *is uniquely determined by the Hilbert space* $\mathcal{F}$.

Proof: Let K be a reproducing kernel of $\mathcal{F}$. Suppose that there exists another reproducing kernel K' of $\mathcal{F}$. Then, for all $\{w, w'\} \in \mathcal{X}$ and $\{h, g\} \in \mathcal{Y}$, applying the reproducing property for K and K' we get

$$\langle K'(w', \cdot)h, K(w, \cdot)g\rangle_{\mathcal{F}} = \langle K'(w', w)h, g\rangle_{\mathcal{Y}}, \tag{1}$$

we have also

$$\begin{array}{rcccl} \langle K'(w', \cdot)h, K(w, \cdot)g\rangle_{\mathcal{F}} & = & \langle K(w, \cdot)g, K'(w', \cdot)h\rangle_{\mathcal{F}} & = & \langle K(w, w')g, h\rangle_{\mathcal{Y}} \\ & = & \langle g, K(w, w')^* h\rangle_{\mathcal{Y}} & = & \langle g, K(w', w)h\rangle_{\mathcal{Y}}. \end{array} \tag{2}$$

(1) and (2) $\Rightarrow K(w, w') \equiv K'(w, w')$, $\forall w, w' \in \mathcal{X}$. ■

A key point for learning with kernels is the ability to express functions in terms of a kernel providing the way to evaluate a function at a given point. This is possible because there exists a bijection relationship between a large class of kernels and associated reproducing kernel spaces which satisfy a regularity property. Bijection between scalar-valued kernels and RKHS was first established by Aronszajn (1950, Part I, Sections 3 and 4). Then Schwartz (1964, Chapter 5) shows that this is a particular case of a more general situation. This bijection in the case where input and output data are continuous and belong to the infinite-dimensional functional spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, is still valid and is given by the following theorem (see also Theorem 4 of Senkene and Tempel'man, 1973).

Theorem 2 *(Bijection between function-valued RKHS and operator-valued kernel)* *A $\mathcal{L}(\mathcal{Y})$-valued Mercer kernel K on $\mathcal{X}^2$ is the reproducing kernel of some Hilbert space $\mathcal{F}$, if and only if it is nonnegative.*

We give a proof of this theorem by extending the scalar-valued case $\mathcal{Y} = \mathbb{R}$ in Aronszajn (1950) to the domain of functional data analysis domain where $\mathcal{Y}$ is $L^2(\Omega_y)$.[6] The proof is performed in two steps. The necessity is an immediate result from the reproducing property. For the sufficiency, the outline of the proof is as follows: we assume $\mathcal{F}_0$ to be the space of all $\mathcal{Y}$-valued functions F of the form $F(\cdot) = \sum_{i=1}^{n} K(w_i, \cdot)u_i$, where $w_i \in \mathcal{X}$ and $u_i \in \mathcal{Y}$, with the following inner product $\langle F(\cdot), G(\cdot)\rangle_{\mathcal{F}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \langle K(w_i, z_j)u_i, v_j\rangle_{\mathcal{Y}}$ defined for any $G(\cdot) = \sum_{j=1}^{m} K(z_j, \cdot)v_j$ with $z_j \in \mathcal{X}$ and $v_j \in \mathcal{Y}$. We show that $(\mathcal{F}_0, \langle\cdot,\cdot\rangle_{\mathcal{F}_0})$ is a pre-Hilbert space. Then we complete this pre-Hilbert space via Cauchy sequences $\{F_n(\cdot)\} \subset \mathcal{F}_0$ to construct the Hilbert space $\mathcal{F}$ of $\mathcal{Y}$-valued functions. Finally, we conclude that $\mathcal{F}$ is a reproducing kernel Hilbert space, since $\mathcal{F}$ is a real inner product space that is complete under the norm $\|\cdot\|_{\mathcal{F}}$ defined by $\|F(\cdot)\|_{\mathcal{F}} = \lim_{n\to\infty} \|F_n(\cdot)\|_{\mathcal{F}_0}$, and has $K(\cdot,\cdot)$ as reproducing kernel.

Proof: *Necessity.* Let K be the reproducing kernel of a Hilbert space $\mathcal{F}$. Using the reproducing property of the kernel K we obtain for any $\{w_i, w_j\} \in \mathcal{X}$ and $\{u_i, u_j\} \in \mathcal{Y}$

$$\sum_{i,j=1}^{n} \langle K(w_i, w_j)u_i, u_j\rangle_Y = \sum_{i,j=1}^{n} \langle K(w_i,\cdot)u_i, K(w_j,\cdot)u_j\rangle_{\mathcal{F}}$$
$$= \langle \sum_{i=1}^{n} K(w_i,\cdot)u_i, \sum_{i=1}^{n} K(w_i,\cdot)u_i\rangle_{\mathcal{F}} = \| \sum_{i=1}^{n} K(w_i,\cdot)u_i\|_{\mathcal{F}}^2 \ \geq \ 0.$$

Sufficiency. Let $\mathcal{F}_0 \subset \mathcal{Y}^{\mathcal{X}}$ be the space of all $\mathcal{Y}$-valued functions F of the form $F(\cdot) = \sum_{i=1}^{n} K(w_i, \cdot)u_i$, where $w_i \in \mathcal{X}$ and $u_i \in \mathcal{Y}$, $i = 1, \ldots, n$. We define the inner product of the functions $F(\cdot) = \sum_{i=1}^{n} K(w_i, \cdot)u_i$ and $G(\cdot) = \sum_{j=1}^{m} K(z_j, \cdot)v_j$ from $\mathcal{F}_0$ as follows

$$\langle F(\cdot), G(\cdot)\rangle_{\mathcal{F}_0} = \langle \sum_{i=1}^{n} K(w_i,\cdot)u_i, \sum_{j=1}^{m} K(z_j,\cdot)v_j\rangle_{\mathcal{F}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \langle K(w_i, z_j)u_i, v_j\rangle_{\mathcal{Y}}.$$

$\langle F(\cdot), G(\cdot)\rangle_{\mathcal{F}_0}$ is a symmetric bilinear form on $\mathcal{F}_0$ and due to the positivity of the kernel K, $\|F(\cdot)\|$ defined by

$$\|F(\cdot)\| = \sqrt{\langle F(\cdot), F(\cdot)\rangle_{\mathcal{F}_0}}$$

is a quasi-norm in $\mathcal{F}_0$. The reproducing property in $\mathcal{F}_0$ is verified with the kernel K. In fact, if $F \in \mathcal{F}_0$ then

$$F(\cdot) = \sum_{i=1}^{n} K(w_i, \cdot)u_i,$$

6. The proof should be applicable to arbitrarily separable output Hilbert spaces $\mathcal{Y}$.

and $\forall\ (w,u) \in \mathcal{X} \times \mathcal{Y}$,

$$\langle F, K(w,\cdot)u\rangle_{\mathcal{F}_0} = \langle \sum_{i=1}^{n} K(w_i,\cdot)u_i, K(w,\cdot)u\rangle_{\mathcal{F}_0} = \langle \sum_{i=1}^{n} K(w_i,w)u_i, u\rangle_{\mathcal{Y}} = \langle F(w), u\rangle_{\mathcal{Y}}.$$

Moreover using the Cauchy-Schwartz inequality, we have: $\forall\ (w,u) \in \mathcal{X} \times \mathcal{Y}$,

$$\langle F(w), u\rangle_{\mathcal{Y}} = \langle F(\cdot), K(w,\cdot)u\rangle_{\mathcal{F}_0} \leq \|F(\cdot)\|_{\mathcal{F}_0}\|K(w,\cdot)u\|_{\mathcal{F}_0}.$$

Thus, if $\|F\|_{\mathcal{F}_0} = 0$, then $\langle F(w), u\rangle_{\mathcal{Y}} = 0$ for any w and u, and hence $F \equiv 0$. Thus $(\mathcal{F}_0, \langle .,. \rangle_{\mathcal{F}_0})$ is a pre-Hilbert space. This pre-Hilbert space is in general not complete, but it can be completed via Cauchy sequences to build the $\mathcal{Y}$-valued Hilbert space $\mathcal{F}$ which has K as reproducing kernel, which concludes the proof. The completion of $\mathcal{F}_0$ is given in Appendix A (we refer the reader to the monograph by Rudin, 1991, for more details about completeness and the general theory of topological vector spaces). ■

We now give an example of a function-valued RKHS and its operator-valued kernel. This example serves to illustrate how these spaces and their associated kernels generalize the standard scalar-valued case or the vector-valued one to functional and infinite-dimensional output data. Thus, we first report an example of a scalar-valued RKHS and the corresponding scalar-valued kernel. We then extend this example to the case of vector-valued Hilbert spaces with matrix-valued kernels, and finally to function-valued RKHS where the output space is infinite dimensional. For the sake of simplicity, the input space $\mathcal{X}$ in these examples is assumed to be a subset of $\mathbb{R}$.

Example 1 *(Scalar-valued RKHS and its scalar-valued kernel; see Canu et al. (2003))*
Let $\mathcal{F}$ be the space defined as follows:

$$\begin{cases} \mathcal{F} = \big\{ f : [0,1] \longrightarrow \mathbb{R} \text{ absolutely continuous, } \exists f' \in L^2([0,1]), f(x) = \displaystyle\int_0^x f'(z)dz \big\}, \\ \langle f_1, f_2\rangle_{\mathcal{H}} = \langle f_1', f_2'\rangle_{L^2([0,1])}. \end{cases}$$

$\mathcal{F}$ is the Sobolev space of degree 1, also called the Cameron-Martin space, and is a scalar-valued RKHS of functions $f : [0,1] \longrightarrow \mathbb{R}$ with the scalar-valued reproducing kernel $k(x,z) = min(x,z)$, $\forall x, z \in \mathcal{X} = [0,1]$.

Example 2 *(Vector-valued RKHS and its matrix-valued kernel)*
Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \mathbb{R}^n$. Consider the matrix-valued kernel K defined by:

$$K(x,z) = \begin{cases} diag(x) & \text{if } \ x \leq z, \\ diag(z) & \text{otherwise}, \end{cases} \tag{3}$$

where, $\forall a \in \mathbb{R}$, $diag(a)$ is the $n \times n$ diagonal matrix with diagonal entries equal to a. Let $\mathcal{M}$ be the space of vector-valued functions from $\mathcal{X}$ onto $\mathbb{R}^n$ whose norm $\|g\|^2_{\mathcal{M}} = \displaystyle\sum_{i=1}^{n} \int_{\mathcal{X}} [g(x)]_i^2 dx$ is finite.

The matrix-valued mapping K is the reproducing kernel of the vector-valued RKHS $\mathcal{F}$ defined as follows:

$$\begin{cases} \mathcal{F} = \big\{ f : [0,1] \longrightarrow \mathbb{R}^n,\ \exists f' = \dfrac{df(x)}{dx} \in \mathcal{M}, [f(x)]_i = \displaystyle\int_0^x [f'(z)]_i dz, \forall i = 1, \ldots, n \big\}, \\ \langle f_1, f_2 \rangle_{\mathcal{F}} = \langle f_1', f_2' \rangle_{\mathcal{M}}. \end{cases}$$

Indeed, K is nonnegative and we have, $\forall x \in \mathcal{X}$, $y \in \mathbb{R}^n$ and $f \in \mathcal{F}$,

$$\begin{aligned} &\langle f, K(x,\cdot)y \rangle_{\mathcal{F}} = \langle f', [K(x,\cdot)y]' \rangle_{\mathcal{M}} \\ &= \sum_{i=1}^{n} \int_0^1 [f'(z)]_i [K(x,z)y]_i' dz \\ &= \sum_{i=1}^{n} \int_0^x [f'(z)]_i y_i dz \quad (dK(x,z)/dz = diag(1) \text{ if } z \leq x, \text{ and } = diag(0) \text{ otherwise}) \\ &= \sum_{i=1}^{n} [f(x)]_i y_i dz \ = \ \langle f(x), y \rangle_{\mathbb{R}^n}. \end{aligned}$$

■

Example 3 *(Function-valued RKHS and its operator-valued kernel)*

Here we extend Example 2 to the case where the output space is infinite dimensional. Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = L^2(\Omega)$ the space of square integrable functions on a compact set $\Omega \subset \mathbb{R}$. We denote by $\mathcal{M}$ the space of $L^2(\Omega)$-valued functions on $\mathcal{X}$ whose norm $\|g\|_{\mathcal{M}}^2 = \displaystyle\int_\Omega \int_{\mathcal{X}} [g(x)(t)]^2 dx dt$ is finite.

Let $(\mathcal{F}; \langle \cdot, \cdot \rangle_{\mathcal{F}})$ be the space of functions from $\mathcal{X}$ to $L^2(\Omega)$ such that:

$$\begin{cases} \mathcal{F} = \big\{ f,\ \exists f' = \dfrac{df(x)}{dx} \in \mathcal{M}, f(x) = \displaystyle\int_0^x f'(z) dz \big\}, \\ \langle f_1, f_2 \rangle_{\mathcal{F}} = \langle f_1', f_2' \rangle_{\mathcal{M}}. \end{cases}$$

$\mathcal{F}$ is a function-valued RKHS with the operator-valued kernel $K(x,z) = M_{\varphi(x,z)}$. M_φ is the multiplication operator associated with the function φ where $\varphi(x,z)$ is equal to x if $x \leq z$ and z otherwise. Since φ is a positive-definite function, K is Hermitian and nonnegative. Indeed,

$$\begin{aligned} \langle K(z,x)^* y, w \rangle_{\mathcal{Y}} = \langle y, K(z,x)w \rangle_{\mathcal{Y}} &= \int_0^1 \varphi(z,x) w(t) y(t) dt = \int_0^1 \varphi(x,z) y(t) z(t) dt \\ &= \langle K(x,z)y, w \rangle_{\mathcal{Y}}, \end{aligned}$$

and

$$\begin{aligned} \sum_{i,j} \langle K(x_i, x_j) y_i, y_j \rangle_{\mathcal{Y}} &= \sum_{i,j} \int_0^1 \varphi(x_i, x_j) y_i(t) y_j(t) dt \\ &= \int_0^1 \sum_{i,j} y_i(t) \varphi(x_i, x_j) y_j(t) dt \ \geq \ 0 \ \text{(since } \varphi \geq 0). \end{aligned}$$

Now we show that the reproducing property holds for any $f \in \mathcal{F}$, $y \in L^2(\Omega)$ and $x \in \mathcal{X}$:

$$\begin{aligned}
&\langle f, K(x,\cdot)y\rangle_{\mathcal{F}} = \langle f', [K(x,\cdot)y]'\rangle_{\mathcal{M}} \\
&= \int_\Omega \int_0^1 [f'(z)](t)[K(x,z)y]'(t)dzdt \\
&\overset{\text{kern. def}}{=} \int_\Omega \int_0^x [f'(z)](t)y(t)dzdt = \int_\Omega [f(x)](t)y(t)dt \\
&= \langle f(x), y\rangle_{L^2(\Omega)}. \qquad \blacksquare
\end{aligned}$$

Theorem 2 states that it is possible to construct a pre-Hilbert space of operators from a nonnegative operator-valued kernel and with some additional assumptions it can be completed to obtain a function-valued reproducing kernel Hilbert space. Therefore, it is important to consider the problem of constructing nonnegative operator-valued kernels. This is the focus of the next section.

5. Operator-valued Kernels for Functional Data

Reproducing kernels play an important role in statistical learning theory and functional estimation. Scalar-valued kernels are widely used to design nonlinear learning methods which have been successfully applied in several machine learning applications (Schölkopf and Smola, 2002; Shawe-Taylor and Cristanini, 2004). Moreover, their extension to matrix-valued kernels has helped to bring additional improvements in learning vector-valued functions (Micchelli and Pontil, 2005a; Reisert and Burkhardt, 2007; Caponnetto and De Vito, 2006). The most common and most successful applications of matrix-valued kernel methods are in multi-task learning (Evgeniou et al., 2005; Micchelli and Pontil, 2005b), even though some successful applications also exist in other areas, such as image colorization (Minh et al., 2010), link prediction (Brouard et al., 2011) and network inference (Lim et al., 2015). A basic, albeit not obvious, question which is always present with reproducing kernels concerns how to build these kernels and what is the optimal kernel choice. This question has been studied extensively for scalar-valued kernels, however it has not been investigated enough in the matrix-valued case. In the context of multi-task learning, matrix-valued kernels are constructed from scalar-valued kernels which are carried over to the vector-valued setting by a positive definite matrix (Micchelli and Pontil, 2005b; Caponnetto et al., 2008).

In this section we consider the problem from a more general point of view. We are interested in the construction of operator-valued kernels, generalization of matrix-valued kernels in infinite dimensional spaces, that perform the mapping between two spaces of functions and which are suitable for functional response data. Our motivation is to build operator-valued kernels that are capable of giving rise to nonlinear FDA methods. It is worth recalling that previous studies have provided examples of operator-valued kernels with infinite-dimensional output spaces (Micchelli and Pontil, 2005a; Caponnetto et al., 2008; Carmeli et al., 2010); however, they did not focus either on building methodological connections with the area of FDA, or on the practical impact of such kernels on real-world applications.

Motivated by building kernels that capture dependencies between samples of functional (infinite-dimensional) response variables, we adopt a FDA modeling formalism. The

design of such kernels will doubtless prove difficult, but it is necessary to develop reliable nonlinear FDA methods. Most FDA methods in the literature are based on linear parametric models. Extending these methods to nonlinear contexts should render them more powerful and efficient. Our line of attack is to construct operator-valued kernels from operators already used to build linear FDA models, particularly those involved in functional response models. Thus, it is important to begin by looking at these models.

5.1 Linear Functional Response Models

FDA is an extension of multivariate data analysis suitable when data are functions. In this framework, a data is a single function observation rather than a collection of observations. It is true that the data measurement process often provides a vector rather than a function, but the vector is a discretization of a real attribute which is a function. Hence, a functional datum i is acquired as a set of discrete measured values, $y_{i1}, \ldots, y_{ip}$; the first task in parametric (linear) FDA methods is to convert these values to a function y_i with values $y_i(t)$ computable for any desired argument value t. If the discrete values are assumed to be noiseless, then the process is interpolation; but if they have some observational error, then the conversion from discrete data to functions is a regression task (*e.g.*, smoothing) (Ramsay and Silverman, 2005).

A functional data model takes the form $y_i = f(x_i) + \epsilon_i$ where one or more of the components y_i, x_i and ϵ_i are functions. Three subcategories of such models can be distinguished: predictors x_i are functions and responses y_i are scalars; predictors are scalars and responses are functions; both predictors and responses are functions. In the latter case, which is the context we face, the function f is a compact operator between two infinite-dimensional Hilbert spaces. Most previous works on this model suppose that the relation between functional responses and predictors is linear; for more details, see Ramsay and Silverman (2005) and references therein.

For functional input and output data, the functional linear model commonly found in the literature is an extension of the multivariate linear one and has the following form:

$$y(t) = \alpha(t) + \beta(t)x(t) + \epsilon(t), \tag{4}$$

where α and β are the functional parameters of the model (Ramsay and Silverman, 2005, Chapter 14). This model is known as the "concurrent model" where "concurrent" means that $y(t)$ only depends on x at t. The concurrent model is similar to the varying coefficient model proposed by Hastie and Tibshirani (1993) to deal with the case where the parameter β of a multivariate regression model can vary over time. A main limitation of this model is that the response y and the covariate x are both functions of the same argument t, and the influence of a covariate on the response is concurrent or point-wise in the sense that x only influences $y(t)$ through its value $x(t)$ at time t. To overcome this restriction, an extended linear model in which the influence of a covariate x can involve a range of argument values $x(s)$ was proposed; it takes the following form:

$$y(t) = \alpha(t) + \int x(s)\beta(s,t)ds + \epsilon(t), \tag{5}$$

where, in contrast to the concurrent model, the functional parameter β is now a function of both s and t, and $y(t)$ depends on $x(s)$ for an interval of values of s (Ramsay and Silverman,

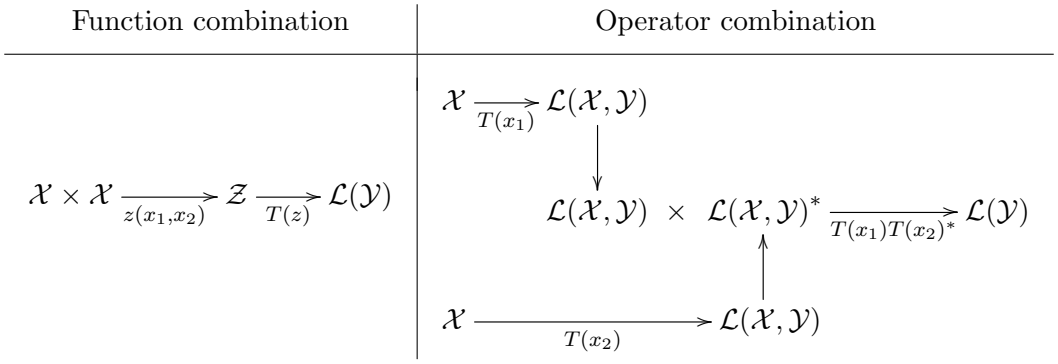

Figure 3: Illustration of building an operator-valued kernel from $\mathcal{X}\times\mathcal{X}$ to $\mathcal{L}(\mathcal{Y})$ using a combination of functions or a combination of operators. (left) The operator-valued kernel is constructed by combining two functions (x_1 and x_2) and by applying a positive $\mathcal{L}(\mathcal{Y})$-valued mapping T to the combination. (right) the operator-valued kernel is generated by combining two operators ($T(x_1)$ and $T(x_2)^*$) built from an $\mathcal{L}(\mathcal{X},\mathcal{Y})$-valued mapping T.

2005, Chapter 16). Estimation of the parameter function $\beta(\cdot,\cdot)$ is an inverse problem and requires regularization. Regularization can be implemented in a variety of ways, for example by penalized splines (James, 2002) or by truncation of series expansions (Müller, 2005). A review of functional response models can be found in Chiou et al. (2004).

The operators involved in the functional data models described above are the multiplication operator (Equation 4) and the integral operator (Equation 5). We think that operator-valued kernels constructed using these operators could be a valid alternative to extend linear FDA methods to nonlinear settings. In Subsection 5.4 we provide examples of multiplication and integral operator-valued kernels. Before that, we identify building schemes that can be common to many operator-valued kernels and applied to functional data.

5.2 Operator-valued Kernel Building Schemes

In our context, constructing an operator-valued kernel turns out to build an operator that maps a couple of functions to a function: in $\mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ from two functions x_1 and x_2 in $\mathcal{X}$. This can be performed in one of two ways: either combining the two functions x_1 and x_2 into a variable $z \in \mathcal{Z}$ and then adding an operator function $T : \mathcal{Z} \longrightarrow \mathcal{L}(\mathcal{Y})$ that performs the mapping from space $\mathcal{Z}$ to $\mathcal{L}(\mathcal{Y})$, or building an $\mathcal{L}(\mathcal{X},\mathcal{Y})$-valued function T, where $\mathcal{L}(\mathcal{X},\mathcal{Y})$ is the set of bounded operators from $\mathcal{X}$ to $\mathcal{Y}$, and then combining the resulting operators $T(x_1)$ and $T(x_2)$ to obtain the operator in $\mathcal{L}(\mathcal{Y})$. In the latter case, a natural way to combine $T(x_1)$ and $T(x_2)$ is to use the composition operation and the kernel $K(x_1,x_2)$ will be equal to $T(x_1)T(x_2)^*$. Figure 3 describes the construction of an operator-valued kernel function using the two schemes which are based on combining functions (x_1 and x_2) or operators ($T(x_1)$ and $T(x_2)$), respectively. Note that separable operator-valued kernels (Álvarez et al., 2012), which are kernels that can be formulated as a product of

a scalar-valued kernel function for the input space alone and an operator that encodes the interactions between the outputs, are a particular case of the function combination building scheme, when we take $\mathcal{Z}$ as the set of real numbers $\mathbb{R}$ and the scalar-valued kernel as combination function. In contrast, the operator combination scheme is particularly amenable to the design of nonseparable operator-valued kernels. This scheme was already used in various problems of operator theory, system theory and interpolation (Alpay et al., 1997; Dym, 1989).

To build an operator-valued kernel and then construct a function-valued reproducing kernel Hilbert space, the operator T is of crucial importance. Choosing T presents two major difficulties. Computing the adjoint operator is not always easy to do, and then, not all operators verify the Hermitian condition of the kernel. On the other hand, since the kernel must be nonnegative, we suggest to construct operator-valued kernels from positive definite scalar-valued kernels which can be the reproducing kernels of real-valued Hilbert spaces. In this case, the reproducing property of the operator-valued kernel allows us to compute an inner product in a space of operators by an inner product in a space of functions which can be, in turn, computed using the scalar-valued kernel. The operator-valued kernel allows the mapping between a space of functions and a space of operators, while the scalar one establishes the link between the space of functions and the space of measured values. It is also useful to define combinations of nonnegative operator-valued kernels that allow to build a new nonnegative one.

5.3 Combinations of Operator-valued Kernels

We have shown in Section 4 that there is a bijection between nonnegative operator-valued kernels and function-valued reproducing kernel Hilbert spaces. So, as in the scalar case, it will be helpful to characterize algebraic transformations, like sum and product, that preserve the nonnegativity of operator-valued kernels. Theorem 3 stated below gives some building rules to obtain a positive operator-valued kernel from combinations of positive existing ones. Similar results for the case of matrix-valued kernels can be found in Reisert and Burkhardt (2007), and for a more general context we refer the reader to Caponnetto et al. (2008) and Carmeli et al. (2010). In our setting, assuming H and G be two nonnegative kernels constructed as described in the previous subsection, we are interested in constructing a nonnegative kernel K from H and G.

Theorem 3 *Let $H : \mathcal{X} \times \mathcal{X} \longrightarrow \mathcal{L}(\mathcal{Y})$ and $G : \mathcal{X} \times \mathcal{X} \longrightarrow \mathcal{L}(\mathcal{Y})$ two nonnegative operator-valued kernels*

(i) *$K \equiv H + G$ is a nonnegative kernel,*

(ii) *if $H(w, z)G(w, z) = G(w, z)H(w, z)$, $\forall w, z \in \mathcal{X}$, then $K \equiv HG$ is a nonnegative kernel,*

(iii) *$K \equiv THT^*$ is a nonnegative kernel for any $\mathcal{L}(\mathcal{Y})$-valued function $T(\cdot)$.*

Proof: Obviously (i) follows from the linearity of the inner product. (ii) can be proved by showing that the "element-wise" multiplication of two positive block operator matrices can be positive (see below). For the proof of (iii), we observe that

$$K(w, z)^* = [T(z)H(w, z)T(w)^*]^* = T(w)H(z, w)T(z)^* = K(z, w),$$

and

$$\sum_{i,j}\langle K(w_i,w_j)u_i,u_j\rangle = \sum_{i,j}\langle T(w_j)H(w_i,w_j)T(w_i)^*u_i,u_j\rangle$$
$$= \sum_{i,j}\langle H(w_i,w_j)T(w_i)^*u_i,T(w_j)^*u_j\rangle,$$

which implies the nonnegativity of the kernel K since H is nonnegative.

To prove (ii), i.e., the kernel $K \equiv HG$ is nonnegative in the case where H and G are nonnegative kernels such that $H(w,z)G(w,z) = G(w,z)H(w,z)$, $\forall w,z \in \mathcal{X}$, we show below that the block operator matrix $\mathbf{K}$ associated to the operator-valued kernel K for a given set $\{w_i\}, i = 1,\ldots,n$ with $n \in \mathbb{N}$, is positive. By construction, we have $\mathbf{K} = \mathbf{H} \circ \mathbf{G}$ where $\mathbf{H}$ and $\mathbf{G}$ are the block operator kernel matrices corresponding to the kernels H and G, and '$\circ$' denotes the "element-wise" multiplication defined by $(\mathbf{H} \circ \mathbf{G})_{ij} = H(w_i,w_j)G(w_i,w_j)$. $\mathbf{K}$, $\mathbf{H}$ and $\mathbf{G}$ are all in $\in \mathcal{L}(\mathcal{Y}^n)$.

Since the kernels H and G are Hermitian and $HG = GH$, it is easy to see that

$$(\mathbf{K}^*)_{ij} = (\mathbf{K}_{ji})^* = K(w_j,w_i)^* = \big(H(w_j,w_i)G(w_j,w_i)\big)^* = G(w_j,w_i)^*H(w_j,w_i)^*$$
$$= G(w_i,w_j)H(w_i,w_j) = H(w_i,w_j)G(w_i,w_j)$$
$$= \mathbf{K}_{ij}.$$

Thus, $\mathbf{K}$ is self-adjoint. It remains, then, to prove that $\langle \mathbf{Ku},\mathbf{u}\rangle \geq 0$, $\forall \mathbf{u} \in \mathcal{Y}^n$, in order to show the positivity of $\mathbf{K}$.

The "element-wise" multiplication can be rewritten as a tensor product. Indeed, we have

$$\mathbf{K} = \mathbf{H} \circ \mathbf{G} = \mathbf{L}^*(\mathbf{H} \otimes \mathbf{G})\mathbf{L},$$

where $\mathbf{L} : \mathcal{Y}^n \longrightarrow \mathcal{Y}^n \otimes \mathcal{Y}^n$ is the mapping defined by $\mathbf{Le_i} = \mathbf{e_i} \otimes \mathbf{e_i}$ for an orthonormal basis $\{\mathbf{e_i}\}$ of the separable Hilbert space $\mathcal{Y}^n$, and $\mathbf{H} \otimes \mathbf{G}$ is the tensor product defined by $(\mathbf{H} \otimes \mathbf{G})(\mathbf{u} \otimes \mathbf{v}) = \mathbf{Hu} \otimes \mathbf{Gv}$, $\forall \mathbf{u},\mathbf{v} \in \mathcal{Y}^n$. To see this, note that

$$\langle \mathbf{L}^*(\mathbf{H} \otimes \mathbf{G})\mathbf{Le_i},\mathbf{e_j}\rangle = \langle (\mathbf{H} \otimes \mathbf{G})\mathbf{Le_i},\mathbf{Le_j}\rangle = \langle (\mathbf{H} \otimes \mathbf{G})(\mathbf{e_i} \otimes \mathbf{e_i}),\mathbf{e_j} \otimes \mathbf{e_j}\rangle$$
$$= \langle \mathbf{He_i} \otimes \mathbf{Ge_i},\mathbf{e_j} \otimes \mathbf{e_j}\rangle = \langle \mathbf{He_i},\mathbf{e_j}\rangle\langle \mathbf{Ge_i},\mathbf{e_j}\rangle$$
$$= \mathbf{H}_{ij}\mathbf{G}_{ij} = \langle (\mathbf{H} \circ \mathbf{G})\mathbf{e_i},\mathbf{e_j}\rangle.$$

Now since H and G are positive, we have

$$\langle \mathbf{Ku},\mathbf{u}\rangle = \langle \mathbf{L}^*(\mathbf{H} \otimes \mathbf{G})\mathbf{Lu},\mathbf{u}\rangle = \langle \mathbf{L}^*(\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}}\mathbf{G}^{\frac{1}{2}})\mathbf{Lu},\mathbf{u}\rangle$$
$$= \langle \mathbf{L}^*(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})\mathbf{Lu},\mathbf{u}\rangle = \langle (\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})\mathbf{Lu},(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})^*\mathbf{Lu}\rangle$$
$$= \langle (\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})\mathbf{Lu},(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})\mathbf{Lu}\rangle = \|(\mathbf{H}^{\frac{1}{2}} \otimes \mathbf{G}^{\frac{1}{2}})\mathbf{Lu}\|^2 \geq 0.$$

This concludes the proof.

■

5.4 Examples of Nonnegative Operator-valued Kernels

We provide here examples of operator-valued kernels for functional response data. All these examples deal with operator-valued kernels constructed following the schemes described above and assuming that $\mathcal{Y}$ is an infinite-dimensional function space. Motivated by building kernels suitable for functional data, the first two examples deal with operator-valued kernels constructed from the multiplication and the integral self-adjoint operators in the case where $\mathcal{Y}$ is the Hilbert space $L^2(\Omega_y)$ of square integrable functions on Ω_y endowed with the inner product $\langle \phi, \psi \rangle = \int_{\Omega_y} \phi(t)\psi(t)dt$. We think that these kernels represent an interesting alternative to extend linear functional models to nonlinear settings. The third example based on the composition operator shows how to build such kernels from non self-adjoint operators (this may be relevant when the functional linear model is based on a non self-adjoint operator). It also illustrates the kernel combination defined in Theorem 3(iii).

1. Multiplication operator:

   In Kadri et al. (2010), the authors attempted to extend the widely used Gaussian kernel to functional data domain using a multiplication operator and assuming that input and output data belong to the same space of functions. Here we consider a slightly different setting, where the input space $\mathcal{X}$ can be different from the output space $\mathcal{Y}$.

   A multiplication operator on $\mathcal{Y}$ is defined as follows:

   $$\begin{array}{rccl} T^h: & \mathcal{Y} & \longrightarrow & \mathcal{Y} \\ & y & \longmapsto & T^h_y \;\; ; \;\; T^h_y(t) \triangleq h(t)y(t). \end{array}$$

   The operator-valued kernel $K(\cdot,\cdot)$ is the following:

   $$\begin{array}{rccl} K: & \mathcal{X} \times \mathcal{X} & \longrightarrow & \mathcal{L}(\mathcal{Y}) \\ & x_1, x_2 & \longmapsto & k_x(x_1,x_2)T^{k_y}, \end{array}$$

   where $k_x(\cdot,\cdot)$ is a positive definite scalar-valued kernel and k_y a positive real function. It is easy to see that $\langle T^h x, y\rangle = \langle x, T^h y\rangle$, then T^h is a self-adjoint operator. Thus $K(x_2,x_1)^* = K(x_2,x_1)$ and K is Hermitian since $K(x_1,x_2) = K(x_2,x_1)$.

   Moreover, we have

   $$\begin{aligned} \sum_{i,j} \langle K(x_i,x_j)y_i, y_j\rangle_{\mathcal{Y}} &= \sum_{i,j} k_x(x_i,x_j)\langle k_y(\cdot)y_i(\cdot), y_j(\cdot)\rangle_{\mathcal{Y}} \\ &= \sum_{i,j} k_x(x_i,x_j) \int k_y(t)y_i(t)y_j(t)dt = \int \sum_{i,j} y_i(t)[k_x(x_i,x_j)k_y(t)]y_j(t)dt \geq 0, \end{aligned}$$

   since the product of two positive-definite scalar-valued kernels is also positive-definite. Therefore K is a nonnegative operator-valued kernel.

2. Hilbert-Schmidt integral operator:

   A Hilbert-Schmidt integral operator on $\mathcal{Y}$ associated with a kernel $h(\cdot,\cdot)$ is defined as follows:

   $$\begin{array}{rccl} T^h: & \mathcal{Y} & \longrightarrow & \mathcal{Y} \\ & y & \longmapsto & T^h_y \;\; ; \;\; T^h_y(t) \triangleq \int h(s,t)y(s)ds. \end{array}$$

In this case, an operator-valued kernel K is a Hilbert-Schmidt integral operator associated with positive definite scalar-valued kernels k_x and k_y, and it takes the following form:

$$\begin{array}{rccc} K(x_1,x_2)[\cdot]: & \mathcal{Y} & \longrightarrow & \mathcal{Y} \\ & f & \longmapsto & g \end{array}$$

where $g(t) = k_x(x_1,x_2)\int k_y(s,t)f(s)ds$.

The Hilbert-Schmidt integral operator is self-adjoint if k_y is Hermitian. This condition is verified and then it is easy to check that K is also Hermitian. K is nonnegative since

$$\sum_{i,j}\langle K(x_i,x_j)y_i,y_j\rangle_{\mathcal{Y}} = \iint \sum_{i,j} y_i(s)[k_x(x_i,x_j)k_y(s,t)]y_j(t)dsdt,$$

which is positive because of the positive-definiteness of the scalar-valued kernels k_x and k_y.

3. Composition operator:

   Let φ be an analytic map. The composition operator associated with φ is the linear map:

   $$C_\varphi : f \longmapsto f\circ\varphi$$

   First, we look for an expression of the adjoint of the composition operator C_φ acting on $\mathcal{Y}$ in the case where $\mathcal{Y}$ is a scalar-valued RKHS of functions on Ω_y and φ an analytic map of Ω_y into itself. For any f in the space $\mathcal{Y}$ associated with the real kernel k,

   $$\begin{array}{rcccl} \langle f, C_\varphi^* k_t(\cdot)\rangle & = & \langle C_\varphi f, k_t\rangle & = & \langle f\circ\varphi, k_t\rangle \\ & = & f(\varphi(t)) & = & \langle f, k_{\varphi(t)}\rangle. \end{array}$$

   This is true for any $f\in\mathcal{Y}$ and then $C_\varphi^* k_t = k_{\varphi(t)}$. In a similar way, $C_\varphi^* f$ can be computed at each point of the function f:

   $$(C_\varphi^* f)(t) = \langle C_\varphi^* f, k_t\rangle = \langle f, C_\varphi k_t\rangle = \langle f, k_t\circ\varphi\rangle$$

   Once we have expressed the adjoint of a composition operator in a reproducing kernel Hilbert space, we consider the following operator-valued kernel:

   $$\begin{array}{rccc} K: & \mathcal{X}\times\mathcal{X} & \longrightarrow & \mathcal{L}(\mathcal{Y}) \\ & x_1,x_2 & \longmapsto & C_{\psi(x_1)}C_{\psi(x_2)}^* \end{array}$$

   where $\psi(x_1)$ and $\psi(x_2)$ are maps of Ω_y into itself. It is easy to see that the kernel K is Hermitian. Using Theorem 3(iii) we obtain the nonnegativity property of the kernel.

5.5 Multiple Functional Data and Kernel Feature Map

Until now, we discussed operator-valued kernels and their corresponding RKHS from the perspective of extending Aronszajn (1950) pioneering work from scalar-valued or vector-valued cases to the function-valued case. However, it is also interesting to explore these kernels from a feature space point of view (Schölkopf et al., 1999; Caponnetto et al., 2008). In this subsection, we provide some ideas targeted at advancing the understanding of feature spaces associated with operator-valued kernels and we show how these kernels can design more suitable feature maps than those associated with scalar-valued kernels, especially when input data are infinite dimensional objects like curves. To explore the potential of adopting an operator-valued kernel feature space approach, we consider a supervised learning problem with multiple functional data where each observation is composed of more than one functional variable (Kadri et al., 2011b,c). Working with multiple functions allows to deal in a natural way with a lot of applications. There are many practical situations where a number of potential functional covariates are available to explain a response variable. For example, in audio and speech processing where signals are converted into different functional features providing information about their temporal, spectral and cepstral characteristics, or in meteorology where the interaction effects between various continuous variables (such as temperature, precipitation, and winds) is of particular interest.

Similar to the scalar case, operator-valued kernels provide an elegant way of dealing with nonlinear algorithms by reducing them to linear ones in some feature space F nonlinearly related to input space. A feature map associated with an operator-valued kernel K is a continuous function

$$\Phi : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathcal{L}(\mathcal{X}, \mathcal{Y}),$$

such that for every $x_1, x_2 \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$

$$\langle K(x_1, x_2)y_1, y_2\rangle_{\mathcal{Y}} = \langle \Phi(x_1, y_1), \Phi(x_2, y_2)\rangle_{\mathcal{L}(\mathcal{X},\mathcal{Y})},$$

where $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ is the set of linear mappings from $\mathcal{X}$ into $\mathcal{Y}$. By virtue of this property, Φ is called a *feature map associated with* K. Furthermore, from the reproducing property, it follows that in particular

$$\langle K(x_1, \cdot)y_1, K(x_2, \cdot)y_2\rangle_{\mathcal{F}} = \langle K(x_1, x_2)y_1, y_2\rangle_{\mathcal{Y}},$$

which means that any operator-valued kernel admits a feature map representation Φ with a feature space $\mathcal{F} \subset \mathcal{L}(\mathcal{X}, \mathcal{Y})$ defined by $\Phi(x_1, y_1) = K(x_1, \cdot)y_1$, and corresponds to an inner product in another space.

From this feature map perspective, we study the geometry of a feature space associated with an operator-valued kernel and we compare it with the geometry obtained by a scalar-valued kernel. More precisely, we consider two reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{H}$. $\mathcal{F}$ is a RKHS of function-valued functions on $\mathcal{X}$ with values in $\mathcal{Y}$. $\mathcal{X} \subset (L^2(\Omega_x))^p$ [7], $\mathcal{Y} \subset L^2(\Omega_y)$ and let K be the reproducing operator-valued kernel of $\mathcal{F}$. $\mathcal{H}$ is also a RKHS, but of scalar-valued functions on $\mathcal{X}$ with values in $\mathbb{R}$, and k its reproducing scalar-valued

7. p is the number of functions that represent input data. In the field of FDA, such data are called multivariate functional data.

kernel. The mappings Φ_K and Φ_k associated, respectively, with the kernels K and k are defined as follows

$$\Phi_K : (L^2)^p \to \mathcal{L}((L^2)^p, L^2), \quad x \mapsto K(x, \cdot)y,$$

and

$$\Phi_k : (L^2)^p \to \mathcal{L}((L^2)^p, \mathbb{R}), \quad x \mapsto k(x, \cdot).$$

These feature maps can be seen as a mapping of the input data x_i, which are vectors of functions in $(L^2)^p$, into a feature space in which the inner product can be computed using the kernel functions. This idea leads to design nonlinear methods based on linear ones in the feature space. In a supervised classification problem for example, since kernels map input data into a higher dimensional space, kernel methods deal with this problem by finding a linear separation in the feature space. We now compare the dimension of feature spaces obtained by the maps Φ_K and Φ_k. To do this, we adopt a functional data analysis point of view where observations are composed of sets of functions. Direct understanding of this FDA viewpoint comes from the consideration of the "atom" of a statistical analysis. In a basic course in statistics, atoms are "numbers", while in multivariate data analysis the atoms are vectors and methods for understanding populations of vectors are the focus. FDA can be viewed as the generalization of this, where the atoms are more complicated objects, such as curves, images or shapes represented by functions (Zhao et al., 2004). Based on this, the dimension of the input space is p since $x_i \in (L^2)^p$ is a vector of p functions. The feature space obtained by the map Φ_k is a space of functions, so its dimension from a FDA viewpoint is equal to one. The map Φ_K projects the input data into a space of operators $\mathcal{L}(\mathcal{X}, \mathcal{Y})$. This means that using the operator-valued kernel K corresponds to mapping the functional data x_i into a higher, possibly infinite, dimensional space $(L^2)^d$ with $d \to \infty$. In a binary functional classification problem, we have higher probability to achieve linear separation between the classes by projecting the functional data into a higher dimensional feature space rather than into a lower one (Cover's theorem), that is why we think that it is more suitable to use operator-valued than scalar-valued kernels in this context.

6. Function-valued Function Learning

In this section, we consider the problem of estimating an unknown function F such that $F(x_i) = y_i$ when observed data $(x_i(s), y_i(t))_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ are assumed to be elements of the space of square integrable functions L^2. $X = \{x_1, \ldots, x_n\}$ denotes the training set with corresponding targets $Y = \{y_1, \ldots, y_n\}$. Since $\mathcal{X}$ and $\mathcal{Y}$ are spaces of functions, the problem can be thought of as an operator estimation problem, where the desired operator maps a Hilbert space of factors to a Hilbert space of targets. Among all functions in a linear space of operators $\mathcal{F}$, an estimate $\widetilde{F} \in \mathcal{F}$ of F may be obtained by minimizing:

$$\widetilde{F} = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{n} \|y_i - F(x_i)\|_{\mathcal{Y}}^2.$$

Depending on $\mathcal{F}$, this problem can be ill-posed and a classical way to turn it into a well-posed problem is to use a regularization term. Therefore, we may consider the solution of

the problem as the function $\widetilde{F} \in \mathcal{F}$ that minimizes:

$$\widetilde{F}_\lambda = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{n} \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda \|F\|_{\mathcal{F}}^2, \tag{6}$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter. Existence of $\widetilde{F}_\lambda$ in the optimization problem (6) is guaranteed for $\lambda > 0$ by the generalized Weierstrass theorem and one of its corollary that we recall from Kurdila and Zabarankin (2005).

Theorem 4 *Let $\mathcal{Z}$ be a reflexive Banach space and $\mathcal{C} \subseteq \mathcal{Z}$ a weakly closed and bounded set. Suppose $J : \mathcal{C} \to \mathbb{R}$ is a proper lower semi-continuous function. Then J is bounded from below and has a minimizer on $\mathcal{C}$.*

Corollary 5 *Let $\mathcal{H}$ be a Hilbert space and $J : \mathcal{H} \to \mathbb{R}$ is a strongly lower semi-continuous, convex and coercive function. Then J is bounded from below and attains a minimizer.*

This corollary can be straightforwardly applied to problem (6) by defining:

$$J_\lambda(F) = \sum_{i=1}^{n} \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda \|F\|_{\mathcal{F}}^2,$$

where F belongs to the Hilbert space $\mathcal{F}$. It is easy to note that J_λ is continuous and convex. Besides, J_λ is coercive for $\lambda > 0$ since $\|F\|_{\mathcal{F}}^2$ is coercive and the sum involves only positive terms. Hence $\widetilde{F}_\lambda = \arg\min_{F \in \mathcal{F}} J_\lambda(F)$ exists.

6.1 Learning Algorithm

We are now interested in solving the minimization problem (6) in a reproducing kernel Hilbert space $\mathcal{F}$ of function-valued functions. In the scalar case, it is well-known that under general conditions on real-valued RKHS, the solution of this minimization problem can be written as:

$$\widetilde{F}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x),$$

where $\alpha_i \in \mathbb{R}$ and k is the reproducing kernel of a real-valued Hilbert space (Wahba, 1990). An extension of this solution to the domain of functional data analysis takes the following form:

$$\widetilde{F}(\cdot) = \sum_{i=1}^{n} K(x_i, \cdot) u_i, \tag{7}$$

where $u_i(\cdot)$ are in $\mathcal{Y}$ and the reproducing kernel K is a nonnegative operator-valued function. With regards to the classical representer theorem, here the kernel K outputs an operator and the "weights" u_i are functions. A proof of the representer theorem in the case of function-valued reproducing kernel Hilbert spaces is given in Appendix B (see also Micchelli and Pontil, 2005a).

Substituting (7) in (6) and using the reproducing property of $\mathcal{F}$, we come up with the following minimization problem over the scalar-valued functions $u_i \in \mathcal{Y}$ ($\mathbf{u}$ is the vector of functions $(u_i)_{i=1,\ldots,n} \in (\mathcal{Y})^n$) rather than the function-valued function (or operator) F:

$$\widetilde{\mathbf{u}}_\lambda = \arg\min_{\mathbf{u}\in(\mathcal{Y})^n} \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{n} K(x_i, x_j)u_j\|_{\mathcal{Y}}^2 + \lambda \sum_{i,j}^{n} \langle K(x_i, x_j)u_i, u_j\rangle_{\mathcal{Y}}. \tag{8}$$

Problem (8) can be solved in three ways:

1. Assuming that the observations are made on a regular grid $\{t_1, \ldots, t_m\}$, one can first discretize the functions x_i and y_i and then solve the problem using multivariate data analysis techniques (Kadri et al., 2010). However, as this is well-known in the FDA domain, this has the drawback of not taking into consideration the relationships that exist between samples.

2. The second way consists in considering the output space $\mathcal{Y}$ to be a scalar-valued reproducing Hilbert space. In this case, the functions u_i can be approximated by a linear combination of a scalar-valued kernel $\hat{u}_i = \sum_{l=1}^{m} \alpha_{il} k(s_l, \cdot)$ and then the problem (8) becomes a minimization problem over the real values α_{il} rather than the discrete values $u_i(t_1), \ldots, u_i(t_m)$. In the FDA literature, a similar idea has been adopted by Ramsay and Silverman (2005) and by Prchal and Sarda (2007) who expressed not only the functional parameters u_i but also the observed input and output data in a basis functions specified a priori (*e.g.*, Fourier basis or B-spline basis).

3. Another possible way to solve the minimization problem (8) is to compute its derivative using the directional derivative and setting the result to zero to find an analytic solution of the problem. It follows that the vector of functions $\mathbf{u} \in \mathcal{Y}^n$ satisfies the system of linear operator equations:

$$(\mathbf{K} + \lambda I)\mathbf{u} = \mathbf{y}, \tag{9}$$

where $\mathbf{K} = [K(x_i, x_j)]_{i,j=1}^{n}$ is a $n \times n$ block operator kernel matrix ($\mathbf{K}_{ij} \in \mathcal{L}(\mathcal{Y})$) and $\mathbf{y} \in \mathcal{Y}^n$ the vector of functions $(y_i)_{i=1}^{n}$. In this work, we are interested in this third approach which extends to functional data analysis domain results and properties known from multivariate statistical analysis. One main obstacle for this extension is the inversion of the block operator kernel matrix $\mathbf{K}$. Block operator matrices generalize block matrices to the case where the block entries are linear operators between infinite dimensional Hilbert spaces. These matrices and their inverses arise in some areas of mathematics (Tretter, 2008) and signal processing (Asif and Moura, 2005). In contrast to the multivariate case, inverting such matrices is not always feasible in infinite dimensional spaces. To overcome this problem, we study the eigenvalue decomposition of a class of block operator kernel matrices obtained from operator-valued kernels having the following form:

$$K(x_i, x_j) = g(x_i, x_j)T, \quad \forall x_i, x_j \in \mathcal{X}, \tag{10}$$

where g is a scalar-valued kernel and T is an operator in $\mathcal{L}(\mathcal{Y})$. This separable kernel construction is adapted from Micchelli and Pontil (2005a,b). The choice of T depends

on the context. For multi-task kernels, T is a finite dimensional matrix which models relations between tasks. In FDA, Lian (2007) suggested the use of the identity operator, while in Kadri et al. (2010) the authors showed that it is better to choose other operators than identity to take into account functional properties of the input and output spaces. They introduced a functional kernel based on the multiplication operator. In this work, we are more interested in kernels constructed from the integral operator. This seems to be a reasonable choice since functional linear model (see Equation 5) are based on this operator (Ramsay and Silverman, 2005, Chapter 16). So we can consider for example the following positive definite operator-valued kernel:

$$(K(x_i, x_j)y)(t) = g(x_i, x_j) \int_{\Omega_y} e^{-|t-s|} y(s) ds, \tag{11}$$

where $y \in \mathcal{Y} = L^2(\Omega_y)$ and $\{s, t\} \in \Omega_y = [0, 1]$. Note that a similar kernel was proposed as an example in Caponnetto et al. (2008) for linear spaces of functions from $\mathbb{R}$ to $\mathcal{G}_y$.

The $n \times n$ block operator kernel matrix $\mathbf{K}$ of operator-valued kernels having the form (10) can be expressed as a Kronecker product between the Gram matrix $G = \big(g(x_i, x_j)\big)_{i,j=1}^{n}$ in $\mathbb{R}^{n \times n}$ and the operator $T \in \mathcal{L}(\mathcal{Y})$, and is defined as follows:

$$\mathbf{K} = \begin{pmatrix} g(x_1, x_1)T & \dots & g(x_1, x_n)T \\ \vdots & \ddots & \vdots \\ g(x_n, x_1)T & \dots & g(x_n, x_n)T \end{pmatrix} = G \otimes T.$$

It is easy to show that basic properties of the Kronecker product between two finite matrices can be restated for this case. So, $\mathbf{K}^{-1} = G^{-1} \otimes T^{-1}$ and the eigendecomposition of the matrix $\mathbf{K}$ can be obtained from the eigendecompositions of G and T (see Algorithm 1).

Theorem 6 *If $T \in \mathcal{L}(\mathcal{Y})$ is a compact, normal operator ($TT^* = T^*T$) on the Hilbert space $\mathcal{Y}$, then there exists an orthonormal basis of eigenfunctions $\{\phi_i, i \geq 1\}$ corresponding to eigenvalues $\{\lambda_i, i \geq 1\}$ such that*

$$Ty = \sum_{i=1} \lambda_i \langle y, \phi_i \rangle \phi_i, \quad \forall y \in \mathcal{Y}.$$

Proof: See Naylor and Sell (1971)[theorem 6.11.2] ■

Let θ_i and $\mathbf{z_i}$ be, respectively, the eigenvalues and the eigenfunctions of $\mathbf{K}$. From Theorem 6 it follows that the inverse operator $\mathbf{K}^{-1}$ is given by

$$\mathbf{K}^{-1}\mathbf{c} = \sum_i \theta_i^{-1} \langle \mathbf{c}, \mathbf{z_i} \rangle \mathbf{z_i}, \quad \forall \mathbf{c} \in \mathcal{Y}^n.$$

Now we are able to solve the system of linear operator equations (9) and the functions u_i can be computed from eigenvalues and eigenfunctions of the matrix $\mathbf{K}$, as described in Algorithm 1.

Algorithm 1 L^2-Regularized Function-valued Function Learning Algorithm

Input
Examples:
(function) data $x_i \in (L^2([0,1]))^p$, size n
(function) labels $y_i \in L^2([0,1])$, size n
Parameters: g, T, κ, λ
Eigendecomposition of G, the Gram matrix of the scalar-valued kernel g
Comment: $G = g(x_i, x_j)_{i,j=1}^n \in \mathbb{R}^{n \times n}$
Let $\alpha_i \in \mathbb{R}$ the n eigenvalues
Let $v_i \in \mathbb{R}^n$ the n eigenvectors
Eigendecomposition of the operator $T \in \mathcal{L}(Y)$
Choose κ, the number of computed eigenfunctions
Compute κ ($\delta_i \in \mathbb{R}, w_i \in L^2([0,1])$) pairs of (eigenvalue, eigenfunction)
Eigendecomposition of $\mathbf{K} = G \otimes T$
Comment: $\mathbf{K} = K(x_i, x_j)_{i,j=1}^n \in (\mathcal{L}(Y))^{n \times n}$
The eigenvalues $\theta_i \in \mathbb{R}$, size $n \times \kappa$, are obtained as: $\theta = \alpha \otimes \delta$
The eigenfunctions $\mathbf{z_i} \in (L^2([0,1]))^n$, size $n \times \kappa$, are obtained as: $\mathbf{z} = v \otimes w$
Solution of problem (8) $\mathbf{u} = (\mathbf{K} + \lambda I)^{-1}\mathbf{y}$
Initialize λ: regularization parameter
$\mathbf{u} = \sum_{i=1}^{n \times \kappa} [(\theta_i + \lambda)^{-1} \sum_{j=1}^n \langle z_{ij}, y_j \rangle \mathbf{z}_i]$

To put our algorithm into context, we remind that a crucial question about the applicability of functional data is how one can find an appropriate space and a basis in which the functions can be decomposed in a computationally feasible way while taking into account the functional nature of the data. This is exactly what Algorithm 1 does. In contrast to parametric FDA methods, the basis function here is not fixed in advance but implicitly defined by choosing a *reproducing operator-valued kernel* acting on both input and output data. The spectral decomposition of the block operator kernel matrix naturally allows the assignment of an appropriate basis function to the learning process for representing input and output functions. Moreover, the formulation is flexible enough to be used with different operators and then to be adapted for various applications involving functional data. Also, in the context of nonparametric FDA where the notion of semi-metric plays an important role in modeling functional data, we note that Algorithm 1 is based on computing and choosing a finite number of eigenfunctions. This is strongly related to the semi-metric building scheme in Ferraty and Vieu (2006) which is based on, for example, functional principal components or successive derivatives. Operator-valued kernels constructed from the covariance operator (Kadri et al., 2013b) or the derivative operator will allow to design semi-metrics similar to those just mentioned. In this sense, the eigendecomposition of the block operator kernel matrix offers a new way of producing semi-metrics.

6.2 Generalization Analysis

Here, we provide an analysis of the generalization error of the function-valued function learning model (6) using the notion of algorithmic stability. For more details and results

with the least squares loss and other loss function (including ϵ-sensitive loss and logistic loss), see Audiffren and Kadri (2013). In the case of vector-valued functions, the effort in this area has already produced several successful results, including Baxter (2000), Ando and Zhang (2005), Maurer (2006), and Maurer and Pontil (2013). Yet, these studies have considered only the case of finite-dimensional output spaces, and have focused rather on linear machines than on nonlinear ones. To our knowledge, the first work investigating the generalization performance of nonlinear vector-valued function learning methods when output spaces can be infinite-dimensional is that of Caponnetto and De Vito (2006). In their study, from a theoretical analysis based on the concept of effective dimension, the authors have derived generalization bounds for the learning model (6) when the hypothesis space is an RKHS with operator-valued kernels.

The convergence rates in Caponnetto and De Vito (2006), although optimal in the case of finite-dimensional output spaces, require assumptions on the kernel that can be restrictive in the infinite-dimensional case. Indeed, their proof depends upon the fact that the trace of the operator $K(x, x)$ is finite (K is the operator-valued kernel function), and this restricts the applicability of their results when the output space is infinite-dimensional. To illustrate this, let us consider the identity operator-valued kernel $K(\cdot, \cdot) = k(\cdot, \cdot)I$, where k is a scalar-valued kernel and I is the identity operator. This simple kernel does not satisfy the finite trace condition and therefore the results of Caponnetto and De Vito (2006) cannot be applied in this case. Regarding the examples of operator-valued kernels given in Subsection 5.4, the kernel built from the integral operator satisfies the finite trace condition, while that based on the multiplication operator does not. To address this issue, we first show that our learning algorithm is uniformly stable, and then we derive under mild assumption on the kernel, using a result from Bousquet and Elisseeff (2002), a generalization bound which holds even when the finite trace condition is not satisfied.

We now state and discuss the main assumptions we need to prove a stability-based bound on the generalization error of our method. In the following, we consider a training set $Z = \{(x_1, y_1), ..., (x_n, y_n)\}$ of size n in $\mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from an unknown distribution P, and we denote by $Z^i = Z \setminus (x_i, y_i)$ the set Z from which the couple (x_i, y_i) is removed. We will use a cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. The loss of an hypothesis F with respect to an example (x, y) is then defined as $\ell(y, F, x) = c(F(x), y)$. The generalization error is defined as:

$$R(F) = \int \ell(y, F(x), x) dP(x, y),$$

and the empirical error as:

$$R_{emp}(F, Z) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, F, x_i).$$

A learning algorithm can be viewed as a function which maps a training set Z onto a function F_Z from $\mathcal{X}$ to $\mathcal{Y}$ (Bousquet and Elisseeff, 2002). In our case, F_Z is the solution of the optimization problem (6) which is an instance of the following scheme

$$F_Z = \arg\min_{F \in \mathcal{F}} R_{reg}(F, Z), \tag{12}$$

where $R_{reg}(F, Z) = R_{emp}(F, Z) + \lambda \|F\|^2_{\mathcal{F}}$.

Assumption 1 *$\exists \kappa > 0$ such that $\forall x \in \mathcal{X}$,*

$$\|K(x,x)\|_{op} \leq \kappa^2,$$

where $\|K(x,x)\|_{op} = \sup_{y \in \mathcal{Y}} \frac{\|K(x,x)y\|_{\mathcal{Y}}}{\|y\|_{\mathcal{Y}}}$ is the operator norm of $K(x,x)$ on $L(\mathcal{Y})$.

Assumption 2 *The real function from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$*

$$(x_1, x_2) \mapsto \langle K(x_1, x_2)y_1, y_2 \rangle_{\mathcal{Y}} \text{ is measurable } \forall y_1, y_2 \in \mathcal{Y}.$$

Assumption 3 *The application $(y, f, x) \mapsto \ell(y, F, x)$ is σ-admissible, i.e. convex with respect to F and Lipschitz continuous with respect to $F(x)$, with σ its Lipschitz constant.*

Assumption 4 *$\exists \xi > 0$ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\forall Z$ a training set,*

$$\ell(y, F_Z, x) \leq \xi.$$

Note that Assumption 1 is a direct extension from the scalar-valued to the operator-valued case of the boundedness condition of the kernel function. It replaces and weakens the finite trace assumption of the operator $K(x,x)$ used in Caponnetto and De Vito (2006); see Remark 1 for more details. Assumption 2 was also used by Caponnetto and De Vito (2006) to avoid problems with measurability. This assumption with the fact that $\mathcal{F}$ is separable ensures that all functions in $\mathcal{F}$ are measurable from $\mathcal{X}$ to $\mathcal{Y}$. Assumptions 3 and 4 are the same as those used by Bousquet and Elisseeff (2002) for learning scalar-valued functions. As a consequence of Assumption 1, we immediately obtain the following elementary lemma which allows to control $\|F(x)\|_{\mathcal{Y}}$ with $\|F\|_{\mathcal{F}}$.

Lemma 1 *Let K be a nonnegative operator-valued kernel satisfying Assumption 1. Then $\forall F \in \mathcal{F}$, $\|F(x)\|_{\mathcal{Y}} \leq \kappa \|F\|_{\mathcal{F}}$.*

Proof:

$$\begin{aligned}
\|F(x)\|_{\mathcal{Y}} &= \sup_{\|y\|=1} |\langle F(x), y \rangle_{\mathcal{Y}}| = \sup_{\|y\|=1} |\langle F(\cdot), K(x,\cdot)y \rangle_{\mathcal{F}}| \\
&\leq \|F(\cdot)\|_{\mathcal{F}} \sup_{\|y\|=1} \|K(x,\cdot)y\|_{\mathcal{F}} \leq \|F(\cdot)\|_{\mathcal{F}} \sup_{\|y\|=1} \sqrt{\langle K(x,x)y, y \rangle_{\mathcal{Y}}} \\
&\leq \|F(\cdot)\|_{\mathcal{F}} \sup_{\|y\|=1} \|K(x,x)y\|_{\mathcal{Y}}^{\frac{1}{2}} \leq \|F(\cdot)\|_{\mathcal{F}} \|K(x,x)\|_{op}^{\frac{1}{2}} \leq \kappa \|F\|_{\mathcal{F}}
\end{aligned}$$

■

Now we are ready to state the stability theorem for our function-valued function learning algorithm. This result is a straightforward extension of Theorem 22 in Bousquet and Elisseeff (2002) to the case of infinite-dimensional output spaces. It is worth pointing out that the proof does not differ much from the scalar-valued case and requires only minor modifications to fit the operator-valued kernel approach. For the convenience of the reader, we present in Appendix C the proof taking into account these modifications. Before stating the theorem we would like to recall the definition of uniform algorithmic stability from Bousquet and Elisseeff (2002).

Definition 6 *A learning algorithm $Z \mapsto F_Z$ has uniform stability β with respect to the loss function ℓ if the following holds*

$$\forall n \geq 1,\ \forall 1 \leq i \leq n,\ \forall Z \text{ a training set},\ \ \|\ell(\cdot, F_Z, \cdot) - \ell(\cdot, F_{Z^i}, \cdot)\|_\infty \leq \beta$$

Theorem 7 *Under Assumptions 1, 2 and 3, a learning algorithm that maps a training set Z to the function F_Z defined in (12) is β stable with*

$$\beta = \frac{\sigma^2 \kappa^2}{2\lambda n}.$$

Proof: See Appendix C. ■

β scales as $1/n$. This allows to get a bound on the generalization error using a result from Bousquet and Elisseeff (2002).

Theorem 8 *Let $Z \mapsto F_Z$ be a learning algorithm with uniform stability β with respect to a loss ℓ that satisfies Assumption 4. Then, $\forall n \geq 1$, $\forall\ 0 \leq \delta \leq 1$, the following bound holds with probability at least $1 - \delta$ over the random draw of training samples*

$$R \leq R_{emp} + 2\beta + (4n\beta + \xi)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Proof: See Theorem 12 in Bousquet and Elisseeff (2002). ■

For our learning model (6), we should note that Assumption 3 is in general not satisfied with the least squares loss function $\ell(y, F, x) = \|y - F(x)\|_{\mathcal{Y}}^2$. To address this issue, one can add a boundedness assumption on $\mathcal{Y}$, which is a sufficient condition to prove the uniform stability when Assumption 1 is satisfied.

Assumption 5 $\exists \sigma_y > 0$ *such that* $\|y\|_{\mathcal{Y}} < \sigma_y$, $\forall y \in \mathcal{Y}$.

Lemma 2 *Let $\ell(y, F, x) = \|y - F(x)\|_{\mathcal{Y}}^2$. If Assumptions 1 and 5 hold, then*

$$|\ell(y, F_Z, x) - \ell(y, F_{Z^i}, x)| \leq \sigma \|F_Z(x) - F_{Z^i}(x)\|_{\mathcal{Y}},$$

with $\sigma = 2\sigma_y(1 + \frac{\kappa}{\sqrt{\lambda}})$.

Proof: See Appendix D. ■

This Lemma can replace the Lipschitz property of ℓ in the proof of Theorem 7. Moreover, Assumptions 1 and 5 are sufficient to satisfy Assumption 4 with $\xi = (\sigma/2)^2$ (see Appendix D). We can then use Theorem 7 to prove the uniform stability of our function-valued function learning algorithm with

$$\beta = \frac{2\kappa^2 \sigma_y^2 (1 + \frac{\kappa}{\sqrt{\lambda}})^2}{\lambda n}.$$

Theorem 8 thus gives us a bound on the generalization error of our method equal, with probability at least $1 - \delta$, to

$$R \leq R_{emp} + \frac{4\kappa^2 \sigma_y^2 (1 + \frac{\kappa}{\sqrt{\lambda}})^2}{\lambda n} + \sigma_y^2 (1 + \frac{\kappa}{\sqrt{\lambda}})^2 (\frac{8\kappa^2}{\lambda} + 1)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Remark 1 *It is important to stress that even though the stability analysis of function-valued function learning algorithms follows in a quite straightforward fashion from the earlier results presented in Bousquet and Elisseeff (2002) and provides convergence rates which are not optimal, it allows to derive generalization error bounds with operator-valued kernels for which the trace of the operator $K(x,x)$ is not necessarily finite. Assumption 1 is weaker than the one used in Caponnetto and De Vito (2006) which requires that the operator K_x is Hilbert-Schmidt*[8] *and $\sup_{x\in\mathcal{X}} Tr(K(x,x)) < \kappa$. While the two assumptions are equivalent when the output space $\mathcal{Y}$ is finite dimensional, this is no longer the case when, as in this paper, $\dim \mathcal{Y} = +\infty$. Moreover, we observe that if the assumption of Caponnetto and De Vito (2006) is satisfied, then our Assumption 1 holds (see proof in Appendix E). The converse is not true (see Remark 2 for a counterexample).*

Remark 2 *Note that the operator-valued kernel based on the multiplication operator and described in Subsection 5.4 satisfies Assumption 1 but not the finite trace condition as assumed in Caponnetto and De Vito (2006). Let k be a positive-definite scalar-valued kernel such that $\sup_{x\in\mathcal{X}} k(x,x) < +\infty$, $\mathcal{I}$ an interval of $\mathbb{R}$, $\mu > 0$, and $\mathcal{Y} = L^2(\mathcal{I},\mathbb{R})$. Let $f \in L^\infty(\mathcal{I},\mathbb{R})$ be such that $\|f\|_\infty < \mu$. Consider the following multiplication operator-valued kernel K:*

$$K(x,z)y(\cdot) = k(x,z)f^2(\cdot)y(\cdot) \in \mathcal{Y}.$$

K is a nonnegative operator-valued kernel. While K always satisfies Assumption 1, the Hilbert-Schmidt property of K_x depends on the choice of f and does not hold in general. For instance, let $f(t) = \frac{\mu}{2}(\exp(-t^2)+1)$, then

$$\|K(x,x)\|_{op} \leq \mu^2 k(x,x),$$

and

$$Tr(K(x,x)) = \sum_{j\in\mathbb{N}} \langle K(x,x)y_j, y_j\rangle \geq k(x,x)\frac{\mu}{2}\sum_{i\in\mathbb{N}} \|y_j\|_2^2 = \infty,$$

where $(y_j)_{j\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{Y}$ (which exists since $\mathcal{Y}$ is separable).

7. Experiments

In this experimental section, we essentially aim at illustrating the potential of adopting a functional data analysis perspective for learning multi-output functions when the data are curves. First, we are interested in the problem of acoustic-to-articulatory speech inversion where the goal is to learn vocal tract (VT) time functions from the acoustic speech signal (Mitra et al., 2010). Then we show, through experiments on sound recognition (Rabaoui et al., 2008), that the proposed framework can be applied beyond functional response regression, for problems like multiple functional classification where each sound to be classified is represented by more than one functional parameters.

8. The operator K_x from $\mathcal{Y}$ to $\mathcal{F}$, defined by $y \mapsto K(x,\cdot)y$, $\forall y \in \mathcal{Y}$, is a Hilbert-Schmidt operator if, for some any basis $(y_j)_{j\in\mathbb{N}}$ of $\mathcal{Y}$, it holds that $Tr(K_x^* K_x) = \sum_j \langle K(x,\cdot)y_j, K(x,\cdot)y_j\rangle_{\mathcal{F}} < +\infty$. This is equivalent to saying that the operator $K(x,x) \in \mathcal{L}(\mathcal{Y})$ is of trace class, since by the reproducing property we have $\langle K(x,\cdot)y_j, K(x,\cdot)y_j\rangle = \langle K(x,x)y_j, y_j\rangle_{\mathcal{Y}}$.

The operator-valued kernel used in these experiments is the kernel K defined by Equation (11). We use the inner product in $\mathcal{X}^p$ for the scalar-valued kernel g, where p is the number of functional parameters of a speech or a sound signal. Also, extending real-valued functional kernel, as in Rossi and Villa (2006), to multiple functional inputs could be possible. Eigenvalues δ_i and eigenfunctions w_i of the Hilbert-Schmidt integral operator T associated with the operator-valued kernel K are equal to $\frac{2}{1+\mu_i^2}$ and $\mu_i \cos(\mu_i x) + \sin(\mu_i x)$ respectively, where μ_i are solutions of the equation $\cot \mu = \frac{1}{2}(\mu - \frac{1}{\mu})$. Eigendecomposition of an infinite dimensional operator T is computed in general by solving a differential equation obtained from the equality $Tw_i = \delta_i w_i$.

In order to choose the regularization parameter λ and the number of eigenfunctions κ that guarantee optimal solutions, one may use the cross-validation score based on the one-curve-leave-out prediction error (Rice and Silverman, 1991). Then we choose λ and κ so as to minimize the cross-validation score based on the squared prediction error

$$CV(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \{y_{ij} - \hat{y}_i^{(-i)}(t_{ij})\}^2, \tag{13}$$

where n is the number of functions y_i, y_{ij} is the observed value at time t_{ij}, N_i the number of measurements made on y_i and $\hat{y}_i^{(-i)}$ the predicted curve for the i^{th} function, computed after removing the data for this function.

7.1 Speech Inversion

The problem of speech inversion has received increasing attention in the speech processing community in the recent years (see Schroeter and Sondhi (1994); Mitra et al. (2010); Kadri et al. (2011a) and references therein). This problem, aka acoustic-articulatory inversion, involves inverting the forward process of speech production (see Figure 4). In other words, for a given acoustic speech signal we aim at estimating the underlying sequence of articulatory configurations which produced it (Richmond, 2002). Speech inversion is motivated by several applications in which it is required to estimate articulatory parameters from the acoustic speech signal. For example, in speech recognition, the use of articulatory information has been of interest since speech recognition efficiency can be significantly improved (Kirchoff, 1999). This is due to the fact that automatic speech recognition (ASR) systems suffer from performance degradation in the presence of noise and spontaneous speech. Moreover, acoustic-to-articulatory speech inversion is also useful in many other interesting applications such as speech analysis and synthesis (Toda et al., 2004) or helping individuals with speech and hearing disorders by providing visual feedback (Toutios and Margaritis, 2005).

Most of current research on acoustic-to-articulatory inversion focuses on learning Electromagnetic Articulography (EMA) trajectories from acoustic parameters and frequently uses the MOCHA `fsew0` data set as training and test data (Richmond, 2002). In a recent work, Mitra et al. (2010) suggest the use of the TAsk Dynamics Application (TADA) model (Nam et al., 2004) to generate acoustic-articulatory database which contains synthetic speech and the corresponding vocal tract time functions. Their results show that tract variables can be better candidates than EMA trajectories for articulatory feature based ASR systems. In our experiments, we follow this work by addressing the issue of

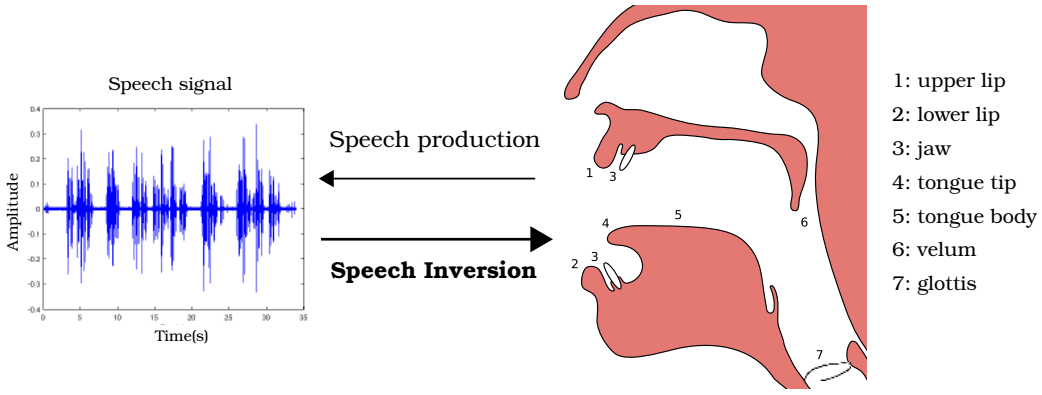

Figure 4: Principle of speech inversion (a.k.a. acoustic-articulatory inversion). Human beings produce an audible speech signal by moving their articulators (*e.g.* tongue, lips, velum, etc.) to modify a source of sound energy in the vocal tract. In performing the inversion mapping, we aim to invert this forward direction of speech production. In other words, we aim to take a speech signal and estimate the underlying articulatory movements which are likely to have created it (Richmond, 2002).

finding the mapping between acoustic parameters and vocal tract variables. In this context, we use Mel-Frequency Cepstral Coefficients (MFCCs) as input and consider as output eight different vocal tract constriction variables, lip aperture (LA), lip protrusion (LP), tongue tip constriction degree (TTCD), tongue tip constriction location (TTCL), tongue body constriction degree (TBCD), tongue body constriction location (TBCL), Velum (VEL) and Glottis (GLO). Table 2 shows the eight vocal tract variables we used in this study and the corresponding constriction organs and articulators (Mitra et al., 2009).

Moreover, articulators move relatively slowly and smoothly, and their movements are continuous. Indeed, the mouth cannot "jump" from one configuration to a completely different one (Richmond, 2002). For this reason, functional data analysis approaches are well suited for the speech inversion task. In other words, even if the measurement process itself is discrete, vocal tract variables are really smooth functions (see Figure 5) rather than vectors and taking into account such prior knowledge on the nature of the data can significantly improve performance. In our proposed method, smoothness is guaranteed by the use of smooth eigenfunctions obtained from the spectral decomposition of the integral operator associated with a Mercer kernel used to construct the operator-valued kernel defined in Equation 11. By this way, our approach does not need the filtering post-processing step, which is necessary in vectorial vocal-tract learning methods to transform the predicted functions on smooth curves and which has the drawback of changing the behavior of the predicted vocal tract time functions.

Various nonlinear acoustic-to-articulatory inversion techniques (Richmond, 2002; Mitra et al., 2010), and particularly kernel-based methods (Toutios and Margaritis, 2005; Mitra

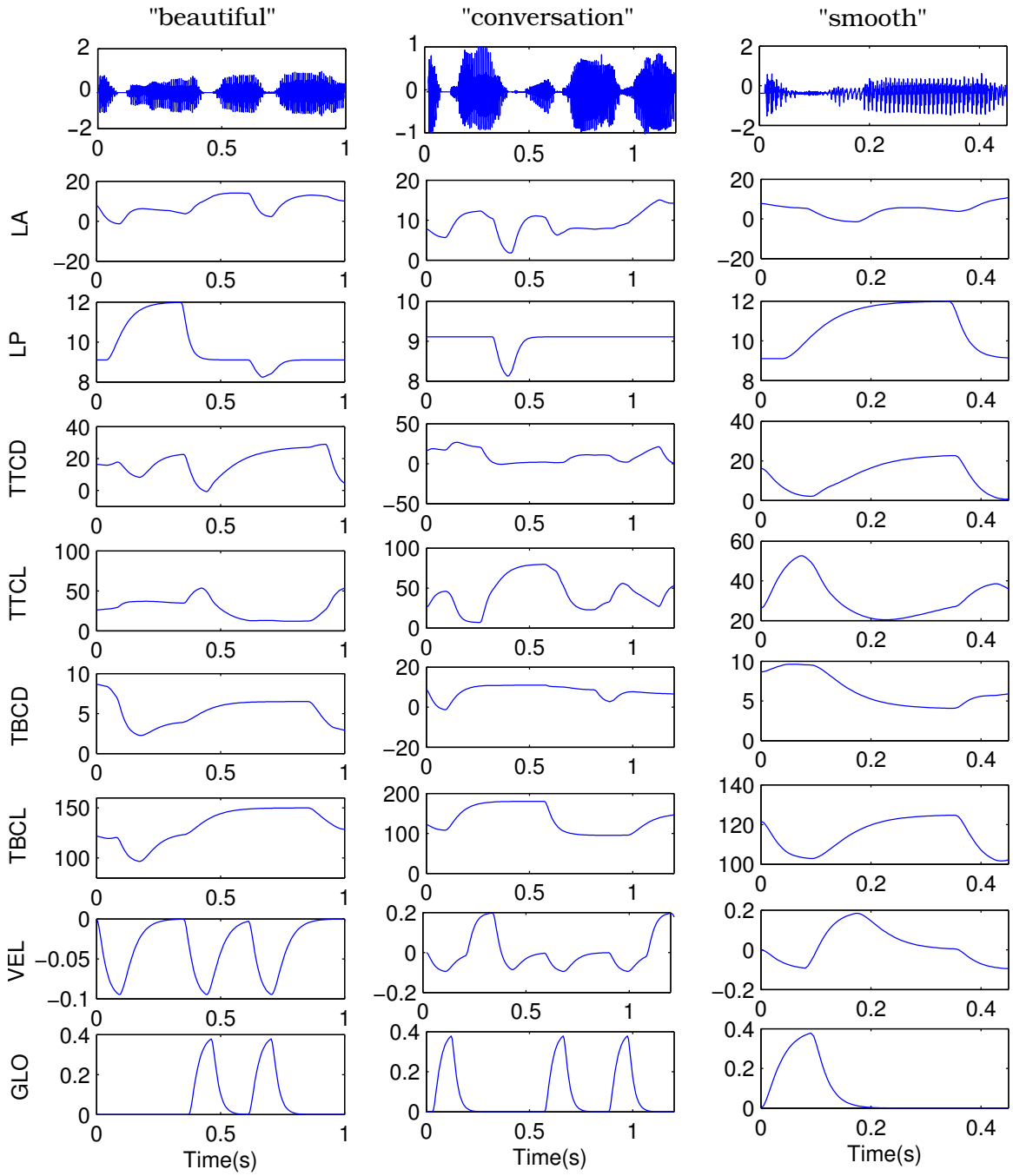

Figure 5: Acoustic waveforms and derived vocal tract time functions for the utterances "beautiful", "conversation" and "smooth". The vocal tract variables are: lip aperture (LA), lip protrusion (LP), tongue tip constriction degree (TTCD), tongue tip constriction location (TTCL), tongue body constriction degree (TBCD), tongue body constriction location (TBCL), Velum (VEL) and Glottis (GLO).

Constriction organ	VT variables	Articulators
lip	lip aperture (LA)	upper lip, lower lip, jaw
	lip protrusion (LP)	
tongue tip	tongue tip constriction degree (TTCD)	tongue body, tip, jaw
	tongue tip constriction location (TTCL)	
tongue body	tongue body constriction degree (TBCD)	tongue body, jaw
	tongue body constriction location (TBCL)	
velum	velum (VEL)	velum
glottis	glottis (GLO)	glottis

Table 2: Constriction organ, vocal-tract (VT) variables and involved articulators (Mitra et al., 2009).

et al., 2009), have been proposed in the literature. In most cases, these works address the articulatory estimation problem within a single-task learning perspective. However, in Richmond (2007) and more recently in Kadri et al. (2011a), the authors put forward the idea that we can benefit from viewing the acoustic-articulatory inversion problem from a multi-task learning perspective. Motivated by comparing our functional operator-valued kernel based approach with multivariate kernel methods, we report on experiments similar to those performed by Mitra et al. (2009) and Kadri et al. (2011a). The tract variables learning technique proposed by Mitra et al. (2009) is based on a hierarchical ε-SVR architecture constructed by associating different SVRs, a SVR for each tract variable. To consider the dependencies between VT time functions, the SVRs corresponding to independent VT variables are first created and then used for constructing the others. Otherwise, the acoustic-to-articulatory method in Kadri et al. (2011a) is based on learning a vector-valued function using a matrix-valued kernel proposed in Caponnetto et al. (2008).

Following Mitra et al. (2010), acoustic-articulatory database is generated by the TADA model (Nam et al., 2004) which is a computational implementation of articulatory phonology. The generated data set consists of acoustic signals for 416 words chosen from the Wisconsin X-ray microbeam data (Westbury et al., 1994) and corresponding Vocal Tract (VT) trajectories sampled at 5 ms. The speech signal was parameterized into 13 Mel-Frequency Cepstral Coefficients. These cepstral coefficients were acquired each 5 ms (synchronized with the TVs) with window duration of 10 ms.

For evaluating the performance of the VT time functions estimation, we use the residual sum of squares error (RSSE) defined as follows

$$RSSE = \int \sum_i \{y_i(t) - \widehat{y}_i(t)\}^2 dt, \tag{14}$$

VT variables	ε-SVR	Multi-task	Functional
LA	2.763	2.341	**1.562**
LP	0.532	**0.512**	0.528
TTCD	3.345	1.975	**1.647**
TTCL	7.752	5.276	**3.463**
TBCD	2.155	2.094	**1.582**
TBCL	15.083	9.763	**7.215**
VEL	0.032	0.034	**0.029**
GLO	**0.041**	0.052	0.064
Total	3.962	2.755	**2.011**

Table 3: Average RSSE for the tract variables using hierarchical ε-SVR (Mitra et al., 2009), the multi-task kernel method (Kadri et al., 2011a) and the proposed functional operator-valued kernel based approach.

where $\widehat{y}_i(t)$ is the prediction of the VT curve $y_i(t)$. Table 3 reports average RSSE results obtained using the hierarchical ε-SVR algorithm (Mitra et al., 2009), the multi-task kernel method (Kadri et al., 2011a) after smoothing the estimated VT trajectories using a Kalman filter as described in Mitra et al. (2009), and the functional operator-valued kernel based approach. The proposed functional approach consistently produced significant performance improvements over the supervised baseline ε-SVR. It also outperforms the discrete multi-task method (Evgeniou et al., 2005; Kadri et al., 2011a) except for the LP and GLO variables. The multi-task and also the ε-SVR methods perform well for these two vocal tract variables and slightly improve our functional approach. This can be explained by the fact that, contrary to other vocal tract variables, LP and GLO time functions are not completely smooth for all times and positions, while our method with the integral operator-valued kernel, as defined in Equation 11, tends to favor the prediction of smooth functions. Building operator-valued kernels suitable for heterogeneous functions, i.e., smooth in some parts and non-smooth in others, could be a good alternative to improve the prediction of these two vocal tract time functions. Note that the number of eigenfunctions κ affects performance. κ has to be well chosen to provide a reasonable approximation of the infinite-dimensional process. In the case of complex output functions, like heterogeneous functions, we need to use many eigenfunctions to have a good approximation, but even for this case, κ remains (very) small compared to the number of examples n.

7.2 Sound Recognition

A second application of the ideas that we present in this paper is sound recognition. Many previous works in the context of sound recognition problem have concentrated on classifying environmental sounds other than speech and music (Dufaux et al., 2000; Peltonen et al., 2002). Such sounds are extremely versatile, including signals generated in domestic, business, and outdoor environments. A system that is able to recognize such sounds may be of great importance for surveillance and security applications (Istrate et al., 2006; Rabaoui et al., 2008). The classification of a sound is usually performed in two steps. First, a pre-

Classes	Number	Train	Test	Total	Duration (s)
Human screams	C1	40	25	65	167
Gunshots	C2	36	19	55	97
Glass breaking	C3	48	25	73	123
Explosions	C4	41	21	62	180
Door slams	C5	50	25	75	96
Phone rings	C6	34	17	51	107
Children voices	C7	58	29	87	140
Machines	C8	40	20	60	184
Total		327	181	508	18mn 14s

Table 4: Classes of sounds and number of samples in the database used for performance evaluation.

processor applies signal processing techniques to generate a set of features characterizing the signal to be classified. Then, in the feature space, a decision rule is implemented to assign a class to a pattern.

Operator-valued kernels can be used in a classification setting by considering the labels y_i to be functions in some function space rather than real values. Similarly to the scalar case, a natural choice for y_i would seem to be the Heaviside step function in $L^2([0,1])$ scaled by a real number. In this context, our method can be viewed as an extension of the Regularized Least Squares Classification (RLSC) algorithm (Rifkin et al., 2003) to the FDA domain (we called it Functional RLSC (Kadri et al., 2011c)). The performance of the proposed algorithm described in Section 6 is evaluated on a data set of sounds collected from commercial databases which include sounds ranging from screams to explosions, such as gun shots or glass breaking, and compared with the RLSC method.

7.2.1 Database Description

As in Rabaoui et al. (2008), the major part of the sound samples used in the recognition experiments is taken from two sound libraries (Leonardo Software; Real World Computing Paternship, 2000). All signals in the database have a 16 bits resolution and are sampled at 44100 Hz, enabling both good time resolution and a wide frequency band, which are both necessary to cover harmonic as well as impulsive sounds. The selected sound classes are given in Table 4, and they are typical of surveillance applications. The number of items in each class is deliberately not equal.

Note that this database includes impulsive and harmonic sounds such as phone rings (C6) and children voices (C7). These sounds are quite likely to be recorded by a surveillance system. Some sounds are very similar to a human listener: in particular, explosions (C4) are pretty similar to gunshots (C2). Glass breaking sounds include both bottle breaking and window breaking situations. Phone rings are either electronic or mechanic alarms.

Temporal representations and spectrograms of some sounds are depicted in Figures 6 and 7. Power spectra are extracted through the Fast Fourier Transform (FFT) every 10 ms from 25 ms frames. They are represented vertically at the corresponding frame indexes.

The frequency range of interest is between 0 and 22 kHz. A lighter shade indicates a higher power value. These figures show that in the considered database we can have both: (1) many similarities between sounds belonging to different classes, (2) diversities within the same class of sounds.

7.2.2 Sound Classification Results

Following Rifkin and Klautau (2004), the 1-vs-all multi-class classifier is selected in these experiments. So we train N (number of classes) different binary classifiers, each one trained to distinguish the data in a single class from the examples in all remaining classes. We run the N classifiers to classify a new example.

The adopted sound data processing scheme is the following. Let $\mathcal{X}$ be the set of training sounds, shared in N classes denoted $\mathcal{C}_1, \ldots, \mathcal{C}_N$. Each class contains m_i sounds, $i = 1, \ldots, N$. Sound number j in class $\mathcal{C}_i$ is denoted $\mathbf{s}_{i,j}$, $(i = 1, \ldots, N, j = 1, \ldots, m_i)$. The pre-processor converts a recorded acoustic signal $\mathbf{s}_{i,j}$ into a time/frequency localized representation. In multivariate methods, this representation is obtained by splitting the signal $\mathbf{s}_{i,j}$ into $T_{i,j}$ overlapping short frames and computing a vector of features $z_{t,i,j}$, $t = 1, \ldots, T_{i,j}$ which characterize each frame. Since the pre-processor is a series of continuous time-localized features, it will be useful to take into account the relationships between feature samples along the time axis and consider dependencies between features. That is why we use a FDA-based approach in which features representing a sound are modeled by functions $z_{i,j}(t)$. In this work, Mel Frequency Cepstral Coefficients (MFCCs) features are used to describe the spectral shape of each signal. These coefficients are obtained using 23 channels Mel filterbank and a Hamming analysis window of length 25 ms with 50% overlap. We also use the energy parameter measured for each window along all the sound signal. So, each sound is characterized by 14 functional parameters: 13 cepstral functions and 1 energy function.

Performance of the proposed functional approach in the context of classification is compared to the results obtained by the RLSC algorithm, see Tables 5 and 6. The performance is measured as the percentage number of sounds correctly recognized and it is given by $(\mathrm{W_r}/\mathrm{T_n}) \times 100\%$, where $\mathrm{W_r}$ is the number of well recognized sounds and $\mathrm{T_n}$ is the total number of sounds to be recognized. The use of the Functional RLSC is fully justified by the results presented here, as it yields consistently a high classification accuracy for the major part of the sound classes.

RLSC setup is similar to that of FLRSC. The major difference is in the modeling of sound features. In RLSC, all the functional parameters which characterize a sound are combined in the same vector which is considered to be in $\mathbb{R}^d$. Functional RLSC considers each input sound as a vector of functions in $(L^2([0,1]))^p$ where p is the number of functional parameters. By using operator-valued kernels rather than scalar-valued ones, we project these functional data into a higher dimensional feature space in which we define a distance measure from the spectral decomposition of the operator-valued kernel, suitable for functions and which allows the learning module to take into account the sequential nature of the data and the dependencies along the time-axis. Moreover, compared to RLSC, κ the number of eigenfunctions in FRLSC can be seen as one more degree of freedom which can be used to improve performance when input data are complex and not represented by a vector in $\mathbb{R}^d$ as usual. Note that the usual scalar case (output space is $\mathbb{R}$ and then $\kappa = 1$) can be recovered

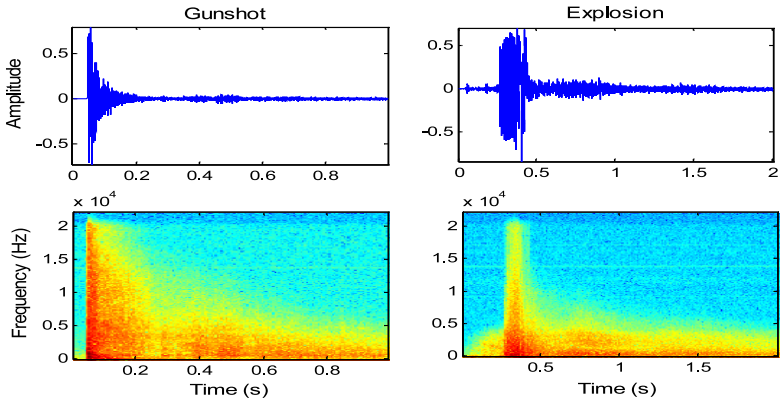

Figure 6: Structural similarities between two different classes. Gunshot and Explosion are two sound signals belonging to two different classes, but they have similar temporal and spectral representations.

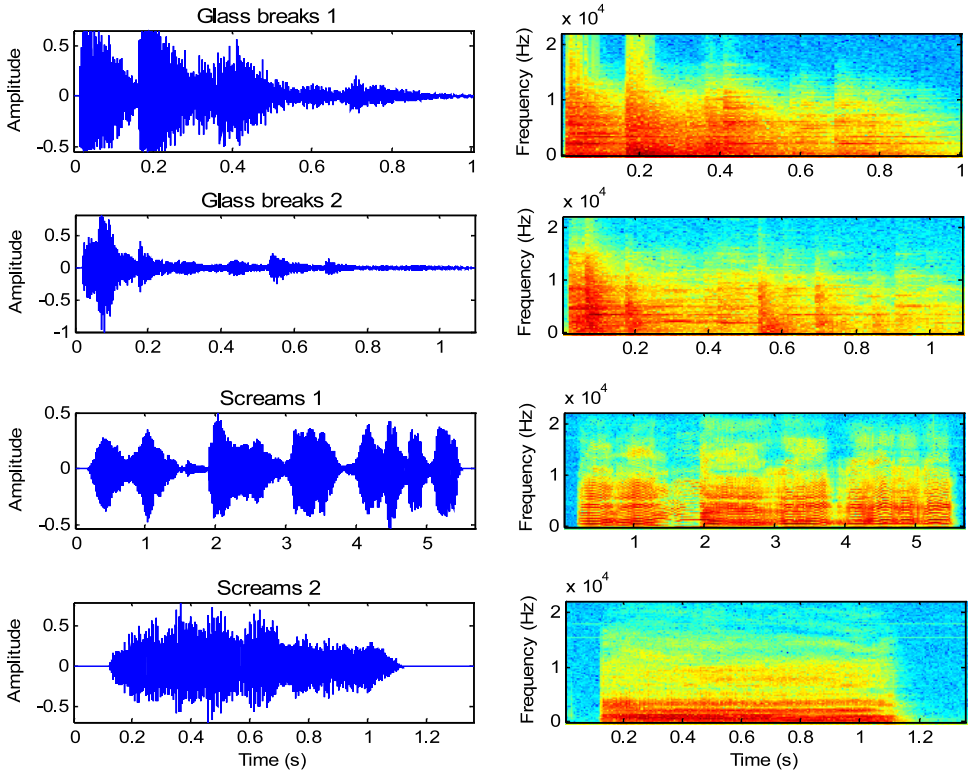

Figure 7: Structural diversity inside the same sound class and between classes. Glass breaks 1 and 2 (resp. Screams 1 and 2) are two sounds from the same class, however they present different temporal (resp. spectral) properties.

	C1	C2	C3	C4	C5	C6	C7	C8
C1	92	4	4.76	0	5.27	11.3	6.89	0
C2	0	52	0	14	0	2.7	0	0
C3	0	20	76.2	0	0	0	17.24	5
C4	0	16	0	66	0	0	0	0
C5	4	8	0	4	84.21	0	6.8	0
C6	4	0	0	0	10.52	86	0	0
C7	0	0	0	8	0	0	69.07	0
C8	0	0	19.04	8	0	0	0	95
Total Recognition Rate = 77.56%								

Table 5: Confusion Matrix obtained when using the Regularized Least Squares Classification (RLSC) algorithm.

	C1	C2	C3	C4	C5	C6	C7	C8
C1	**100**	0	0	2	0	5.3	3.4	0
C2	0	**82**	0	8	0	0	0	0
C3	0	14	**90.9**	8	0	0	3.4	0
C4	0	4	0	**78**	0	0	0	0
C5	0	0	0	1	**89.47**	0	6.8	0
C6	0	0	0	0	10.53	**94.7**	0	0
C7	0	0	0	0	0	0	**86.4**	0
C8	0	0	9.1	3	0	0	0	**100**
Total Recognition Rate = ***90.18%***								

Table 6: Confusion Matrix obtained when using the proposed Functional Regularized Least Squares Classification (FRLSC) algorithm.

from the functional case; for example when the operator-valued kernel is constructed from the identity operator or/and the output space is the space of constant functions.

8. Conclusion

We have presented a learning methodology for nonlinear functional data analysis, which is an extension of scalar-valued and matrix-valued kernel based methodologies to the functional response setting. The problem of functional supervised learning is formalized as the problem of learning an operator between two infinite dimensional scalar-valued Hilbert spaces in a reproducing kernel Hilbert space of function-valued functions. We have introduced a set of rigorously defined operator-valued kernels that can be valuably applied to nonparametric operator learning when input and output data are continuous smooth functions, and we have

showed their use for solving the problem of minimizing a regularized risk functional in the case of functional outputs without the need to discretize covariate and target functions. Our fully functional approach has been successfully applied to the problems of speech inversion and sound recognition, showing that the proposed framework is particularly relevant for audio signal processing applications where attributes are functions and dependent of each other.

In future work, it would be interesting to explore further the potential of our proposed method in other machine learning problems such as collaborative filtering (Abernethy et al., 2009) and structured output prediction (Brouard et al., 2011; Kadri et al., 2013b) by building operator-valued kernels that can capture not only the functional information of responses, but also other types of output structure. In this context, learning the operator-valued kernel would be interesting to find the right model of dependencies between outputs. Recent works in this direction includes the papers of Dinuzzo et al. (2011), Kadri et al. (2012), Sindhwani et al. (2013) and Lim et al. (2015), but further investigations are needed in this area. On the algorithmic side, possible extensions of this work include on-line implementations to deal with the case where the functional data set is made available step by step (Audiffren and Kadri, 2015). Learning, sequentially and without re-training from scratch at each iteration, a new function-valued function for each new observed pair of functional samples would be of practical interest. Finally, although not covered in this paper, the analysis we present is likely to be applicable to learning problems that involve functional data with different functional profiles (*e.g.*, both smooth and spiky functions). Designing nonseparable operator-valued kernels that can exhibit better ability to characterize different smoothing levels is also an interesting future research direction.

Acknowledgments

We thank the anonymous reviewers for several insightful comments that helped improve the original version of this paper. We also thank M. Mbekhta for fruitful discussions on the spectral theory of block operator matrices and A. Rabaoui for providing the sound recognition data set. A large part of this research was done while H.K. was at SequeL (INRIA-Lille) and J.A. at Aix-Marseille Université and LIF. This work was supported by Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER through the 'Contrat de Projets Etat Region (CPER) 2007-2013'. H.K. acknowledges the support of a Junior Researcher Contract No. 4297 from the Nord-Pas-de-Calais region. H.K. and P.P. acknowledge the support of the French National Research Agency (ANR-09-EMER-007 project LAMPADA). H.K. was also supported in part by French grants from the CNRS-LAGIS (ANR KernSig Project). A.R. was supported by the PASCAL2 Network of Excellence, ICT-216886, ANR Project ASAP ANR-09-EMER-001 and the INRIA ARC MABI. P.P. also acknowledges the support of INRIA.

Appendix A. Proof of Theorem 2 - Completion of $\mathcal{F}_0$

We show below how to construct the Hilbert space $\mathcal{F}$ of $\mathcal{Y}$-valued functions, that is the completion of the function-valued pre-Hilbert space $\mathcal{F}_0$. $\mathcal{F}_0 \subset \mathcal{Y}^{\mathcal{X}}$ is the space of all $\mathcal{Y}$-

valued functions F of the form $F(\cdot) = \sum_{i=1}^{n} K(w_i, \cdot)u_i$, where $w_i \in \mathcal{X}$ and $u_i \in \mathcal{Y}$, $i = 1, \ldots, n$. Consider the inner product of the functions $F(\cdot) = \sum_{i=1}^{n} K(w_i, \cdot)u_i$ and $G(\cdot) = \sum_{j=1}^{m} K(z_j, \cdot)v_j$ from $\mathcal{F}_0$ defined as follows

$$\langle F(\cdot), G(\cdot)\rangle_{\mathcal{F}_0} = \langle \sum_{i=1}^{n} K(w_i, \cdot)u_i, \sum_{j=1}^{m} K(z_j, \cdot)v_j\rangle_{\mathcal{F}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \langle K(w_i, z_j)u_i, v_j\rangle_{\mathcal{Y}}.$$

We have shown that $(\mathcal{F}_0, \langle ., .\rangle_{\mathcal{F}_0})$ is a pre-Hilbert space. This pre-Hilbert space is in general not complete, but It can be completed via Cauchy sequences to build the $\mathcal{Y}$-valued reproducing kernel Hilbert space $\mathcal{F}$.

Consider any Cauchy sequence $\{F_n(\cdot)\} \subset \mathcal{F}_0$, for every $w \in \mathcal{X}$ the functional $F(w)$ is bounded, since

$$\begin{aligned}
\|F(w)\|_{\mathcal{Y}} &= \sup_{\|u\|=1} |\langle F(w), u\rangle_{\mathcal{Y}}| = \sup_{\|u\|=1} |\langle F(\cdot), K(w, \cdot)u\rangle_{\mathcal{F}_0}| \\
&\leq \|F(\cdot)\|_{\mathcal{F}_0} \sup_{\|u\|=1} \|K(w, \cdot)u\|_{\mathcal{F}_0} \leq \|F(\cdot)\|_{\mathcal{F}_0} \sup_{\|u\|=1} \sqrt{\langle K(w, w)u, u\rangle_{\mathcal{Y}}} \\
&\leq M_w \|F(\cdot)\|_{\mathcal{F}_0} \text{ with } M_w = \sup_{\|u\|=1} \sqrt{\langle K(w, w)u, u\rangle_{\mathcal{Y}}}.
\end{aligned}$$

Moreover, if the kernel K is Mercer, it is locally bounded (see Carmeli et al., 2010, Proposition 2). It is easy to see that in this case $\|F(w)\|_{\mathcal{Y}} \leq M\|F(\cdot)\|_{\mathcal{F}_0}$, where M here does not depend on w. Consequently,

$$\|F_n(w) - F_m(w)\|_{\mathcal{Y}} \leq M\|F_n(\cdot) - F_m(\cdot)\|_{\mathcal{F}_0}.$$

It follows that $\{F_n(w)\}$ is a Cauchy sequence in $\mathcal{Y}$ and by the completeness of the space $\mathcal{Y}$, there exists a $\mathcal{Y}$-valued function F where, $\forall w \in \mathcal{X}$, $F(w) = \lim_{n\to\infty} F_n(w)$. So the Cauchy sequence $\{F_n(\cdot)\}$ defines a function $F(\cdot)$ to which it is convergent at every point of $\mathcal{X}$.

Let us denote $\mathcal{F}$ the linear space containing all the functions $F(\cdot)$, the limits of Cauchy sequences $\{F_n(\cdot)\} \subset \mathcal{F}_0$, and consider the norm in $\mathcal{F}$ defined by $\|F(\cdot)\|_{\mathcal{F}} = \lim_{n\to\infty} \|F_n(\cdot)\|_{\mathcal{F}_0}$, where $F_n(\cdot)$ is a Cauchy sequence of $\mathcal{F}_0$ converging to $F(\cdot)$. This norm is well defined since it does not depend on the choice of the Cauchy sequence. In fact, suppose that two Cauchy sequences $\{F_n(\cdot)\}$ and $\{G_n(\cdot)\}$ in $\mathcal{F}_0$ define the same function $F(\cdot) \in \mathcal{F}$. Then $\{F_n(\cdot) - G_n(\cdot)\}$ is also a Cauchy sequence and $\forall w \in \mathcal{X}$, $\lim_{n\to\infty} F_n(w) - G_n(w) = 0$. Hence, $\lim_{n\to\infty} \langle F_n(w) - G_n(w), u\rangle_{\mathcal{Y}} = 0$ for any $u \in \mathcal{G}_y$ and using the reproducing property, it follows that $\lim_{n\to\infty} \langle F_n(\cdot) - G_n(\cdot), H(\cdot)\rangle_{\mathcal{F}_0} = 0$ for any function $H(\cdot) \in \mathcal{F}_0$ and thus

$$\lim_{n\to\infty} \|F_n(\cdot) - G_n(\cdot)\|_{\mathcal{F}_0} = 0.$$

Consequently,

$$|\lim_{n\to\infty} \|F_n\| - \lim_{n\to\infty} \|G_n\| | = \lim_{n\to\infty} | \|F_n\| - \|G_n\| | \leq \lim_{n\to\infty} \|F_n - G_n\| = 0.$$

So that for any function $F(\cdot) \in \mathcal{F}$ defined by two different Cauchy sequences $\{F_n(\cdot)\}$ and $\{G_n(\cdot)\}$ in $\mathcal{F}_0$, we have $\lim_{n\to\infty} \|F_n(\cdot)\|_{\mathcal{F}_0} = \lim_{n\to\infty} \|G_n(\cdot)\|_{\mathcal{F}_0} = \|F(\cdot)\|_{\mathcal{F}}$. $\|.\|_{\mathcal{F}}$ has all the

properties of a norm and defines in $\mathcal{F}$ an inner product which on $\mathcal{F}_0$ coincides with $\langle .,. \rangle_{\mathcal{F}_0}$ already defined. It remains to be shown that $\mathcal{F}_0$ is dense in $\mathcal{F}$ which is a complete space.

For any $F(\cdot)$ in $\mathcal{F}$ defined by the Cauchy sequence $F_n(\cdot)$, we have $\lim_{n\to\infty} \|F(\cdot)-F_n(\cdot)\|_{\mathcal{F}} = \lim_{n\to\infty} \lim_{m\to\infty} \|F_m(\cdot)-F_n(\cdot)\|_{\mathcal{F}_0} = 0$. It follows that $F(\cdot)$ is a strong limit of $F_n(\cdot)$ in $\mathcal{F}$ and then $\mathcal{F}_0$ is dense in $\mathcal{F}$. To prove that $\mathcal{F}$ is a complete space, we consider $\{F_n(\cdot)\}$ any Cauchy sequence in $\mathcal{F}$. Since $\mathcal{F}_0$ is dense in $\mathcal{F}$, there exists a sequence $\{G_n(\cdot)\} \subset \mathcal{F}_0$ such that $\lim_{n\to\infty} \|G_n(\cdot) - F_n(\cdot)\|_{\mathcal{F}} = 0$. Besides $\{G_n(\cdot)\}$ is a Cauchy sequence in $\mathcal{F}_0$ and thus defines a function $H(\cdot) \in \mathcal{F}$ which verifies $\lim_{n\to\infty} \|G_n(\cdot) - h(\cdot)\|_{\mathcal{F}} = 0$. So $\{G_n(\cdot)\}$ converges strongly to $H(\cdot)$ and then $\{F_n(\cdot)\}$ also converges strongly to $H(\cdot)$ which means that the space $\mathcal{F}$ is complete. In addition, $K(\cdot,\cdot)$ has the reproducing property in $\mathcal{F}$. To see this, let $F(\cdot) \in \mathcal{F}$, then $F(\cdot)$ is defined by a Cauchy sequence $\{F_n(\cdot)\} \subset \mathcal{F}_0$ and we have from the continuity of the inner product in $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ (endowed with the uniform topology) that, for all $w \in \mathcal{X}$ and $u \in \mathcal{Y}$,

$$\begin{aligned} \langle F(w), u\rangle_{\mathcal{Y}} &= \langle \lim_{n\to\infty} F_n(w), u\rangle_{\mathcal{Y}} = \lim_{n\to\infty} \langle F_n(w), u\rangle_{\mathcal{Y}} = \lim_{n\to\infty} \langle F_n(\cdot), K(w,\cdot)u\rangle_{\mathcal{F}_0} \\ &= \langle \lim_{n\to\infty} F_n(\cdot), K(w,\cdot)u\rangle_{\mathcal{F}} = \langle F(\cdot), K(w,\cdot)u\rangle_{\mathcal{F}}. \end{aligned}$$

Finally, we conclude that $\mathcal{F}$ is a reproducing kernel Hilbert space since $\mathcal{F}$ is a real inner product space that is complete under the norm $\|.\|_{\mathcal{F}}$ defined above, and has $K(\cdot,\cdot)$ as reproducing kernel. ■

Appendix B. Representer Theorem

We provide here a proof of the analog of the representer theorem in the case of function-valued reproducing kernel Hilbert spaces.

Theorem 9 *(representer theorem)*
Let K a nonnegative Mercer operator-valued kernel and $\mathcal{F}$ its corresponding function-valued reproducing kernel Hilbert space. The solution $\widetilde{F}_\lambda \in \mathcal{F}$ of the regularized optimization problem

$$\widetilde{F}_\lambda = \arg\min_{F\in\mathcal{F}} \sum_{i=1}^{n} \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda \|F\|_{\mathcal{F}}^2$$

has the following form

$$\widetilde{F}_\lambda(\cdot) = \sum_{i=1}^{n} K(x_i,\cdot)u_i,$$

where $u_i \in \mathcal{Y}$.

Proof: We use the Frechet derivative which is the strongest notion of derivative in a normed linear space; see, for example, Chapter 4 of Kurdila and Zabarankin (2005). We use the standard notation D_F for the Frechet derivative operator. Let $J_\lambda(F) = \sum_{i=1}^{n} \|y_i - F(x_i)\|_{\mathcal{Y}}^2 + \lambda\|F\|_{\mathcal{F}}^2$ be the functional to be minimized. $\widetilde{F}$ is the operator in $\mathcal{F}$ such that

$\widetilde{F} = \arg\min_{F\in\mathcal{F}} J_\lambda(F) \Rightarrow D_F J_\lambda(\widetilde{F}) = 0$. To compute $D_F J_\lambda(F)$, we use the Gateaux derivative D_G of J_λ with respect to F in the direction H, which is defined by:

$$D_G J_\lambda(F,H) = \lim_{\tau\longrightarrow 0} \frac{J_\lambda(F+\tau H) - J_\lambda(F)}{\tau}.$$

J_λ can be written as $J_\lambda(F) = \sum_{i=1}^{n} G_i(F) + \lambda L(F)$ and using the fact that $D_G J_\lambda(F,H) = \langle D_G J_\lambda(F), H\rangle$ we obtain

i. $L(F) = \|F\|_{\mathcal{F}}^2$

$$\lim_{\tau\longrightarrow 0} \frac{\|F+\tau H\|_{\mathcal{F}}^2 - \|F\|_{\mathcal{F}}^2}{\tau} = 2\langle F, H\rangle \implies D_G L(F) = 2F.$$

ii. $G_i(F) = \|y_i - F(x_i)\|_{\mathcal{Y}}^2$

$$\lim_{\tau\longrightarrow 0} \frac{\|y_i - F(x_i) - \tau H(x_i)\|_{\mathcal{Y}}^2 - \|y_i - F(x_i)\|_{\mathcal{Y}}^2}{\tau} = -2\langle y_i - F(x_i), H(x_i)\rangle_{\mathcal{Y}}$$
$$= -2\langle K(x_i,\cdot)(y_i - F(x_i)), H\rangle_{\mathcal{F}} = -2\langle K(x_i,\cdot)u_i, H\rangle_{\mathcal{F}} \text{ with } u_i = y_i - F(x_i)$$
$$\implies D_G G_i(F) = -2K(x_i,\cdot)u_i.$$

When the kernel K is Mercer, Corollary 4.1.1 in Kurdila and Zabarankin (2005) can be applied to show that J_λ is Fréchet differentiable and that, $\forall F \in \mathcal{F}$, $D_F J_\lambda(F) = D_G J_\lambda(F)$. Using (i), (ii), and $D_F J_\lambda(\widetilde{F}) = 0 \implies \widetilde{F}(\cdot) = \frac{1}{\lambda}\sum_{i=1}^{n} K(x_i,\cdot)u_i$. ■

Appendix C. Proof of Theorem 7

We show here that under Assumptions 1, 2 and 3, a learning algorithm that maps a training set Z to the function F_Z defined in (12) is β stable with $\beta = \frac{\sigma^2\kappa^2}{2n\lambda}$. First, since ℓ is convex with respect to F, we have $\forall 0 \le t \le 1$

$$\ell(y, F_Z + t(F_{Z^i} - F_Z), x) - \ell(y, F_Z, x) \le t\left(\ell(y, F_{Z^i}, x) - \ell(y, F_Z, x)\right).$$

Then, by summing over all couples (x_k, y_k) in Z^i,

$$R_{emp}(F_Z + t(F_{Z^i} - F_Z), Z^i) - R_{emp}(F_Z, Z^i) \le t\left(R_{emp}(F_{Z^i}, Z^i) - R_{emp}(F_Z, Z^i)\right). \quad (15)$$

Symmetrically, we also have

$$R_{emp}(F_{Z^i} + t(F_Z - F_{Z^i}), Z^i) - R_{emp}(F_{Z^i}, Z^i) \le t\left(R_{emp}(F_Z, Z^i) - R_{emp}(F_{Z^i}, Z^i)\right). \quad (16)$$

Thus, by summing (15) and (16), we obtain

$$\begin{aligned} &R_{emp}(F_Z + t(F_{Z^i} - F_Z), Z^i) - R_{emp}(F_Z, Z^i) \\ &\qquad + R_{emp}(F_{Z^i} + t(F_Z - F_{Z^i}), Z^i) - R_{emp}(F_{Z^i}, Z^i) \le 0. \end{aligned} \quad (17)$$

Now, by definition of F_Z and F_{Z^i},

$$\begin{aligned} R_{reg}(F_Z, Z) - R_{reg}(F_Z + t(F_{Z^i} - F_Z), Z) \\ + R_{reg}(F_{Z^i}, Z^i) - R_{reg}(F_{Z^i} + t(F_Z - F_{Z^i}), Z^i) \leq 0. \end{aligned} \tag{18}$$

Combining (17) and (18), we find

$$\begin{aligned} &\ell(y_i, F_Z, x_i) - \ell(y_i, F_Z + t(F_{Z^i} - F_Z), x_i) \\ &\quad + n\lambda \left(\|F_Z\|_{\mathcal{F}}^2 - \|F_Z + t(F_{Z^i} - F_Z)\|_{\mathcal{F}}^2 + \|F_{Z^i}\|_{\mathcal{F}}^2 - \|F_{Z^i} + t(F_Z - F_{Z^i})\|_{\mathcal{F}}^2 \right) \leq 0. \end{aligned} \tag{19}$$

Moreover, we have

$$\begin{aligned} &\|F_Z\|_{\mathcal{F}}^2 - \|F_Z + t(F_{Z^i} - F_Z)\|_{\mathcal{F}}^2 + \|F_{Z^i}\|_{\mathcal{F}}^2 - \|F_{Z^i} + t(F_Z - F_{Z^i})\|_{\mathcal{F}}^2 \\ &\quad = \|F_Z\|_{\mathcal{F}}^2 - \|F_Z\|_{\mathcal{F}}^2 - t^2\|F_{Z^i} - F_Z\|_{\mathcal{F}}^2 - 2t\langle F_Z, F_{Z^i} - F_z\rangle_{\mathcal{F}} \\ &\quad + \|F_{Z^i}\|_{\mathcal{F}}^2 - \|F_{Z^i}\|_{\mathcal{F}}^2 - t^2\|F_Z - F_{Z^i}\|_{\mathcal{F}}^2 - 2t\langle F_{Z^i}, F_Z - F_{Z^i}\rangle_{\mathcal{F}} \\ &\quad = -2t^2\|F_{Z^i} - F_Z\|_{\mathcal{F}}^2 - 2t\langle F_Z, F_{Z^i} - F_z\rangle_{\mathcal{F}} - 2t\langle F_{Z^i}, F_Z - F_{Z^i}\rangle_{\mathcal{F}} \\ &\quad = -2t^2\|F_{Z^i} - F_Z\|_{\mathcal{F}}^2 + 2t\|F_{Z^i} - F_Z\|_{\mathcal{F}}^2 \\ &\quad = 2t(1-t)\|F_{Z^i} - F_Z\|_{\mathcal{F}}^2. \end{aligned} \tag{20}$$

Hence, since ℓ is σ-Lipschitz continuous with respect to $F(x)$, we obtain from (19) and (20), $\forall t \in]0, 1[$,

$$\begin{aligned} \|F_Z - F_{Z^i}\|_{\mathcal{F}}^2 &\leq \frac{1}{2t(1-t)} \left(\|F_Z\|_{\mathcal{F}}^2 - \|F_Z + t(F_{Z^i} - F_Z)\|_{\mathcal{F}}^2 + \|F_{Z^i}\|_{\mathcal{F}}^2 - \|F_{Z^i} + t(F_Z - F_{Z^i})\|_{\mathcal{F}}^2 \right) \\ &\leq \frac{1}{2t(1-t)n\lambda} \big(\ell(y_i, F_Z + t(F_{Z^i} - F_Z), x_i) - \ell(y_i, F_Z, x_i) \big) \\ &\leq \frac{\sigma}{2(1-t)n\lambda} \|F_{Z^i}(x_i) - F_Z(x_i)\|_{\mathcal{Y}}. \end{aligned}$$

In particular, when t tends to 0, we have

$$\|F_Z - F_{Z^i}\|_{\mathcal{F}}^2 \leq \frac{\sigma}{2n\lambda} \|F_{Z^i}(x_i) - F_Z(x_i)\|_{\mathcal{Y}} \leq \frac{\sigma\kappa}{2n\lambda} \|F_{Z^i} - F_Z\|_{\mathcal{F}},$$

which gives that

$$\|F_Z - F_{Z^i}\|_{\mathcal{F}} \leq \frac{\sigma\kappa}{2n\lambda}.$$

This implies that, $\forall (x, y)$,

$$|\ell(y, F_Z, x) - \ell(y, F_{Z^i}, x)| \leq \sigma \|F_Z(x) - F_{Z^i}(x)\|_{\mathcal{Y}} \leq \sigma\kappa \|F_Z - F_{Z^i}\|_{\mathcal{F}} \leq \frac{\sigma^2\kappa^2}{2n\lambda},$$

which concludes the proof. ■

Appendix D. Proof of Lemma 2

We show here that Assumption 4 is satisfied for the least squares loss function when Assumption 5 holds and use that to prove Lemma 2. First, note that ℓ is convex with respect to its second argument. Since $\mathcal{F}$ is a vector space, $0 \in \mathcal{H}$. Thus,

$$\lambda \|F_Z\|^2 \leq R_{reg}(F_Z, Z) \leq R_{reg}(0, Z) \leq \frac{1}{n}\sum_{k=1}^{n} \|y_k\|^2 \leq \sigma_y^2, \tag{21}$$

where we used the definition of F_Z (see Equation 12) and the bound on Y (Assumption 5). This inequality is uniform over Z, and thus holds for F_{Z^i}. Moreover, $\forall x \in \mathcal{X}$,

$$\|F_Z(x)\|_{\mathcal{Y}}^2 = \langle F_Z(x), F_Z(x)\rangle_{\mathcal{Y}} = \langle K(x,x)F_Z, F_Z\rangle_{\mathcal{F}} \leq \|K(x,x)\|_{op}\|F_Z\|_{\mathcal{F}}^2 \leq \kappa^2 \frac{\sigma_y^2}{\lambda}.$$

Hence, using Lemma 1 and (21), we obtain

$$\|y - F_Z(x)\|_{\mathcal{Y}} \leq \|y\|_{\mathcal{Y}} + \|f_Z(x)\|_{\mathcal{Y}} \leq \sigma_y + \kappa\frac{\sigma_y}{\sqrt{\lambda}},\ \forall (x,y) \in \mathcal{X}\times\mathcal{Y}.$$

Then it follows that

$$\begin{aligned}
&\left|\|y - F_Z(x)\|_{\mathcal{Y}}^2 - \|y - F_{Z^i}(x)\|_{\mathcal{Y}}^2\right| \\
&= \left|\|y - F_Z(x)\|_{\mathcal{Y}} - \|y - F_{Z^i}(x)\|_{\mathcal{Y}}\right| \left|\|y - F_Z(x)\|_{\mathcal{Y}} + \|y - F_{Z^i}(x)\|_{\mathcal{Y}}\right| \\
&\leq 2\sigma_y(1 + \frac{\kappa}{\sqrt{\lambda}})\|F_Z(x) - F_{Z^i}(x)\|_{\mathcal{Y}}.
\end{aligned}$$

■

Appendix E. Proof of Remark 1

We show here that if the finite trace assumption of the operator $K(x,x)$ in Caponnetto and De Vito (2006) is satisfied, then our Assumption 1 on the kernel holds. Let K be an operator-valued kernel satisfying the hypotheses of Caponnetto and De Vito (2006), i.e K_x is Hilbert-Schmidt and $\sup_{x\in\mathcal{X}} Tr(K(x,x)) < +\infty$. Then, $\exists \eta > 0$, $\forall x \in \mathcal{X}$, $\exists \left(e_j^x\right)_{j\in\mathbb{N}}$ an orthonormal basis of $\mathcal{Y}$, $\exists \left(h_j^x\right)_{j\in\mathbb{N}}$ an orthogonal family of $\mathcal{F}$ with $\sum_{j\in\mathbb{N}} \|h_j^x\|_{\mathcal{F}}^2 \leq \eta$ such that $\forall y \in \mathcal{Y}$,

$$K(x,x)y = \sum_{j,\ell} \left\langle h_j^x, h_\ell^x\right\rangle_{\mathcal{F}} \left\langle y, e_j^x\right\rangle_{\mathcal{Y}} e_\ell^x.$$

Thus, $\forall i \in \mathbb{N}$,

$$K(x,x)e_i^x = \sum_{j,\ell} \left\langle h_j^x, h_\ell^x\right\rangle_{\mathcal{F}} \left\langle e_i^x, e_j^x\right\rangle_{\mathcal{Y}} e_\ell^x = \sum_{\ell} \langle h_i^x, h_\ell^x\rangle_{\mathcal{F}}\, e_\ell^x.$$

Hence

$$\begin{aligned}
\|K(x,x)\|_{op}^2 &= \sup_{i\in\mathbb{N}} \|K(x,x)e_i^x\|_{\mathcal{Y}}^2 = \sup_{i\in\mathbb{N}} \sum_{j,\ell} \langle h_i^x, h_\ell^x\rangle_{\mathcal{F}} \left\langle h_i^x, h_j^x\right\rangle_{\mathcal{F}} \left\langle e_j^x, e_\ell^x\right\rangle_{\mathcal{Y}} \\
&= \sup_{i\in\mathbb{N}} \sum_{\ell} (\langle h_i^x, h_\ell^x\rangle_{\mathcal{F}})^2 \leq \sup_{i\in\mathbb{N}} \|h_i^x\|_{\mathcal{F}}^2 \sum_{\ell} \|h_\ell^x\|_{\mathcal{F}}^2 \leq \eta^2.
\end{aligned}$$

■

References

J. Abernethy, F. Bach, T. Evgeniou, and J. P. Vert. A new approach to collaborative filtering: operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

D. Alpay, A. Dijksma, J. Rovnyak, and H. de Snoo. *Schur Functions, Operator Colligations, and Reproducing Kernel Pontryagin Spaces*, volume 96 of *Operator theory: Advances and Applications*. Birkhäuser Verlag, 1997.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. *Foundation and Trends in Machine Learning*, 4(3):195–266, 2012.

R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

A. Asif and J. Moura. Block matrices with L-block-banded inverse: inversion algorithms. *IEEE Transactions on Signal Processing*, 53(2):630–642, February 2005.

J. Audiffren and H. Kadri. Stability of multi-task kernel regression algorithms. In *Asian Conference on Machine Learning (ACML)*, volume 29, pages 1–16, 2013.

J. Audiffren and H. Kadri. Online learning with operator-valued kernels. In *European symposium on artificial neural networks (ESANN)*, 2015.

L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

S. Ben-david and R. Schuller-Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, 73:273–287, 2008.

P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube. Articulatory synthesis and perception of plosive-vowel syllables with virtual consonant targets. In *Interspeech*, pages 1017–1020. ISCA, 2010.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

L. Breiman and J. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society, Series B*, 59:3–54, 1997.

C. Brouard, F. d'Alché-Buc, and M. Szafranski. Semi-supervised penalized output kernel regression for link prediction. In *International Conference on Machine Learning (ICML)*, 2011.

S. Canu, X. Mary, and A. Rakotomamonjy. Functional learning through kernel. *in Advances in Learning Theory: Methods, Models and Applications. NATO Science Series III: Computer and Systems Sciences*, 2003.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 68:1615–1646, 2008.

C. Carmeli, E. De Vito, and A. Toigo. Vector-valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

C. Carmeli, E. De Vito, and A. Toigo. Vector-valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

J. M. Chiou, H. G. Müller, and J. L. Wang. Functional response models. *Statistica Sinica*, 14:675–693, 2004.

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*, 12:136–154, 1982.

F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. In *International Conference on Machine Learning (ICML)*, 2011.

A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. In *European Signal Processing Conference (EUSIPCO)*, pages 1033–1036, 2000.

H. Dym. *J Contractive Matrix Functions, Reproducing Kernel Spaces and Interpolation*. American Mathematical Society, 1989.

T. Evgeniou and M. Pontil. Regularized multi-task learning. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, 2004.

T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

J. Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997.

F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2):161–173, 2003.

F. Ferraty and P. Vieu. Nonparametric models for functional data, with applications in regression, time series prediction and curves discrimination. *Journal of Nonparametric Statistics*, 16:111–125, 2004.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer Verlag, 2006.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society B*, 55:757–796, 1993.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

T. C. Hesterberg, N. H. Choi, L. Meier, and C. Fraley. Least angle and ℓ_1 penalized regression: a review. *Statistics Surveys*, 2:61–93, 2008.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, 2012.

D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J. F. Serignat. Information extraction from sound for medical telemonitoring. *IEEE Transactions on Information Technology in Biomedicine*, 10:264–274, 2006.

G. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society Series B*, 64(3):411–432, 2002.

T. Jebara. Multi-task feature and kernel selection for SVMs. In *International Conference on Machine Learning (ICML)*, 2004.

H. Kadri, E. Duflos, Ph. Preux, S. Canu, and M. Davy. Nonlinear functional regression: a functional RKHS approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 111–125, 2010.

H. Kadri, E. Duflos, and Ph. Preux. Learning vocal tract variables with multi-task kernels. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011a.

H. Kadri, E. Duflos, Ph. Preux, and S. Canu. Multiple functional regression with both discrete and continuous covariates. In *International Workshop on Functional and Operatorial Statistics (IWFOS)*. Springer, 2011b.

H. Kadri, A. Rabaoui, Ph. Preux, E. Duflos, and A. Rakotomamonjy. Functional regularized least squares classification with operator-valued kernels. In *International Conference on Machine Learning (ICML)*, pages 993–1000, 2011c.

H. Kadri, A. Rakotomamonjy, F. Bach, and Ph. Preux. Multiple operator-valued kernel learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

H. Kadri, S. Ayache, C. Capponi, S. Koço, F. X. Dupé, and E. Morvant. The multi-task learning view of multimodal data. In *Asian Conference on Machine Learning (ACML)*, pages 261–276, 2013a.

H. Kadri, M. Ghavamzadeh, and Ph. Preux. A generalized kernel approach to structured output learning. In *International Conference on Machine Learning (ICML)*, 2013b.

K. Kirchoff. *Robust Speech Recognition Using Articulatory Information*. PhD thesis, University of Bielefeld, 1999.

A. Kurdila and M. Zabarankin. *Convex Functional Analysis*. Birkhauser Verlag, 2005.

Leonardo Software. http://www.leonardosoft.com.

D. J. Levitin, R. L. Nuzzo, B. W. Vines, and J. O. Ramsay. Introduction to functional data analysis. *Canadian Psychology*, 48(3):135–155, 2007.

H. Lian. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *The Canadian Journal of Statistics*, 35:597–606, 2007.

N. Lim, Y. Senbabaoglu, G. Michailidis, and F. d'Alché-Buc. OKVAR-Boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks. *Bioinformatics*, 29(11):1416–1423, 2013.

N. Lim, F. d'Alché-Buc, C. Auliac, and G. Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3):489–513, 2015.

A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.

A. Maurer and M. Pontil. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory (COLT)*, pages 55–76, 2013.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005a.

C. A. Micchelli and M. Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 921–928, 2005b.

H. Q. Minh and V. Sindhwani. Vector-valued manifold regularization. In *International Conference on Machine Learning (ICML)*, 2011.

H. Q. Minh, S. H. Kang, and T. M. Le. Image and video colorization using vector-valued reproducing kernel Hilbert spaces. *Journal of Mathematical Imaging and Vision*, 37(1): 49–65, 2010.

H. Q. Minh, L. Bazzani, and V. Murino. A unifying framework for vector-valued manifold regularization and multi-view learning. In *International Conference on Machine Learning (ICML)*, 2013.

V. Mitra, Y. Ozbek, H. Nam, X. Zhou, and C. Y. Espy-Wilson. From acoustics to vocal tract time functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4497–4500, 2009.

V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein. Retrieving tract variables from acoustics: a comparison of different machine learning strategies. *IEEE Journal of Selected Topics in Signal Processing*, pages 1027–1045, 2010.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18, 2012.

H. G. Müller. Functional modeling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32:223–240, 2005.

H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. TADA: an enhanced, portable task dynamics model in MATLAB. *Journal of the Acoustical Society of America*, 115:2430, 2004.

A. W. Naylor and G. R. Sell. *Linear Operator Theory in Engineering and Science*. Holt, Rinehart and Winston, Inc., New York, 1971.

J. Oliva, B. Poczos, and J. Schneider. Distribution to distribution regression. In *International Conference on Machine Learning (ICML)*, 2013.

V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa. Computational audiroty scene recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

B. Poczos, L. Xiong, D. Sutherland, and J. Schneider. Support distribution machines. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 2012.

B. Poczos, A. Rinaldo, A. Singh, and L. Wasserman. Distribution-free distribution regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 507–515, 2013.

L. Prchal and P. Sarda. Spline estimator for the functional linear regression with functional response. *Preprint*, 2007.

C. Preda. Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137:829–840, 2007.

A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze. Using one-class SVMs and wavelets for audio surveillance. *IEEE Transactions on Information Forensics and Security*, 3(4): 763–775, 2008.

J. O. Ramsay. When the data are functions. *Psychometrika*, 47:379–396, 1982.

J. O. Ramsay and J. L. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society*, B(53):539–572, 1991.

J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer Verlag, New York, 2002.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, 2nd ed.* Springer Verlag, New York, 2005.

Real World Computing Paternship. Cd-sound scene database in real acoustical environments, 2000. URL `http://tosa.mri.co.jp/sounddb/indexe.htm`.

M. Reisert and H. Burkhardt. Learning equivariant functions with matrix-valued kernels. *Journal of Machine Learning Research*, 8:385–408, 2007.

J. A. Rice. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14:613–629, 2004.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B*, 53(1):233–243, 1991.

K. Richmond. *Estimating Articulatory Parameters from the Acoustic Speech Signal*. PhD thesis, The Center for Speech Technology Research, Edinburgh University, 2002.

K. Richmond. A multitask learning perspective on acoustic-articulatory inversion. In *Interspeech*. ISCA, 2007.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

R. Rifkin, G. Yeo, and T. Poggio. Regularized least squares classification. *Advances in Learning Theory: Methods, Model and Applications NATO Science Series III: Computer and Systems Sciences*, 190:131–153, 2003.

F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7–9):730–742, 2006.

W. Rudin. *Functional Analysis*. McGraw-Hill Science, 1991.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2002.

B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Müller, G. Rätsch, and A. J. Smola. Input space vs. feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5):1000–1017, 1999.

J. Schroeter and M. M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2:133–150, 1994.

L. Schwartz. Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathématique*, 13:115–256, 1964.

E. Senkene and A. Tempel'man. Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 1973.

J. Shawe-Taylor and N. Cristanini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

J. Q. Shi and T. Choi. *Gaussian Process Regression Analysis for Functional Data*. CRC Press, 2011.

T. Simila and J. Tikka. Input selection and shrinkage in multiresponse linear regression. *Computational Statistics and Data Analysis*, 52:406–422, 2007.

V. Sindhwani, H. Q. Minh, and A. C. Lozano. Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and granger causality. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

S. Szedmak, J. Shawe-Taylor, and E. Parado-Hernandez. Learning via linear operators: maximum margin regression; multiclass and multiview learning at one-class complexity. Technical report, PASCAL, Southampton, UK, 2006. URL `http://arxiv.org/pdf/1106.6251v1`.

T. Toda, A. W. Black, and K. Tokuda. Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In *ISCA Speech Synthesis Workshop*, 2004.

A. Toutios and K. Margaritis. A support vector approach to the acoustic-to-articulatory mapping. In *Interspeech*, pages 3221–3224. ISCA, 2005.

C. Tretter. *Spectral theory of block operator matrices and applications*. Imperial College Press, London, 2008.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47:349–363, 2005.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics (SIAM), 1990.

J. R. Westbury, G. Turner, and J. Dembovski. X-ray microbeam speech production database user's handbook, 1994.

F. Yao, H. G. Müller, and J. L. Wang. Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33:2873–2903, 2005.

H. Zhang, Y. Xu, and Q. Zhang. Refinement of operator-valued reproducing kernels. *Journal of Machine Learning Research*, 13:91–136, 2012.

X. Zhao, J. S. Marron, and M. T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14:789–808, 2004.