Quadratically Constrained Quadratic Programming for Subspace Selection in Kernel Regression Estimation

Marco Signoretto, Kristiaan Pelckmans, and Johan A.K. Suykens

K.U. Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)
{Marco.Signoretto,Kristiaan.Pelckmans,Johan.Suykens}@esat.kuleuven.be

Abstract. In this contribution we consider the problem of regression estimation. We elaborate on a framework based on functional analysis giving rise to structured models in the context of reproducing kernel Hilbert spaces. In this setting the task of input selection is converted into the task of selecting functional components depending on one (or more) inputs. In turn the process of learning with embedded selection of such components can be formalized as a convex-concave problem. This results in a practical algorithm that can be implemented as a quadratically constrained quadratic programming (QCQP) optimization problem. We further investigate the mechanism of selection for the class of linear functions, establishing a relationship with LASSO.

1 Introduction

Problems of model selection constitute one of the major research challenges in the field of machine learning and pattern recognition. In particular, when modelling in the presence of multivariate data, the problem of input selection is largely unsolved for many classes of algorithms [1]. An interesting convex approach was found in the use of l_1-norm, resulting in sparseness amongst the optimal coefficients. This sparseness is then interpreted as an indication of non-relevance. The *Least Absolute Shrinkage and Selection Operator* (LASSO) [2] was amongst the first to advocate this approach, but also the literature on basis pursuit [3] employs a similar strategy. However, these procedures mainly deal with the class of linear models. On the other hand, when the functional dependency is to be found in a broader class, there is a lack of principled methodologies for modeling with embedded approaches for pruning irrelevant inputs.

On a different track, recent advances in convex optimization have been exploited to tackle general problems of model selection and many approaches have been proposed [4],[5]. The common denominator of the latter is conic programming, a class of convex problems broader than linear and quadratic programming [6].

In the present paper we focus on regression estimation. By taking a functional analysis perspective we present an optimization approach, based on a QCQP problem, that permits to cope with the problem of selection in a principled way. In our framework the search of a functional dependency is performed

in an hypothesis space formed by the direct sum of orthogonal subspaces. In particular, functional ANOVA models, a popular way of modelling in presence of multivariate data, are a special case of this setting. Roughly speaking, since each subspace is composed of functions which depend on a subset of inputs, these models provide a natural framework to deal with the issue of input selection. Subsequently we present our optimization approach, inspired by [4]. In a nutshell, it provides at the same time a way to optimize the hypothesis space and to find in it the minimizer of a regularized least squares functional. This is achieved by optimizing a family of parametrized reproducing kernels each corresponding to an hypothesis space equipped with an associated inner-product. This approach leads to sparsity in the sense that it corresponds to select a subset of subspaces forming the class of models. This paper is organized as follows. In Section 2 we introduce the problem of regression estimation. Section 3 presents the abstract class of hypothesis spaces we deal with. Section 4 presents our approach for learning the functional dependency with embedded selection of relevant components. Before illustrating the method on a real-life dataset (Section 6), we uncover in Section 5 the mechanism of selection for a simple case, highlighting the relation with LASSO.

2 Regression Estimation in Reproducing Kernel Hilbert Spaces

In the following we use boldface for vectors and capital letters for matrices, operators and functionals. We further denote with $[\boldsymbol{w}]_i$ and $[A]_{ij}$ respectively the $i-$th component of $\boldsymbol{w}$ and the element (i,j) of A.

The general search for functional dependency in the context of regression can be modelled as follows [7]. It is generally assumed that for each $\boldsymbol{x} \in \mathcal{X}$, drawn from a probability distribution $p(\boldsymbol{x})$, a corresponding $y \in \mathcal{Y}$ is attributed by a *supervisor* according to an unknown conditional probability $p(y|\boldsymbol{x})$. The typical assumption is that the underlying process is deterministic but the output measurements are affected by (Gaussian) noise. The learning algorithm is then required to estimate the *regression* function i.e. the expected value $h(\boldsymbol{x}) = \int yp(y|\boldsymbol{x})dy$, based upon a training set $\mathcal{Z}^N = \{(\boldsymbol{x}_1, y_1) \times \ldots \times (\boldsymbol{x}_N, y_N)\}$ of N i.i.d. observations drawn according to $p(\boldsymbol{x}, y) = p(y|\boldsymbol{x})p(\boldsymbol{x})$. This is accomplished by finding from an appropriate set of functions $\mathcal{H}$, the minimizer $\hat{f}$ of the *empirical risk* associated to $\mathcal{Z}^N$: $R_{\text{emp}}f \triangleq \frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i) - y_i)^2$. When the learning process takes place in a *reproducing kernel Hilbert space* $\mathcal{H}$ and, more specifically, in $\mathcal{I}_r = \{f \in \mathcal{H} \ : \ \|f\| \leq r\}$ for some $r \geq 0$, the minimizer of the empirical risk is guaranteed to exist and the crucial choice of r can be guided by a number of probabilistic bounds [8]. For fixed r the minimizer of $R_{\text{emp}}f$ in $\mathcal{I}_r$ can be equivalently found by minimizing the regularized risk functional:

$$M_\lambda f \triangleq R_{\text{emp}}f + \lambda\|f\|^2_{\mathcal{H}} \tag{1}$$

where λ is related to r via a decreasing global homeomorphism for which there exists a precise closed form [8]. In turn the unique minimizer $\hat{f}_\lambda$ of (1) can

be easily computed. Before continuing let's recall some aspects of reproducing kernel Hilbert spaces. For the general theory on this subject we make reference to the literature [9],[10],[8],[11]. In a nutshell reproducing kernel Hilbert spaces are spaces of functions[1], in the following denoted by $\mathcal{H}$, in which for each $\boldsymbol{x} \in \mathcal{X}$ the evaluation functional $L_{\boldsymbol{x}} : \ f \mapsto f(\boldsymbol{x})$ is a bounded linear functional. In this case the Riesz theorem guarantees that there exists $k_{\boldsymbol{x}} \in \mathcal{H}$ which is the representer of the evaluation functional $L_{\boldsymbol{x}}$ i.e. it satisfies $L_{\boldsymbol{x}} f = \langle f, k_{\boldsymbol{x}} \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the inner-product in $\mathcal{H}$. We then call *reproducing kernel* (r.k.) the symmetric bivariate function $k(\boldsymbol{x}, \boldsymbol{y})$ which for fixed $\boldsymbol{y}$ is the representer of $L_{\boldsymbol{y}}$. We state explicitly a well-known result which can be found in one of its multiple forms e.g. in [8].

Theorem 1. *Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. For a given sample $\mathcal{Z}^N$ the regularized risk functional* (1) *admits over $\mathcal{H}$ a unique minimizer $\hat{f}_\lambda$ that can be expressed as*

$$\hat{f}_\lambda = \sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i k_{\boldsymbol{x}_i} \tag{2}$$

where $\hat{\boldsymbol{\alpha}}$ is the unique solution of the well-posed linear system:

$$(\lambda N I + K)\boldsymbol{\alpha} = \boldsymbol{y}, \tag{3}$$

$[K]_{ij} = \langle k_{\boldsymbol{x}_i}, k_{\boldsymbol{x}_j} \rangle$ and $\langle \cdot, \cdot \rangle$ refers to the inner-product defined in $\mathcal{H}$.

Corollary 1. *The function* (2) *minimizes* (1) *if and only if $\hat{\boldsymbol{\alpha}}$ is the solution of*

$$\max_{\boldsymbol{\alpha}} \ 2\boldsymbol{\alpha}^\top \boldsymbol{y} - \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} - \lambda N \boldsymbol{\alpha}^\top \boldsymbol{\alpha}. \tag{4}$$

Notice that minimizing (1) is an optimization problem in a possibly infinite dimensional Hilbert Space. The crucial aspect of theorem 1 is that it actually permits to compute the solution $\hat{f}$ by resorting to standard methods in finite dimensional Euclidean spaces. We now turn into the core of structured hypothesis space which represents the theoretical basis for our modeling with embedded selection approach.

3 Structured Hypothesis Space

3.1 General Framework

Assume $\{\mathcal{F}^{(i)}\}_{i=1}^{d}$ are orthogonal subspaces of a RKHS $\mathcal{F}$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Denoted with $P^{(i)}$ the projection operator mapping $\mathcal{F}$ onto $\mathcal{F}^{(i)}$, their r.k. is given by $k^{(i)} = P^{(i)} k$ [2]. It can be shown [11],[12],[13] that

[1] In the following we always deal with real functions.

[2] Projection operators here generalize the notion of the ordinary projection operators in Euclidean spaces. By Pk we mean that P is applied to the bivariate function k as a function of one argument keeping the other one fixed. More precisely if $k_{\boldsymbol{x}}$ is the representer of $L_{\boldsymbol{x}}$ then $Pk_{\boldsymbol{x}} = Pk(\cdot, \boldsymbol{x})$.

the space $\mathcal{H} = \mathcal{F}^{(1)} \oplus \mathcal{F}^{(2)} \oplus \ldots \mathcal{F}^{(d)}$ of functions $\{f = f^{(1)} + \ldots + f^{(d)},\ f^{(i)} \in \mathcal{F}^{(i)}\}$ equipped with the inner product:

$$\langle f, g \rangle_{\mathcal{H},\boldsymbol{\gamma}} = [\boldsymbol{\gamma}]_i^{-1} \langle f^{(1)}, g^{(1)} \rangle_{\mathcal{F}} + \ldots + [\boldsymbol{\gamma}]_d^{-1} \langle f^{(d)}, g^{(d)} \rangle_{\mathcal{F}},\ \boldsymbol{\gamma} \in \mathbb{R}^d : \boldsymbol{\gamma} \succ \mathbf{0} \quad (5)$$

is a RKHS with reproducing kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} [\boldsymbol{\gamma}]_i k^{(i)}(\boldsymbol{x}, \boldsymbol{y})$. Indeed for each $\boldsymbol{x}$ the evaluation functional is bounded and by applying the definition of r.k., which is based on the inner product and the Riesz theorem, it is easy to check that $k_{\boldsymbol{x}}$ as above is its representer.

Notice that, even if the inner-product (5) requires $\boldsymbol{\gamma} \succ \mathbf{0}$, for any choice of $\boldsymbol{\gamma} \succeq \mathbf{0}$ in the kernel expansion k there always exists a corresponding space such that if $[\boldsymbol{\gamma}]_j = 0, \forall\ j \in \mathcal{I}$ then $k = \sum_{j \notin \mathcal{I}} [\boldsymbol{\gamma}]_j k^{(j)}$ is the reproducing kernel of $\mathcal{H} = \oplus_{j \notin \mathcal{I}} \mathcal{F}^{(j)}$ equipped with the inner-product $\langle f, g \rangle_{\mathcal{H},\boldsymbol{\gamma}} = \sum_{j \notin \mathcal{I}} [\boldsymbol{\gamma}]_j^{-1} \langle f^{(j)}, g^{(j)} \rangle_{\mathcal{F}}$.

Let $\hat{f}_\lambda = \sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i k_{\boldsymbol{x}_i}$ be the solution of (1) in $\mathcal{H}$. Notice that its projection in $\mathcal{F}^{(j)}$ is $P^{(j)} \hat{f}_\lambda = P^{(j)} (\sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i k_{\boldsymbol{x}_i}) = \sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i P^{(j)} \left(\sum_{l=1}^{d} [\boldsymbol{\gamma}]_l k_{\boldsymbol{x}_i}^{(l)} \right) = [\boldsymbol{\gamma}]_j \sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i k_{\boldsymbol{x}_i}^{(j)}$ and for the norm induced by (5) we easily get $\|\hat{f}_\lambda\|_{\mathcal{H},\boldsymbol{\gamma}}^2 = \sum_{j=1}^{d} [\boldsymbol{\gamma}]_j^{-1} \|P^{(j)} \hat{f}_\lambda\|_{\mathcal{F}}^2$ where $\|P^{(j)} \hat{f}_\lambda\|_{\mathcal{F}}^2 = [\boldsymbol{\gamma}]_j^2 \hat{\boldsymbol{\alpha}}^\top K^{(j)} \hat{\boldsymbol{\alpha}}$, $[K^{(j)}]_{lk} = k^{(j)}(\boldsymbol{x}_l, \boldsymbol{x}_k)$.

Besides the tensor product formalism of ANOVA models, which we present in Subsection 3.3, the present ideas naturally apply to the class of linear functions.

3.2 Space of Linear Functions

Consider the Euclidean space $\mathcal{X} = \mathbb{R}^d$. The space of linear functions $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ forms its dual space $\mathcal{X}^\star$. Since $\mathcal{X}$ is finite dimensional with dimension d, also its dual has dimension d and the dual base is simply given by[3]:

$$e_i(\boldsymbol{e}_j) = \begin{cases} 1\ , j = i \\ 0\ , j \neq i \end{cases}. \quad (6)$$

where $\{\boldsymbol{e}_j\}_{j=1}^{d}$ denotes the canonical basis of $\mathbb{R}^d$. We can further turn $\mathcal{X}^\star$ into an Hilbert space by posing:

$$\langle e_j, e_j \rangle_{\mathcal{X}^\star} = \begin{cases} 1\ , j = i \\ 0\ , j \neq i \end{cases} \quad (7)$$

which defines the inner-product for any couple $(f, g) \in \mathcal{X}^\star \times \mathcal{X}^\star$. Since $\mathcal{X}^\star$ is finite dimensional, it is a RKHS with r.k. [4]: $k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} e_i(\boldsymbol{x}) e_i(\boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{y}$ where for the last equality we exploited the linearity of e_i and the representation of vectors of $\mathbb{R}^d$ in terms of the canonical basis.

[3] Notice that, as an element of $\mathcal{X}^\star$, e_i is a linear function and (6) defines the value of $e_i(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. Since any element of $\mathcal{X}^\star$ can be represented as a linear combination of the basis vectors e_j, (6) also defines the value of $f(\boldsymbol{x})$ for all $f \in \mathcal{X}^\star,\ \boldsymbol{x} \in \mathcal{X}$.

[4] Indeed, any finite dimensional Hilbert space of functions $\mathcal{H}$ is a RKHS (see e.g. [10]).

Consider now the spaces $\mathcal{X}_j^\star$ spanned by e_j for $j = 1, \ldots, d$ and denote with $P^{(j)}$ the projection operator mapping $\mathcal{X}^\star$ into $\mathcal{X}_j^\star$. For their r.k. we simply have: $k^{(j)}(\boldsymbol{x}, \boldsymbol{y}) = P^{(j)} k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top A^{(j)} \boldsymbol{y}$ where $A^{(j)} = \boldsymbol{e}_j \boldsymbol{e}_j^\top$. Finally, redefining the standard dot-product according to (5), it is immediate to get $k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top A \boldsymbol{y}$ where $A = \mathrm{diag}(\boldsymbol{\gamma})$.

3.3 Functional ANOVA Models

ANOVA models represent an attempt to overcome the curse of dimensionality in non-parametric multivariate function estimation. Before introducing them we recall the concept of tensor product of RKHSs.

We refer the reader to the literature for the basic definitions and properties about tensor product Hilbert spaces [14],[15],[16],[17]. Their importance is mainly due to the fact that any reasonable multivariate function can be represented as a tensor product [15]. We just recall the following. Let $\mathcal{H}_1, \mathcal{H}_2$ be RKHSs defined respectively on $\mathcal{X}_1$ and $\mathcal{X}_2$ and having r.k. k_1 and k_2. Then the tensor product $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ equipped with the inner-product $\langle f_1 \otimes f_1', f_2 \otimes f_2', \rangle = \langle f_1, f_1' \rangle_{\mathcal{H}_1} \langle f_2, f_2' \rangle_{\mathcal{H}_2}$ is a reproducing kernel Hilbert space with reproducing kernel: $k_1 \otimes k_2 : ((\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2)) \to k_1(\boldsymbol{x}_1, \boldsymbol{y}_1) k_2(\boldsymbol{x}_2, \boldsymbol{y}_2)$ where $(\boldsymbol{x}_1, \boldsymbol{y}_1) \in \mathcal{X}_1 \times \mathcal{X}_1$, $(\boldsymbol{x}_2, \boldsymbol{y}_2) \in \mathcal{X}_2 \times \mathcal{X}_2$ [9]. Notice that the previous can be immediately extended to the case of tensor product of a finite number of spaces.

Let now $\mathcal{I}$ be an index set and $\mathcal{W}_j$, $j \in \mathcal{I}$ be reproducing kernel Hilbert spaces. Consider the tensor product $\mathcal{F} = \otimes_{j \in \mathcal{I}} \mathcal{W}_j$. An orthogonal decomposition in the factor spaces $\mathcal{W}_j$ induces an orthogonal decomposition of $\mathcal{F}$ [13],[12]. Consider e.g. the twofold tensor product $\mathcal{F} = \mathcal{W}_1 \otimes \mathcal{W}_2$. Assume now $\mathcal{W}_1 = \mathcal{W}_1^{(1)} \oplus \mathcal{W}_1^{(2)}$, $\mathcal{W}_2 = \mathcal{W}_2^{(1)} \oplus \mathcal{W}_2^{(2)}$. Then $\mathcal{F} = \left(\mathcal{W}_1^{(1)} \otimes \mathcal{W}_2^{(1)}\right) \oplus \left(\mathcal{W}_1^{(2)} \otimes \mathcal{W}_2^{(1)}\right) \oplus \left(\mathcal{W}_1^{(1)} \otimes \mathcal{W}_2^{(2)}\right) \oplus \left(\mathcal{W}_1^{(2)} \otimes \mathcal{W}_2^{(2)}\right)$ and each element $f \in \mathcal{F}$ admits a unique expansion:

$$f = f^{(11)} + f^{(21)} + f^{(12)} + f^{(22)}. \tag{8}$$

Each factor: $\mathcal{F}^{(jl)} \triangleq \mathcal{W}_1^{(j)} \otimes \mathcal{W}_2^{(l)}$ is a RKHS with r.k. $k^{(jl)} = P_1^{(j)} k_1 \otimes P_2^{(l)} k_2$ where $P_i^{(j)}$ denotes the projection of $\mathcal{W}_i$ onto $\mathcal{W}_i^{(j)}$. In the context of functional ANOVA models [13],[12] one considers the domain $\mathcal{X}$ as the Cartesian product of sets: $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$ where typically $\mathcal{X}_i \subset \mathcal{R}^{d_i}$, $d_i \geq 1$. In this context for $i = 1, \ldots, d$, $\mathcal{W}_i$ is a space of functions on $\mathcal{X}_i$, $\mathcal{W}_i^{(1)}$ is typically the subspace of constant functions and $\mathcal{W}_i^{(2)}$ is its orthogonal complement (i.e. the space of functions having zero projection on the constant function). In such a framework equation (8) is typically known as ANOVA decomposition.

Concrete constructions of these spaces are provided e.g. in [10] and [11]. Due to their orthogonal decomposition, they are a particular case of the general framework presented in the beginning of this section.

4 Subspace Selection via Quadratically Constrained Quadratic Programming

Consider an hypothesis space $\mathcal{H}$ defined as in the previous section and fix $\lambda > 0$ and $\boldsymbol{\gamma} \succeq \mathbf{0}$. It is immediate to see, in view of theorem 1, that the minimizer of the regularized risk functional $M_{\lambda,\boldsymbol{\gamma}} f = R_{\text{emp}} f + \lambda \sum_{j=1}^{d} [\boldsymbol{\gamma}]_j^{-1} \|P^{(j)} f\|_{\mathcal{F}}^2$ admits a representation as $\sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i \left(\sum_{j=1}^{d} [\boldsymbol{\gamma}]_j k_{\boldsymbol{x}_i}^{(j)} \right)$ where $\hat{\boldsymbol{\alpha}}$ solves the unconstrained optimization problem given by corollary 1. Adapted to the present setting (4) becomes: $\max_{\boldsymbol{\alpha}} 2\boldsymbol{\alpha}^\top \boldsymbol{y} - \sum_{j=1}^{d} [\boldsymbol{\gamma}]_j \boldsymbol{\alpha}^\top K^{(j)} \boldsymbol{\alpha} - \lambda N \boldsymbol{\alpha}^\top \boldsymbol{\alpha}$.

Consider the following optimization problem, consisting in an outer and an inner part:

$$\begin{bmatrix} \min\limits_{\boldsymbol{\gamma} \in \mathbb{R}^d} \max\limits_{\boldsymbol{\alpha}} 2\boldsymbol{\alpha}^\top \boldsymbol{y} - \sum_{j=1}^{d} [\boldsymbol{\gamma}]_j \boldsymbol{\alpha}^\top K^{(j)} \boldsymbol{\alpha} - \lambda N \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \\ \text{s.t.} \qquad \mathbf{1}^\top \boldsymbol{\gamma} = p \\ \boldsymbol{\gamma} \succeq \mathbf{0} \end{bmatrix}. \tag{9}$$

The idea behind it is that of optimizing simultaneously both $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Correspondingly, as we will show, we get at the same time the selection of subspaces and the optimal coefficients $\hat{\boldsymbol{\alpha}}$ of the minimizer (2) in the actual structure of the hypothesis space. In the meantime notice that, if we pose $\hat{f} = \sum_{i=1}^{N} [\hat{\boldsymbol{\alpha}}]_i \left(\sum_{j=1}^{d} [\hat{\boldsymbol{\gamma}}]_j k_{\boldsymbol{x}_i}^{(j)} \right)$ where $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\alpha}}$ solve (9), $\hat{f}$ is clearly the minimizer of $M_{\lambda,\hat{\boldsymbol{\gamma}}} f$ in $\mathcal{H} = \oplus_{j \in \mathcal{I}^+} \mathcal{F}^{(j)}$ where $\mathcal{I}^+$ is the set corresponding to non-zero $[\hat{\boldsymbol{\gamma}}]_j$.

The parameter p in the linear constraint controls the total sum of parameters $[\boldsymbol{\gamma}]_j$. While we adapt the class of hypothesis spaces according to the given sample, the capacity, which critically depends on the training set, is controlled by the regularization parameter λ selected outside the optimization. Problem (9) was inspired by [4] where similar formulations were devised in the context of transduction i.e. for the task of completing the labeling of a partially labelled dataset. By dealing with the more general search for functional dependency, our approach starts from the minimization of a risk functional in a Hilbert space. In this sense problem (9) can also be seen as a convex relaxation to the model selection problem arising from the design of the hypothesis space. See [13],[11] for an account on different non-linear non-convex heuristics for the search of the best $\boldsymbol{\gamma}$.

The solution of (9) can be carried out solving a QCQP problem. Indeed we have the following result that, due to space limitations, we state without proof.

Proposition 1. *Let p and λ be positive parameters and denote with $\hat{\boldsymbol{\alpha}}$ and $\hat{\nu}$ the solution of problem:*

$$\begin{bmatrix} \min\limits_{\boldsymbol{\alpha} \in \mathbb{R}^N,\ \nu \in \mathbb{R},\ \beta \in \mathbb{R}} \lambda N \beta - 2\boldsymbol{\alpha}^\top \boldsymbol{y} + \nu p \\ \boldsymbol{\alpha}^\top I \boldsymbol{\alpha} \quad \leq \beta \\ \text{s.t.} \\ \boldsymbol{\alpha}^\top K^{(j)} \boldsymbol{\alpha} \leq \nu \ j = 1, \ldots, d \end{bmatrix}. \tag{10}$$

Define the sets: $\hat{\mathcal{I}}^+ = \{i \ : \hat{\boldsymbol{\alpha}}^\top K^{(i)} \hat{\boldsymbol{\alpha}} = \hat{\nu}\}$, $\hat{\mathcal{I}}^- = \{i : \hat{\boldsymbol{\alpha}}^\top K^{(i)} \hat{\boldsymbol{\alpha}} < \hat{\nu}\}$ *and let* h *be a bijective mapping between index sets:* $h : \hat{\mathcal{I}}^+ \to \{1, \ldots, |\hat{\mathcal{I}}^+|\}$. *Then:*

$$[\hat{\boldsymbol{\gamma}}]_i = \begin{cases} [\boldsymbol{b}]_{h(i)} \ , i \in \hat{\mathcal{I}}^+ \\ 0 \quad\quad , i \in \hat{\mathcal{I}}^- \end{cases} , \ [\boldsymbol{b}]_{h(i)} > 0 \ \forall \ i \in \hat{\mathcal{I}}^+, \ \sum_{i \in \hat{\mathcal{I}}^+} [\boldsymbol{b}]_{h(i)} = p \qquad (11)$$

and $\hat{\boldsymbol{\alpha}}$ *are the solutions of* (9).

Basically, once problem (9) is recognized to be convex [4], it can be converted into (10) which is a QCQP problem [6]. In turn the latter can be efficiently solved e.g. with CVX [18] and the original variables $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ can be computed. Notice that, while one gets the unique solution $\hat{\boldsymbol{\alpha}}$, the non-zero $[\hat{\boldsymbol{\gamma}}]_{i \in \hat{\mathcal{I}}^+}$ are just constrained to be strictly positive and to sum up to p. This seems to be an unavoidable side effect of the selection mechanism. Any choice (11) is valid since for each of them the optimal value of the objective function in (9) is attained. Despite of that one solution among the set (11) corresponds to the optimal value of the dual variables associated to the constraints $\boldsymbol{\alpha}^\top K^{(j)} \boldsymbol{\alpha} \leq \nu \ j = 1, \ldots, d$ in (10). Since interior point methods solve both the primal and dual problem, by solving (10) one gets at no extra price an optimal $\hat{\boldsymbol{\gamma}}$.

5 Relation with LASSO and Hard Thresholding

Concerning the result presented in proposition 1 we highlighted that the solution $\hat{f}$ has non-zero projection only on the subspaces whose associated quadratic constraints are active i.e. for which the boundary $\hat{\nu}$ is attained. This is why in the following we refer to (10) as *Quadratically Constraint Quadratic Programming Problem for Subspace Selection* (QCQPSS). Interestingly this mechanism of selection is related to other thresholding approaches that were studied in different contexts. Among them the most popular is LASSO, an estimate procedure for linear regression that shrinks some coefficients to zero [2]. If we pose $X = \begin{bmatrix} \boldsymbol{x}_1 \ldots \boldsymbol{x}_N \end{bmatrix}$[5] the LASSO estimate $[\hat{\boldsymbol{w}}]_j$ of the coefficients corresponds to the solution of problem:

$$\begin{bmatrix} \min\limits_{\boldsymbol{w} \in \mathbb{R}^d} \|\boldsymbol{y} - X^\top \boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad\quad \|\boldsymbol{w}\|_1 \leq t \end{bmatrix}.$$

It is known [2] that, when $XX^\top = I$, the solution of LASSO corresponds to $[\hat{\boldsymbol{w}}]_j = \text{sign}([\tilde{\boldsymbol{w}}]_j) \left(|[\tilde{\boldsymbol{w}}]_j| - \xi_t\right)^+$ where $\tilde{\boldsymbol{w}}$ is the least squares estimate $\tilde{\boldsymbol{w}} \triangleq \arg\min_{\boldsymbol{w}} \|\boldsymbol{y} - X^\top \boldsymbol{w}\|_2^2$ and ξ_t depends upon t. Thus in the orthogonal case LASSO selects (and shrinks) the coefficients corresponding to the LS estimate that are bigger than the threshold. This realizes the so called *soft thresholding* which represents an alternative to the hard threshold estimate (subset selection)

[5] It is typically assumed that the observations are normalized i.e. $\sum_i [\boldsymbol{x}_i]_j = 0$ and $\sum_i [\boldsymbol{x}_i]_j^2 = n \ \forall \ j$.

[19]: $[\hat{\boldsymbol{w}}]_j = [\tilde{\boldsymbol{w}}]_j \; I(|[\tilde{\boldsymbol{w}}]_j| > \gamma)$ where γ is a number and $I(\cdot)$ denotes here the indicator function.

When dealing with linear functions as in Subsection 3.2 it is not too hard to carry out explicitly the solution of (9) when $XX^\top = I$. We were able to demonstrate the following. Again we do not present here the proof for space limitations.

Proposition 2. *Assume* $XX^\top = I$, $k^{(j)}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top A^{(j)} \boldsymbol{y}$, $A^{(j)} = \boldsymbol{e}_j \boldsymbol{e}_j^\top$ *and let* $[K^{(j)}]_{lm} = k^{(j)}(\boldsymbol{x}_l, \boldsymbol{x}_m)$. *Let* $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\alpha}}$ *be the solution of* (9) *for fixed values of* p *and* λ. *Then the function* $\hat{f}_{\lambda,p} = \sum_{i=1}^N [\hat{\boldsymbol{\alpha}}]_i \left(\sum_{j=1}^d [\hat{\boldsymbol{\gamma}}]_j k_{\boldsymbol{x}_i}^{(j)} \right)$ *can be equivalently restated as:*

$$\hat{f}_{\lambda,p}(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \boldsymbol{x}, \; [\hat{\boldsymbol{w}}]_j = I([\tilde{\boldsymbol{w}}]_j > \xi) \frac{[\tilde{\boldsymbol{w}}]_j}{1 + \frac{N\lambda}{[\hat{\boldsymbol{\gamma}}]_j}}$$

where $\xi = \lambda N \sqrt{\hat{\nu}}$, $\hat{\nu}$ *being the solution of* (10).

In comparison with l_1-norm selection approaches, when dealing with linear functions, problem (9) provides a principled way to combine sparsity and l_2 penalty. Notice that the coefficients associated to the set of indices $\hat{\mathcal{I}}^+$ are a shrinked version of the corresponding LS estimate and that the amount of shrinkage depends upon the parameters λ and $\hat{\boldsymbol{\gamma}}$.

6 Experimental Results

We illustrate here the method for the class of linear functions.

In near-infrared spectroscopy absorbances are measured at a large number of evenly-spaced wavelengths. The publicly available Tecator data set (`http://lib.stat.cmu.edu/datasets/tecator`) donated by the Danish Meat Research Institute, contains the logarithms of the absorbances at 100 wavelengths of finely chopped meat recorded using a Tecator Infrared Food and Feed Analyzer. It consists of sets for training, validation and testing. The logarithms of absorbances are used as predictor variables for the task of predicting the percentage of fat. The result on the same test set of QCQPSS as well as the LASSO, were compared with the results obtained by other subset selection algorithms[6] as reported (with discussion) in [20]. For both QCQPSS and the LASSO the data were normalized. The values of p and λ in (10) were taken from a predefined grid. For each value of $|\hat{\mathcal{I}}_{p,\lambda}^+|$ as a function of (p, λ), a couple $(\hat{p}, \hat{\lambda})$ was selected if $\hat{f}_{\hat{p},\hat{\lambda}}$ achieved the smallest residual sum of squares (RSS) computed on the base of the validation set. A corresponding criterion was used in the LASSO for the choice of $\hat{t}$ associated to any cardinality of the selected subsets. QCQPSS selected at most 9 regressors while by tuning the parameter in the LASSO one can select subsets of variables of an arbitrary cardinality. However very good results were achieved for models depending on small subsets of variables, as reported on Table 1.

[6] In the cases after the selection, the parameters were fitted according to least squares.

Table 1. Best RSS on test dataset for each cardinality of subset of variables. QCQPSS achieves very good results for models depending on small subsets of variables. The results for the first 5 methods are taken from [20].

Num. of vars.	Forward sel.	Backward elim.	Sequential replac.	Sequential 2-at-a-time	Random 2-at-a-time	QCQPSS	LASSO
1	14067.6	14311.3	14067.6	14067.6	14067.6	1203100	6017.1
2	2982.9	3835.5	2228.2	2228.2	2228.2	**1533.8**	1999.5
3	1402.6	1195.1	1191.0	1156.3	1156.3	**431.0**	615.8
4	1145.8	1156.9	833.6	799.7	799.7	**366.1**	465.6
5	1022.3	1047.5	711.1	610.5	610.5	402.3	391.4
6	913.1	910.3	614.3	475.3	475.3	390.5	388.7
7	–	–	–	–	–	358.9	356.1
8	852.9	742.7	436.3	417.3	406.2	376.0	351.7
9	–	–	–	–	–	369.4	345.3
10	746.4	553.8	348.1	348.1	340.1	–	344.3
12	595.1	462.6	314.2	314.2	295.3	–	322.9
15	531.6	389.3	272.6	253.3	252.8	–	331.5

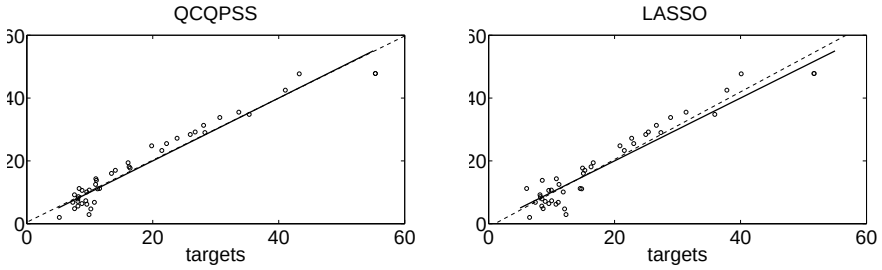

Fig. 1. Targets vs. outputs (test dataset) for the best-fitting 3 wavelengths models. The best linear fit is indicated by a dashed line. The perfect fit (output equal to targets) is indicated by the solid line. The correlation coefficients are respectively .97 (QCQPSS - left panel) and .959 (LASSO - right panel).

7 Conclusions

We have presented an abstract class of structured reproducing kernel Hilbert spaces which represents a broad set of models for multivariate function estimation. Within this framework we have elaborated on a convex approach for selecting relevant subspaces forming the structure of the approximating space. Subsequently we have focused on the space of linear functions, in order to gain a better insight on the selection mechanism and to highlight the relation with LASSO.

Acknowledgment This work was sponsored by the Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES-/4CHEM, several PhD/postdoc and fellow grants; Flemish Government: FWO: PhD

postdoc grants, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.03-/02.07, G.0320.08, G.0558.08, G.0557.08, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite+; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011) ; EU: ERNSI; Contract Research: AMINAL.

References

1. Pelckmans, K., Goethals, I., De Brabanter, J., Suykens, J., De Moor, B.: Componentwise Least Squares Support Vector Machines. In: Support Vector Machines: Theory and Applications (Wang L., ed.). Springer (2005) 77–98
2. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) **58**(1) (1996) 267–288
3. Chen, S.: Basis Pursuit. PhD thesis, Department of Statistics, Stanford University (November 1995)
4. Lanckriet, G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.: Learning the Kernel Matrix with Semidefinite Programming. The Journal of Machine Learning Research **5** (2004) 27–72
5. Tsang, I.: Efficient hyperkernel learning using second-order cone programming. IEEE transactions on neural networks **17**(1) (2006) 48–58
6. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. Society for Industrial Mathematics (2001)
7. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
8. Cucker, F., Zhou, D.: Learning Theory: An Approximation Theory Viewpoint (Cambridge Monographs on Applied & Computational Mathematics). Cambridge University Press New York, NY, USA (2007)
9. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society **68** (1950) 337 – 404
10. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers (2004)
11. Wahba, G.: Spline Models for Observational Data. Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia (1990)
12. Gu, C.: Smoothing spline ANOVA models. Series in Statistics (2002)
13. Chen, Z.: Fitting Multivariate Regression Functions by Interaction Spline Models. J. of the Royal Statistical Society. Series B (Methodological) **55**(2) (1993) 473–491
14. Light, W., Cheney, E.: Approximation theory in tensor product spaces. Lecture Notes in Math **1169** (1985)
15. Takemura, A.: Tensor Analysis of ANOVA Decomposition. Journal of the American Statistical Association **78**(384) (1983) 894–900
16. Huang, J.: Functional ANOVA models for generalized regression. J. Multivariate Analysis **67** (1998) 49–71
17. Lin, Y.: Tensor Product Space ANOVA Models. The Annals of Statistics **28**(3) (2000) 734–755
18. Grant, M., Boyd, S., Ye, Y.: CVX: Matlab Software for Disciplined Convex Programming (2006)
19. Donoho, D., Johnstone, J.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3) (2003) 425–455
20. Miller, A.: Subset Selection in Regression. CRC Press (2002)