Data-Dependent Norm Adaptation for Sparse Recovery in Kernel Ensembles Learning

Marco Signoretto MARCO.SIGNORETTO@ESAT.KULEUVEN.BE
ESAT-SCD/SISTA
Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, B-3001 Leuven (BELGIUM)

Kristiaan Pelckmans KP@IT.UU.SE
Division of Systems and Control
Department of Information Technology, Uppsala University
Box 337 SE-751, 05 Uppsala (SWEDEN)

Lieven De Lathauwer LIEVEN.DELATHAUWER@KULEUVEN-KORTRIJK.BE
Group Science, Engineering and Technology
Katholieke Universiteit Leuven, Campus Kortrijk
E. Sabbelaan 53, 8500 Kortrijk (BELGIUM)

Johan A.K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
ESAT-SCD/SISTA
Katholieke Universiteit Leuven,
Kasteelpark Arenberg 10, B-3001 Leuven (BELGIUM)

Editor:

Abstract

We study the problem of learning sparse nonlinear models given observations drawn from an unknown distribution. This problem can be recasted into the Multiple Kernel Learning (MKL) framework. It has also been studied in the context of functional ANOVA models, under the name of COmponent Selection and Smoothing Operator (COSSO). Our primary focus is on detecting the structure behind the data. We give an oracle-type inequality that relates sparse recovery with a measure of in-sample dependence. This result motivates a new type of data-dependent regularization that adapts to the observed dependence structure. The idea is then translated into an algorithm based on COSSO. Experimental results are provided to show the effectiveness of the proposed approach.

1. Introduction

In real-life problems there is often need for unrevealing the structure behind the data. Functional ANOVA models have emerged as a class of structured models useful to capture nonlinear relations, while still providing insight in the model and dealing appropriately with the curse of dimensionality. The general principle is to approximate the underlying functional relation by an additive expansion, where components are functions on subsets of variables. This setting has been studied especially in nonparametric regression by smoothing splines (Wahba (1990); Gu (2002)). A special case is represented by additive models as defined in Buja et al. (1989), Hastie and Tibshirani (1990), which provide an extension of the

standard linear regression model while retaining the non interaction assumption. In linear modeling, a given target variable is modelled via a linear function of the input covariates and l_1 regularization is often used for input selection. In this context, it is known that the asymptotic probability of selecting the correct sparsity structure depends upon conditions on the covariance structure of the input data (as shown, e.g., in Zhao and Yu (2006)). In daily practice these results are confirmed by finite sample studies. It is shown that the empirical probability of correct sparse recovery decreases with the degree of collinearity in the data. One popular workaround is the use of suitable preprocessing techniques, like principal component analysis. However, a transformation of the original coordinate system has often the drawback of hindering the interpretation of the successively fitted models. This is the case, in particular, when data have different physical interpretations, such as lengths and weights. When nonlinear transforms of the covariates are taken, as in additive type of models, the problem of collinearity between the original variables often translates into the collinearity of the corresponding transforms. This phenomenon is traditionally referred to in the literature as *concurvity*, and represents the nonparametric analog of collinearity (Buja et al. (1989); Hastie and Tibshirani (1990); Gu (1992)). Our contribution in this paper is to characterize sparse recovery in presence of concurvity and to propose a new data-dependent approach to improve the detection of the underlying sparsity.

The structure of the paper is as follows. We begin by presenting our notation and some preliminaries (Section 2). In Section 3 we introduce the formulation for nonparametric sparse recovery and discuss the MKL framework. In Section 4 we give an oracle result relating concurvity and sparse recovery. We then propose to replace the original block-wise penalties used in MKL, with data-dependent norms adapted to the in-sample dependence (Section 5). Successively, we elaborate on an algorithm which exploits the adaptation to learn a sparse nonparametric model (Section 6). We then present in Section 7 experimental results on both simulation studies and real life examples. Finally, in Section 8 we draw our conclusions.

2. Notation and preliminaries

For any $p \in \mathbb{N}$ we use the convention of denoting the set $\{1, \ldots, p\}$ by $\mathbb{N}_p$. We indicate an arbitrary vector $(x_j : j \in \mathbb{N}_d) \in \mathbb{R}^d$ by x and we write e.g. x_j^i to mean the $j-$th component of $x^i \in \mathbb{R}^d$. When needed the notation will be extended to arbitrary sequences, by replacing $\mathbb{N}_d$ with some abstract index set $\mathcal{I}$. The Euclidean norm of $\mathbb{R}^d$ is denoted by $\|\cdot\|_{\mathbb{R}^d}$. We assume the input space $\mathcal{X}$ is a compact domain in $\mathbb{R}^d$ and let the output space $\mathcal{Y}$ be a closed subset of $\mathbb{R}$. We further denote by ρ the underlying probability measure on $\mathcal{X} \times \mathcal{Y}$ and by ξ the corresponding marginal probability measure on $\mathcal{X}$. These measures are both unknown but a finite set of input-output pairs $\mathcal{D}_n := \left\{\left(x^i, y_i\right) : i \in \mathbb{N}_n\right\} \subset \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. according to ρ is available. Associated to the latter we have the empirical measure $\rho_n = \frac{1}{n}\sum_{i\in\mathbb{N}_n} \delta_{(x^i,y_i)}$ where $\delta_{(x^i,y_i)}$ denotes the Dirac measure at $(x^i, y_i) \in \mathcal{D}_n$. Let $l : \mathbb{R} \times \mathcal{Y} \to [0, +\infty]$ be a convex loss function. For a function $h : \mathcal{X} \to \mathbb{R}$ the empirical risk is defined upon $\mathcal{D}_n$ as

$$C_n(h) = \int_{\mathcal{X}\times\mathcal{Y}} l\left(h\left(x\right), y\right) d\rho_n(x, y) = \frac{1}{n} \sum_{(x^i,y_i)\in\mathcal{D}_n} l\left(h\left(x^i\right), y_i\right) \ .$$

It is the discrete version of its theoretical counterpart, the expected risk measured upon the underlying probability:

$$C(h) = \int_{\mathcal{X}\times\mathcal{Y}} l\left(h(x), y\right) d\rho(x, y) \ . \tag{1}$$

Let $\mathcal{H}$ be a space of real-valued functions defined on $\mathcal{X}$. In the following we often write $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ to stress that $\mathcal{H}$ is endowed with the Hilbert space structure defined according to the inner product $\langle\cdot,\cdot\rangle$. In particular, a significant role will be played by the space of square integrable real-valued functions $L_2(\mathcal{X}, \xi)$ (often simply denoted by L_2) endowed with the inner product

$$\langle f, g\rangle_{L_2(\mathcal{X},\xi)} = \int_{\mathcal{X}} f(x)g(x)d\xi(x)$$

and having associated norm $\|f\|_{L_2(\mathcal{X},\xi)} = \sqrt{\langle f, f\rangle_{L_2(\mathcal{X},\xi)}}$. Let now $\left\{\left(\mathcal{H}^j, \langle\cdot,\cdot\rangle_j\right) \ : \ j \in \mathbb{N}_m\right\}$ be a set of reproducing kernel Hilbert spaces of functions defined on $\mathcal{X}$ (Aronszajn (1950); Wahba (1990)). For any $j \in \mathbb{N}_m$ denote the reproducing kernel (r.k.) of $\mathcal{H}^j$ by $k^j : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. In the following we shall assume that for any $j \in \mathbb{N}_m$, $\mathcal{H}^j$ is a subspace of $L_2(\mathcal{X}, \xi)$ with continuos inclusion. This is the case whenever $\int_{\mathcal{X}} k^j(x, x)d\xi(x) < +\infty$. Since $\mathcal{X}$ is compact this holds true as soon as k^j is continuous. In this case for any $h^j \in \mathcal{H}^j$ we have, by application of the Cauchy-Schwarz inequality

$$\|h^j\|^2_{L_2(\mathcal{X},\xi)} = \int_{\mathcal{X}} |h^j(x)|^2 d\xi(x) = \int_{\mathcal{X}} \left|\left\langle h^j, k^j(x,\cdot)\right\rangle_j\right|^2 d\xi(x) \leq \|h^j\|_j^2 \int_{\mathcal{X}} k^j(x, x)d\xi(x) < +\infty \, .$$

For an index set $\mathcal{A} \subseteq \mathbb{N}_m$ let $\mathcal{H}_{\mathcal{A}} := \oplus_{j\in\mathcal{A}}\mathcal{H}^j$ be the space of (possibly non uniquely defined) additive models $h = \sum_{j\in\mathcal{A}} h^j$. Further denote simply by $\mathcal{H}$ the *model space* $\mathcal{H} = \oplus_{j\in\mathbb{N}_m}\mathcal{H}^j$. The latter space can be endowed with the inner product

$$\langle h, g\rangle = \inf\left\{\sum_{j\in\mathbb{N}_m} \langle h^j, g^j\rangle_j \ : \ h = \sum_{j\in\mathbb{N}_m} h^j; \ g = \sum_{j\in\mathbb{N}_m} g^j; \ h^j, g^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m\right\} \tag{2}$$

having associated norm $\|f\| = \sqrt{\langle f, f\rangle}$. It follows from the arguments in Aronszajn (1950) that, when endowed with such an inner product, $\mathcal{H}$ admits r.k.

$$k(x, y) = \sum_{j\in\mathbb{N}_m} k^j(x, y) \ . \tag{3}$$

In the following we often write k_x to mean $k(x,\cdot)$. We say that an arbitrary function $h \in \mathcal{H}$ is *known via its additive representation* if a specific decomposition $h = \sum_{j\in\mathbb{N}_m} h^j$ is considered given. We denote by $\mathcal{A}_h$ its *sparsity pattern* i.e. we define

$$\mathcal{A}_h := \left\{j \in \mathbb{N}_m \ : \ h^j \neq 0\right\} \ . \tag{4}$$

Under mild conditions on the kernel function, the Mercer spectral decomposition of k (Riesz and Sz.-Nagy (1955)) is

$$k(x, y) = \sum_{i\in\mathbb{N}} \lambda_i\phi_i(x)\phi_i(y)$$

where $\{\phi_i \ : \ i \in \mathbb{N}\}$ is an orthonormal system of eigenfunctions of $L_2(\mathcal{X}, \xi)$ and $\{\lambda_i \ : \ i \in \mathbb{N}\}$ is a corresponding sequence of positive eigenvalues. Let $\mathcal{H}_D \subset \mathcal{H}$ be defined as

$$\mathcal{H}_D = \left\{ h \in \mathcal{H} \ : \ \sum_{i \in \mathbb{N}} \frac{\langle h, \phi_i \rangle_{L_2}^2}{\lambda_i^2} \leq +\infty \right\} .$$

We follow Koltchinskii and Yuan (2008) and introduce the operator $D : \mathcal{H}_D \to L_2$,

$$D(h) = \sum_{i \in \mathbb{N}} \frac{\langle h, \phi_i \rangle_{L_2}}{\lambda_i} \phi_i$$

which is easily seen to satisfy $\langle h, g \rangle = \langle D(h), g \rangle_{L_2}$. In the same fashion we define a set of operators for the component spaces. That is, for $j \in \mathbb{N}_m$ we define $\mathcal{H}_D^j \subset \mathcal{H}^j$ and $D^j : \mathcal{H}_D^j \to L_2$ with analog definition as for D.

The penalty functional we shall consider is the sum of the norms in the component spaces, namely

$$J(h) = \inf \left\{ \sum_{j \in \mathbb{N}_m} \|h^j\|_j \ : \ h = \sum_{j \in \mathbb{N}_m} h^j, \ h^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m \right\} . \tag{5}$$

3. The Multiple Kernel Learning Problem

Here we study for $\tau > 0$ the variational problem

$$\inf \left\{ \frac{1}{n} \sum_{(x^i, y_i) \in \mathcal{D}_n} l(h(x^i), y_i) + \tau \sum_{j \in \mathbb{N}_m} \|h^j\|_j \ : \ h = \sum_{j \in \mathbb{N}_m} h^j, \ h^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m \right\} \tag{6}$$

that we can compactly restate as:

$$\inf \left\{ C_n(h) + \tau J(h) \ : \ h = \sum_{j \in \mathbb{N}_m} h^j, \ h^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m \right\} . \tag{7}$$

This formulation trades off the empirical error, as measured by C_n, with the complexity penalty J encouraging sparsity of the solution with respect to the component functions. The functional extends the standard LASSO (Tibshirani (1996)) to a non-parametric setting. This problem has attracted a lot of attention and it has been investigated both from a functional viewpoint (Micchelli and Pontil (2005); Micchelli and Pontil (2007)) and from the point of view of optimization (Lanckriet et al. (2004b); Bach et al. (2004)). Although conceptually different the two perspectives converge when the selection is tailored to the data, since in both cases it boils down to learn a sparse combination of kernel matrices arising from the observed sample. The formulation (7) is rather general and has been considered to accomplish different tasks. Here we are especially interested in uncovering the structure underlying the input-output data. We shall assume that every kernel in the ensemble encodes a specific feature from the input domain. A particular case that fits into this general idea

is found in functional ANOVA models (Gu (2002)). In this case component functions refer either to main effects or to higher order interactions among subsets of variables.

Below, we focus on the property of the sparse recovery associated to an estimate

$$\hat{h}^{\tau} := \arg\inf \left\{ C_n(h) + \tau J(h) \ : \ h = \sum_{j \in \mathbb{N}_m} h^j, \ h^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m \right\} \tag{8}$$

which amounts to uncover the relevant features. When studying the performance of an estimator it is common in the statistical literature to compare it with some theoretical representation of the process under study. According to Candès (2006), "an *oracle inequality* relates the performance of a real estimator with that of an ideal estimator which relies on perfect information supplied by an oracle". In this spirit we call the probabilistic inequality we are about to formulate an oracle inequality, as we assume we are given by an oracle a function $f \in \mathcal{H}$ (the oracle function) having additive representation $\sum_{j \in \mathbb{N}_m} f^j, \ f^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m$ and sparsity pattern $\mathcal{A}_f$. There is no need to specify further the nature of the oracle as it is not required in our derivation. This is in line with Koltchinskii and Yuan (2008) to which the result we are about to present is related. We just point out that an interesting oracle to target is the best additive representation entailing a certain number of components, i.e. given an integer $0 < d < m$ the function

$$f^d = \arg\inf \left\{ C(h) \ : \ h = \sum_{j \in \mathcal{A}} h^j \in \mathcal{H}_{\mathcal{A}}, \ \mathcal{A} \subset \mathbb{N}_m, \ \mathrm{card}(A) = d \right\} \ .$$

Our first goal here is to relate the model selection consistency of $\hat{h}^{\tau}$ with respect to f in terms of the dependence structure among the transforms of random variables contained in the model space $\mathcal{H}$.

4. Oracle result

Before stating our oracle result we point out the traditional notion of *concurvity* and then recall some indices that can be used to quantitatively characterize it.

4.1 Concurvity in Function Spaces

Traditionally the term *concurvity* has been used to describe the (approximate) collinearity of nonlinear transforms of predictors. This denomination was first introduced in Buja et al. (1989) in the context of additive models. The name *prospective concurvity* was used in Gu (1992) in connection to the empirical distribution of the data. It alludes to the existence of a family of functions $\{h^j \in \mathcal{H}^j \ : \ j \in \mathbb{N}_m\}$ such that $\{(h^j(x^i) \ : \ i \in \mathbb{N}_n) \ : \ j \in \mathbb{N}_m\}$ is a set of (approximatively) collinear vectors. In this case exact concurvity corresponds to the fact that the decomposition is not uniquely defined at the input sites in the dataset $\mathcal{D}_n$. An analog concept at the population level is discussed in Buja et al. (1989). In the case of a simple additive model $h = \sum_{j \in \mathbb{N}_m} h^j$ where the $j-$th component function depends upon the $j-$th variable, approximate concurvity corresponds to have the covariates distributed in the proximity of a low dimensional manifold embedded in the domain $\mathcal{X}$. Finally *observed*

concurvity is used when a model in $\mathcal{H}$ has already been fitted to the data. It refers to the collinearity at the design sites of the components of the model (the so called *retrospective* linear model). In this case the collinearity indices of Stewart (1987) can be used as a diagnostic tool. In Gu (1992) this leads to perform model selection by simply dropping fitted components, after performing estimation via a functional with squared norm penalty. This strategy is to be contrasted with sparsity inducing estimators, which perform model fitting and selection at the same time.

4.2 Concurvity Indices and Oracle Result

We shall now introduce indices to measure the degree of concurvity. Let $\mathcal{A} \subseteq \mathbb{N}_m$, define $q = \text{card}(\mathcal{A})$ and denote by $\mathcal{A}^c$ its complement in $\mathbb{N}_m$. Consider a given dictionary of functions on $\mathcal{X}$, $\{h^j \ : \ j \in \mathcal{A}\}$. The $q \times q$ Gram matrix of $\{h^j \ : \ j \in \mathcal{A}\}$ with respect to the inner product of L_2 is

$$G\left(\{h^j \ : \ j \in \mathcal{A}\}\right) := \left(\left\langle h^i, h^j \right\rangle_{L_2} \ : \ (i,j) \in \mathcal{A} \times \mathcal{A}\right) \ .$$

Denote by $\lambda_{\min}(A)$ the minimum eigenvalue of a square matrix A. A index of central importance for a worst-case characterization of dependence is

$$\kappa(\mathcal{A}) := \inf\left\{\lambda_{\min}(G(\{h^j \ : \ j \in \mathcal{A}\})) \ : \ h^j \in \mathcal{H}^j, \|h^j\|_{L_2} = 1, \ j \in \mathbb{N}_m\right\} \ . \tag{9}$$

In words, $\kappa(\mathcal{A})$ is the infimum eigenvalue of all the possible Gram matrices of functions running into the predefined set of spaces.

Remark 4.1 *It holds that:*

$$\lambda_{\min}(G(\{h^j \ : \ j \in \mathcal{A}\})) = \inf\left\{\left\|\sum_{j\in\mathcal{A}} \alpha_j h^j\right\|_{L_2}^2 \ : \ \alpha \in \mathbb{R}^q, \ \|\alpha\|_{\mathbb{R}^q} = 1\right\}$$

as $\left\{\| \sum_{j\in\mathcal{A}} \alpha_j h^j \|_{L_2}^2 \ : \ \alpha \in \mathbb{R}^q, \ \| \alpha\|_{\mathbb{R}^q} = 1\right\}$ *corresponds to the Rayleigh quotient*

$$R(\alpha) = \frac{\langle \alpha, G(\{h^j \ : \ j \in \mathcal{A}\})\alpha\rangle_{\mathbb{R}^q}}{\langle \alpha, \alpha\rangle_{\mathbb{R}^q}} \ .$$

Another central index is the supremum of the cosine of the angle between functions running in $\mathcal{H}_{\mathcal{A}}$ and $\mathcal{H}_{\mathcal{A}^c}$ respectively:

$$\delta(\mathcal{A}) = \sup\left\{\frac{\langle h, g\rangle_{L_2}}{\|h\|_{L_2}\|g\|_{L_2}} \ : \ h \in \mathcal{H}_{\mathcal{A}}, \ g \in \mathcal{H}_{\mathcal{A}^c}, \ h \neq 0, \ g \neq 0\right\} \ . \tag{10}$$

We call this the *mutual coherence in the model space*, as it is similar to the ordinary notion of mutual coherence of a predefined finite dictionary of functions (Donoho and Huo (2001); Donoho and Elad (2003)). If the functions are zero mean it corresponds to a correlation coefficient. At last we pose for $h = \sum_{j\in\mathbb{N}_m} h^j \in \mathcal{H}$

$$\psi(\mathcal{A}_h) = \sum_{j\in\mathcal{A}_h} \frac{\|D^j(h^j)\|_{L_2}^2}{\|h^j\|_j^2} \tag{11}$$

which we call the *relative smoothness of* h, details are given below.

We now collect in the following definition the assumptions on the loss function that we need within this Section.

Definition 4.1 *A function* $l : \mathbb{R} \times \mathcal{Y} \to [0, +\infty]$ *such that*

1. *for all* $y \in \mathcal{Y}$ *the function* $l(\cdot, y)$ *is convex on* $\mathbb{R}$*;*
2. l *is measurable on* $\mathbb{R} \times \mathcal{Y}$*;*
3. *for all* $y \in \mathcal{Y}$ *the function* $l(\cdot, y)$ *is differentiable on* $\mathbb{R}$*;*
4. *there are* $b \in [0, +\infty]$ *and* $a : \mathcal{Y} \to [0, +\infty]$ *such that*
   (a) $l(u, y) \leq a(y) + b|u|^2 \ \forall\, (u, y) \in \mathbb{R} \times \mathcal{Y}$
   (b) $\int_{\mathcal{X}\times\mathcal{Y}} a(y) d\rho(x, y) \ < \ +\infty$*;*

is called a well-behaved loss function w.r.t. ρ*.*

The convexity and measurability requirements are standard working assumptions. The other conditions are needed within the proofs of the following statements to rely on known results from convex variational analysis. With the exception of differentiability, the requirements in our definition correspond to the ones in the definition of $2-$losses in De Vito et al. (2004). In the context of regression, a well-behaved loss is for example the quadratic loss $l(u, y) = \frac{1}{2}(y - u)^2$ as soon as $\int_{\mathcal{X}\times\mathcal{Y}} y^2 d\rho(x, y) < +\infty$. Under similar conditions, for $\mathcal{Y} = \{-1, 1\}$ *margin-based losses* (Steinwart and Christmann (2008)) like the quadratic loss $l(u, y) = \frac{1}{2}(1 - yu)^2$, truncated quadratic loss $l(u, y) = (\max\{0, 1 - yu\})^2$ and logistic loss $l(u, y) = \log(1 + \exp(-yu))$, are also well-behaved.

For a functional $F : \mathcal{H} \to \mathbb{R}$ we denote by $\partial F(u)$ its subdifferential at u. Loosely speaking, subdifferentials are good surrogates of gradients and to some extent the ordinary differential calculus can be extended into subdifferential calculus. The reader is referred to Appendix A for the necessary definitions and results on which our statements rely on. Before stating our main Theorem connecting $\hat{h}^\tau$ with an oracle f as above, we state two Propositions which are merely instrumental to its proof.

Proposition 4.2 *Let* $h, f \in \mathcal{H}$ *be arbitrary functions known via their additive representations. The following upper bound holds:*

$$\sum_{j \in \mathcal{A}_f} \left(\|f^j\|_j - \|h^j\|_j \right) \leq \sqrt{\frac{\psi(\mathcal{A}_f)}{\kappa(\mathcal{A}_f)\left(1 - \delta^2(\mathcal{A}_f)\right)}} \|f - h\|_{L_2}. \tag{12}$$

Proof See Appendix B. ■

Proposition 4.3 *Let* $l : \mathbb{R} \times \mathcal{Y} \to [0, +\infty]$ *be a well-behaved loss function and* $f \in \mathcal{H}$ *be an arbitrary function known via its additive representation. Further let* C_n *be defined upon an i.i.d. sample* $\mathcal{D}_n$*. Let*

$$v_f(x, y, z) := \partial_u l(u, y)|_{u=f(x)} \, k(x, z)$$

and define

$$M_f := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \|D(v_f(x,y,\cdot))\|_{L_2} \ .$$

For $\nu > 0$ define

$$\eta := 2\exp\left(-\frac{n}{2\nu}\log\left(\frac{1+\nu}{\nu}\right)\right) \ .$$

Then with probability exceeding $1-\eta$ it holds that

$$\|D(\partial C_n(f))\|_{L_2} < \frac{\nu+1}{\nu} M_f \ .$$

Proof See Appendix B. ■

We are now ready to state the main Theorem.

Theorem 4.4 *Let $l : \mathbb{R}\times\mathcal{Y} \to [0,+\infty]$ be a well-behaved loss function and C_n in (7) be defined upon an i.i.d. sample $\mathcal{D}_n$. Denote by $f \in \mathcal{H}$ any oracle function known via its additive representation. Fix $\nu > 0$ and let η and M_f be defined as in Proposition 4.3. Further pose*

$$\mu(\mathcal{A}_f) := \frac{\sqrt{\psi(\mathcal{A}_f)}}{\sqrt{\kappa(\mathcal{A}_f)\,(1-\delta^2(\mathcal{A}_f))}}$$

and for $c > 0$ fix $\tau = cM_f$. Assume $\hat{h}^\tau$ defined by (8) satisfies $\hat{h}^\tau \neq f$. Then with probability exceeding $1-\eta$ it holds that

$$\frac{\sum_{j\notin\mathcal{A}_f}\|\hat{h}^{\tau,j}\|_j}{\|f-\hat{h}^\tau\|_{L_2(\mathcal{X},\xi)}} \leq \mu(\mathcal{A}_f) + \frac{1}{c}\left(\frac{2}{\eta}\right)^{\frac{2\nu}{n}} \ . \tag{13}$$

Proof See Appendix B. ■

4.3 Comments on the Oracle Result

The probabilistic inequality (13) emphasizes the role of concurvity in sparse recovery in the following sense. It is assumed that τ is chosen to be proportional to M_f which depends upon the oracle, the decay of the eigenspectrum of k (through D) and the loss. Under this circumstance the left-hand side is probabilistically bounded by the sum of two terms. The first term, namely $\mu(\mathcal{A}_f)$, is related to the structure of the statistical dependence in the additive decomposition of f and on the relative smoothness of the oracle. It is of primary interest here. The second, namely $\frac{1}{c}\left(\frac{2}{\eta}\right)^{\frac{2\nu}{n}}$, is a function of the constant of proportionality, the confidence η and the sample size n. Our interest is not in asymptotics but in relating concurvity and sparse recovery. Thus we shall consider n fixed and the constant c adequately chosen. We stress here that the result does not depend on any specific property of the oracle which can be arbitrarily chosen. However, the bound is of interest for f sufficiently close to the estimate $\hat{h}^\tau$, in the sense of the norm $\|\cdot\|_{L_2(\mathcal{X},\xi)}$. Under this condition the distance can be appropriately bounded. The norm of the "irrelevant" components (according to the oracle), and thus the consistent determination of $\mathcal{A}_f^c$, is then related to $\mu(\mathcal{A}_f)$. When

this is small, the bound is tighter and the irrelevant components are likely to have norm (close to) zero. On the contrary, when this term is big, the solution might include some irrelevant components. This is confirmed by empirical evidence as we shall illustrate in the experimental section. It is also supported by the findings in Meier et al. (2008) which deals with nonparametric modeling via finite dimensional B-splines.

The numerator of $\mu(\mathcal{A}_f)$ depends upon the relative smoothness of the oracle f. We write *relative* here, as the smoothness of each components enters (11) both via the norm of L_2 and the norm of the component spaces. As an intuitive example, consider the following. Let $\overline{\mathbb{Z}} = \mathbb{Z}\backslash\{0\}$, denote by $0 < q < 1$ a fixed *decay index* and let $\mathcal{C}$ be the set of complex sequences $\mathcal{C} := \left\{ \left(c_k \ : \ k \in \overline{\mathbb{Z}}\right) \ : \ c_k \in \mathbb{C},\ c_k = \overline{c_{-k}} \text{ for all } k \in \overline{\mathbb{Z}} \right\}$. Finally assume $\mathcal{X} = \{[0,1]\}^d$ and let the marginal probability measure ξ be corresponding to the Lebesgue product measure. For any $j \in \mathbb{N}_d$, the space

$$\mathcal{H}^j := \left\{ h : \mathcal{X} \to \mathbb{R} \text{ s.t. } h(x) = \sum_{k \in \overline{\mathbb{Z}}} c_k e^{2\pi i k x_j},\ c \in \mathcal{C},\ \|h\|_j^2 := \sum_{k \in \overline{\mathbb{Z}}} \frac{c_k^2}{q^{|k|}} < +\infty \right\}$$

can be shown to be a well-defined RKHS with r.k. $k^j(x,y) = \sum_{k \in \overline{\mathbb{Z}}} q^{|k|} e^{2\pi i k x_j} e^{-2\pi i k y_j}$ taking values in $\mathbb{R}$. Functions in $\mathcal{H}^j$ are smoothly varying along the $j-$th coordinate. Indeed, for any function in the space, the series defining the norm needs to converge and thus the amplitude $2\sqrt{c_k c_{-k}}$ of the $k-$th harmonic $c_k e^{2\pi i k x_j} + c_{-k} e^{-2\pi i k x_j}$ must decay with k in a way determined by q. Notice that, by definition, the functions in $\mathcal{H}^j$ are constant along the coordinates different than j. For any $j \in \mathbb{N}_d$, D^j is defined as

$$\left[D^j(h^j)\right](x) = \sum_{k \in \overline{\mathbb{Z}}} \frac{c_k}{q^{|k|}} e^{2\pi i k x_j} \ .$$

It is not difficult to see now that the relative smoothness $\psi(\mathcal{A})$ of $f = \sum_{j \in \mathcal{A}} f^j,\ f^j \in \mathcal{H}^j$ is small if all the components have energy concentrated at low harmonics while it is usually larger on the opposite case.

For what concerns the denominator of $\mu(\mathcal{A}_f)$, interestingly it resembles in a way the quantity appearing in the conditions for parametric sparse recovery via l_1 norm (see e.g. the *Irrepresentable Condition* of Zhao and Yu (2006)). As in that case, here the dependence structure appears via the *intra*-dependence (within the relevant components) and the *inter*-dependence (between functions in $\mathcal{H}_{\mathcal{A}_f}$ and $\mathcal{H}_{\mathcal{A}_f^c}$). To illustrate the main idea consider again the set of spaces introduced above. If ξ is indeed the uniform measure then no matter the specific choice of the oracle f, $\mu(\mathcal{A}_f) = \sqrt{\psi(\mathcal{A}_f)}$ as the mutual coherence is zero and $\kappa(\mathcal{A}_f) = 1$. This is because the component spaces are by definition mutually orthogonal subspaces of $L_2(\mathcal{X}, \xi)$. In this case the bound in equation (13) is just in terms of the relative smoothness of the oracle, n and the confidence. On the contrary, when the marginal measure ξ is not uniform, the component spaces are no longer orthogonal in $\left(L_2(\mathcal{X}, \xi), \langle \cdot, \cdot \rangle_{L_2(\mathcal{X},\xi)}\right)$ and both $\delta(\mathcal{A})$ and $\kappa(\mathcal{A})$ are generally non-zero. Notice that the same comments apply to tensor product smoothing splines (Wahba (1990); Gu (2002)). In this case functions in the additive decompositions might also correspond to interaction between covariates. As in the example above, the components lie in orthogonal subspaces of $L_2\left(\{[0,1]\}^d, \nu\right)$ where ν is again the Lebesgue product measure.

To the best of our knowledge Koltchinskii and Yuan (2008) is until now the only reference for learning bound in sparse recovery in infinite dimensional RKHS's. Their main result is more elaborated and the role of the excess risk is also highlighted. This comes at a price of sacrifying interpretability as the dependence between concurvity and sparse recovery is somewhat hidden. Being simpler (even though less explicit) proposition 4.4 addresses this inconvenience.

5. Data-Dependent Norm Adaptation

Motivated by the oracle result, in this Section we propose a data-dependent approach to improve the sparse recovery in presence of concurvity. The framework we are about to discuss is in a similar spirit as the data-dependent geometric regularization, introduced in the context of semi-supervised learning in Belkin et al. (2006). The aim is to incorporate in the learning task the geometry of the underlying probability distribution of the input data. We begin by introducing a class on norms for the component spaces, which is instrumental for our approach.

5.1 Data-Dependent Adaptive Norms

Our starting point is the set of spaces $\left\{\left(\mathcal{H}^j, \langle\cdot,\cdot\rangle_j\right) \ : \ j \in \mathbb{N}_m\right\}$ introduced above. It is assumed that $m \leq n$ that is, we have at most as many components as points in the sample. For $j \in \mathbb{N}_m$ denote by $S^j : \mathcal{H}^j \to \mathbb{R}^n$ a bounded linear operator. In particular we shall focus on the composition of operators $S^j := A^j \bullet I_{\mathcal{D}_n,j}$ where $I_{\mathcal{D}_n,j}\left(h^j\right) = \left(h^j\left(x^i\right) \ : \ (x^i, y_i) \in \mathcal{D}_n\right)$ is the vector of evaluations of h^j according to the input points in $\mathcal{D}_n$ and A^j denotes both a linear operator $A^j : \mathbb{R}^n \to \mathbb{R}^n$ and the associated square matrix. Notice that the we have not specify at this stage the nature of A^j; in what follows a specific choice will play a central role in our derivation. We now modify the original norm by incorporating a new term depending upon the sample. Namely for $h^j \in \mathcal{H}^j$ and $\mu > 0$ we define the new norm

$$\left\|h^j\right\|^2_{\mathcal{D}_n,j} := \left\|h^j\right\|^2_j + \mu \left\|S^j(h^j)\right\|^2_{\mathbb{R}^n} \ . \tag{14}$$

When endowed with this norm, $\mathcal{H}^j$ can be shown to be a RKHS with the r.k. $k^j_{\mathcal{D}_n} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ defined by:

$$k^j_{\mathcal{D}_n}(x,y) = k^j(x,y) - \left\langle \bar{k}^j(x), \left(\frac{1}{\mu} I + M^j K^j\right)^{-1} M^j \bar{k}^j(y) \right\rangle_{\mathbb{R}^n} \tag{15}$$

where $\bar{k}^j(x)$ is the vector $\left(k^j(x_i, x) \ : \ i \in \mathbb{R}^n\right)$, M^j is the positive semi-definite matrix ${A^j}^\top A^j$ and K^j is the kernel matrix arising from the kernel function k^j and the points in the given sample. A proof is essentially based on simple orthogonality arguments and can be found in Sindhwani et al. (2005). In the context of semi-supervised learning the idea is to deform the (single) RKHS along a finite-dimensional subspace spanned by the data. This is typically done based upon the Laplacian associated to the input points in the dataset. Our approach differs substantially at this point: in the problem we are about to formulate, the set of matrices $\{M^j \ : \ j \in \mathbb{N}_m\}$ is not fixed. In (14) the parameter μ balances the trade-off between the two contributions. Notice that we adopt the same parameter for all the

component spaces. This choice is clearly non restrictive and it is made here purely for computational reasons. Indeed in practice μ needs to be chosen according to the data via some (usually) expensive model selection procedure.

Based on this set of norms we finally define the new norm for the space of additive models $\mathcal{H} = \oplus_{j \in \mathbb{N}_m} \mathcal{H}^j$:

$$\|h\|_{\mathcal{D}_n}^2 = \inf \left\{ \sum_{j \in \mathbb{N}_m} \|h^j\|_{\mathcal{D}_n, j}^2 \ : \ h = \sum_{j \in \mathbb{N}_m} h^j, \ h^j \in \mathcal{H}^j, \ j \in \mathbb{N}_m \right\} . \tag{16}$$

Correspondingly the associated reproducing kernel is now

$$k_{\mathcal{D}_n}(x, y) = \sum_{j \in \mathbb{N}_m} k_{\mathcal{D}_n}^j(x, y) \ . \tag{17}$$

Let now $\{\mathcal{R}^j \ : \ j \in \mathbb{N}_m\}$ be a set of one-dimensional mutually orthogonal subspaces of $\mathbb{R}^n$. Denote by $P_\perp^j$ the projector operator mapping $x \in \mathbb{R}^n$ onto the orthogonal complement of $\mathcal{R}^j$. We address the adaptation in a supervised way by formulating the following *data-dependent norm adaptation problem.* For $\lambda_0 > 0$ consider finding $\left\{ P_\perp^j \ : \ j \in \mathbb{N}_m \right\}$ and $h = \sum_{j \in \mathbb{N}_m} h^j \in \mathcal{H}$ which minimize:

$$Q\left(h, \left\{ P_\perp^j \ : \ j \in \mathbb{N}_m \right\}\right) := C_n(h) + \lambda_0 \sum_{j \in \mathbb{N}_m} \left(\|h^j\|_j^2 + \mu \left\| P_\perp^j \bullet I_{\mathcal{D}_n, j}\left(h^j\right) \right\|_{\mathbb{R}^n}^2 \right) . \tag{18}$$

By minimizing (18) we fit an additive model to the data meanwhile penalizing the coherence with respect to the empirical measure. This is achieved by driving the vectors in $\left\{ I_{\mathcal{D}_n, j}\left(h^j\right) \ : \ j \in \mathbb{N}_m \right\}$ to live in orthogonal subspaces. The empirical measure is used as a proxy for the underlying measure and the approach is readily motivated in the spirit of the oracle inequality. Notice however that while the analysis in the latter is based on a worst-case scenario, the "decorrelation" of the components we aim at is data-driven. When μ is set to zero the problem correspond to a standard penalized empirical risk minimization. As μ is increased, the role of the second penalty becomes more and more relevant. The smoothness of the solution is traded for the geometric property of having less inter-dependence with respect to the empirical measure.

5.2 An Algorithm for the Data-Dependent Norm Adaptation Problem

Once the set of projectors $\left\{ P_\perp^j \ : \ j \in \mathbb{N}_m \right\}$ has been fixed, (18) can be restated as

$$Q(h) = C_n(h) + \lambda_0 \|h\|_{\mathcal{D}_n}^2 \ ,$$

which is a regularized empirical risk functional with data-dependent squared norm penalty. It is well known that a minimizer of the latter admits a representation as $\sum_i \alpha_i k_{\mathcal{D}_n}\left(x^i, \cdot\right)$ where $\alpha \in \mathbb{R}^n$ and $k_{\mathcal{D}_n}$ is given by (17) in terms of the modified kernel functions (15) with $M^j = P_\perp^j$ and μ fixed at a predefined value. Such a minimizer can be computed by solving an optimization problem which depends on the choice of the loss function. In the case of quadratic loss a solution is available in closed form and can be computed solving a system of

linear equations (Wahba (1990); Suykens et al. (2002)). On the other hand, if we fix $h \in \mathcal{H}$, the minimization of (18) amounts to find projection matrices $\{P^j \ : \ j \in \mathbb{N}_d\}$ that minimize

$$\sum_j \left\| \left(I_{\mathcal{D}_n,j} - P^j \bullet I_{\mathcal{D}_n,j} \right) \left(h^j \right) \right\|_{\mathbb{R}^n}^2$$

where for any $j \in \mathbb{N}_m$ P^j is the projection onto $\mathcal{R}^j$. Denote now by H the $n \times m$ matrix with $j-$th column $I_{\mathcal{D}_n,j}\left(h^j\right)$ and let $\| \cdot \|_F$ be the Frobenious norm. Further, for a positive integer s, denote by I_s the $s \times s$ identity matrix and by $\mathrm{diag}(x)$ a diagonal matrix with diagonal $x \in \mathbb{R}^s$. The latter problem can be equivalently formulated as

$$\min \left\{ \|H - Q\Lambda\|_F \ : \ Q^\top Q = I_m, \ \Lambda = \mathrm{diag}(\lambda), \ \lambda \in \mathbb{R}^m \right\} . \tag{19}$$

Once an approximation $Q\Lambda$ is found, it is immediate to find the set of projection matrices $\left\{ P_\perp^j \ : \ j \in \mathbb{N}_m \right\}$. Indeed if we denote by q^i the vector corresponding to the $i-$th column of Q, we clearly have $P_\perp^j = I_n - q^i (q^i)^\top$. In turn the minimization of (19) can be performed efficiently with the algorithm in the next Subsection. Therefore a reasonable scheme for the minimization of (18) consists in alternating between the minimization of a regularized empirical risk functional and the minimization of (19). The latter is referenced in the following as the *component-wise orthogonal approximation* problem.

Algorithm Implementation

A natural initial value for the set of projection matrices can be obtained as follows. First (18) is solved for h with μ set to 0. Denoted the obtained solution by $\hat{h}$, we then find $(\hat{Q}, \hat{\Lambda})$ by solving (19) for the value of the matrix H associated to $\hat{h}$. For $j \in \mathbb{N}_m$, an initial value for $P_\perp^j$ can now be readily computed, based upon the $j-$th column of $\hat{Q}$. It is now easy to find a set of initial kernel functions by means of (15) and, from these, the composite kernel (17) associated to the norm (16). Our algorithm then starts by the solution of the regularized risk functional where the norm is determined by this set of initial projection matrices. We summarize the procedure in Table 1. In our experience the value of the global objective functional usually decreases rapidly in the first iterations and then converges slowly. In applications, though, an exact solution is not needed. By applying a limited number of iterations of our algorithm, we get what can be seen as an iterative data-dependent improvement on the original set of norms. Thus our stopping criterion is simply determined by a predefined number of iterations (10 in our experiments).

Although alternating schemes can be slow in convergence, they are easy to implement and flexible. The flexibility is underpinned by the fact that it is easy to replace one block with another one. This means, for example, that one can solve the optimization problem associated to different loss functions in (18) without modifying the overall solution strategy. Moreover, efficient standard solvers can be employed as the overall and specialized problem is divided into simpler and standard ones. These type of algorithms are common in statistics. Two popular examples are the Backfitting Algorithm of Buja et al. (1989) and the COSSO of Lin and Zhang (2006) about which we shall say more in Section 6.

We conclude by stressing that the problem we are addressing here is simply the data-dependent norm adaptation. The sparse recovery problem will be discussed later and it will

take advantage of such an adaptation. Hence, the final outcome of the procedure we have been discussing in this Section is the set of modified kernels $\{k^j_{\mathcal{D}_n} \ : \ j \in \mathbb{N}_m\}$. We shall consider the function minimizing (18) as a by-product of the optimization.

Table 1: Outline of the algorithm in Section 5.2.

Pseudocode for the data-dependent norm adaptation problem	
Input:	Dataset $\mathcal{D}_n$; parameters $\lambda_0,\ \mu$; set of *kernel functions* $\left\{k^j \ : \ j \in \mathbb{N}_m\right\}$; stopping criterion.
Output:	Set of projection matrices $\left\{P^j_\perp \ : \ j \in \mathbb{N}_m\right\}$ and corresponding set of *data-dependent kernel functions* $\left\{k^j_{\mathcal{D}_n} \ : \ j \in \mathbb{N}_m\right\}$.
Algorithm:	1. Find the minimizer $\hat{h}$ of (18) for the given value of λ_0 and $\mu = 0$. 2. Component-wise orthogonal approximation step: call the algorithm in Section 5.3 to find a solution $(\hat{Q}, \hat{\Lambda})$ of (19) where H is set based on $\hat{h}$. 3. For q^i $i-$th column of $\hat{Q}$ and I_n $n \times n$ identity matrix, set $P^j_\perp = I_n - q^i(q^i)^\top$. 4. Find via (15) the *data-dependent kernel functions* associated to the projection matrices in the last step. 5. Iterate until stopping criterion is met: (a) Find the minimizer $\hat{h}$ of (18) for the given values of λ_0, μ and norm (16) corresponding to the data-dependent kernel functions found at point 4. (b) Execute step 2, 3 and 4 above.

5.3 Solving the component-wise orthogonal approximation problem

In this Section, we shall elaborate more on problem (19) which, given H, seeks the best approximation $Q\Lambda$ with the prescribed structure.

For a $n \times m$ matrix $\bar{Q}$ with orthonormal columns and a upper-triangular $m \times m$ matrix $\bar{R}$, let $H = \bar{Q}\bar{R}$ be the *thin QR factorization* of H (Golub and Van Loan (1996)). Denote by $\text{ran}(S)$ the range of a $n \times m$ matrix S i.e. $\text{ran}(S) = \{y \in \mathbb{R}^n \ : \ y = Sx,\ x \in \mathbb{R}^m\}$. In order to achieve the minimum of $\|H - Q\Lambda\|^2_F$, we must have that $\text{ran}(Q) \subseteq \text{ran}(H)$. Since $\text{ran}(H) = \text{ran}(\bar{Q})$ we can write $Q = \bar{Q}\tilde{Q}$ where $\tilde{Q}$ is an orthogonal $m \times m$ matrix so that $Q^\top Q = I_m$. Now we have the chain of equalities

$$\left\|H - Q\Lambda\right\|^2_F = \left\|\bar{Q}\bar{R} - \bar{Q}\tilde{Q}\Lambda\right\|^2_F = \left\|\bar{R} - \tilde{Q}\Lambda\right\|^2_F = \left\|\Lambda - \tilde{Q}^\top\bar{R}\right\|^2_F = \left\|\Lambda - \bar{R}^\top\tilde{Q}\right\|^2_F . \quad (20)$$

For given $\bar{R}$ and fixed Λ, the minimization over $\tilde{Q}$ of the last term in the chain amounts to find an orthogonal transformation that most closely rotates $\bar{R}^\top$ into Λ. This is an instance of the so-called *orthogonal Procrustes problem* (Golub and Van Loan (1996)), studied in

linear algebra. A solution can be computed via the SVD. More precisely, if

$$U^\top(\bar{R}\Lambda)V = \operatorname{diag}\left((\sigma_i \ : \ i \in \mathbb{N}_m)\right)$$

is the SVD of $\bar{R}\Lambda$, then the minimum of $\|H - Q\Lambda\|_F^2$ conditioned on Λ is attained at $Q = \bar{Q}UV^\top$, as shown in Golub and Van Loan (1996). On the other hand, conditioned on the current value of $\tilde{Q}$, the minimization w.r.t. Λ can be carried out again based upon (20). Indeed the first order optimality condition $\partial_\Lambda \|\bar{R} - \tilde{Q}\Lambda\|_F = 0$ easily leads to

$$\Lambda = \operatorname{diag}\left(\left(\left(\tilde{Q}^\top \bar{R}\right)_{jj} \ : \ j \in \mathbb{N}_m\right)\right) \ .$$

These facts naturally give rise to an alternating least squares approach preceded by the QR factorization of H. The alternating part closely resembles algorithms that were proposed in the context of Multi-way Analysis (see e.g. Smilde et al. (2004)). Any time the problem is solved in either one of the unknown, the least-squares error $\|H - Q\Lambda\|_F^2$ is not increased and usually decreases. Because the problem is bounded below by zero, convergence usually follows (De Leeuw et al. (1976)).

The algorithm we have been discussing here, summarized in Table 2, might be of interest by itself. In the present setting it forms the component-wise orthogonal approximation step in the iterative procedure for solving (18). In the practical implementation we compute the QR factorization followed by a single update of $\tilde{Q}$ and Λ (i.e., we execute just once the steps 3.(a) and 3.(b) in Table 2). This already leads to satisfactory results for the overall minimization of (18).

Table 2: Outline of the algorithm in Section 5.3.

Pseudocode for the component-wise orthogonal approximation problem	
Input:	$n \times m$ matrix H; stopping criterion.
Output:	$n \times m$ matrix Q such that $Q^\top Q = I_m$ and $m \times m$ diagonal matrix Λ.
Algorithm:	1. Compute the thin QR decomposition of H, $H = \bar{Q}\bar{R}$. 2. Initialization: set $\Lambda = \operatorname{diag}\left(\left(\bar{R}_{jj} \ : \ j \in \mathbb{N}_m\right)\right)$. 3. Iterate until stopping criterion is met: (a) $\tilde{Q} = UV^\top$ where $U^\top(\bar{R}\Lambda)V = \operatorname{diag}((\sigma_i \ : \ i \in \mathbb{N}_m))$ is the SVD of $\bar{R}\Lambda$. (b) Set $\Lambda = \operatorname{diag}\left(\left(\left(\tilde{Q}^\top \bar{R}\right)_{jj} \ : \ j \in \mathbb{N}_m\right)\right)$. 4. Set $Q = \bar{Q}\tilde{Q}$.

6. Sparse Functional ANOVA Models with Data-Dependent Norms

In this section we go back to the estimator (6) used in combination with quadratic loss. In this case the problem is almost the same as the one studied in Lin and Zhang (2006) under the name of Component Selection and Smoothing Operator (COSSO). The main difference with the latter is that we have not considered in the derivation an unpenalized bias term. Including the bias involves minor changes and it has not been reported above for simplicity of presentation. We also remark that the analysis in Lin and Zhang (2006) is done explicitly in terms of tensor product smoothing splines (Wahba (1990)). In principle we are not restricted to this framework, although the latter represents a solid formalism to deal with multivariate data. More comments about this are postponed until the next Section. A discussion on the link between the MKL framework and the COSSO can also be found in Micchelli and Pontil (2007). Our interest in here is in comparing the performances of the estimator with and without data-dependent norm adaptation (respectively ADACOSSO and COSSO here below).

6.1 The Adaptive COSSO

It is known that for fixed τ a solution for (7) amounts to find in a predefined model space $\mathcal{H} \oplus \{1\}$ the solution of :

$$\min \left\{ C_n(h) + \lambda_0 \sum_{j \in \mathbb{N}_m} \theta_j^{-1} \|h^j\|_{\star,j}^2 + \lambda \sum_{j \in \mathbb{N}_m} \theta_j \ : \ h \in \mathcal{H} \oplus \{1\},\ \theta_j \geq 0,\ j \in \mathbb{N}_m \right\} \qquad (21)$$

where $\{1\}$ is the space of unpenalized constant functions, the pair (λ_0, λ) has a known relation with τ and λ_0 can be fixed at any value. Details and proofs can be found in Lin and Zhang (2006). In our setting $\|\cdot\|_{\star,j}$ is either the standard norm $\|\cdot\|_j$ or the data-dependent one $\|\cdot\|_{\mathcal{D}_n,j}$, as introduced in Section 5. Associated to these norms we have respectively the original kernel or the modified one, which we denote in a unified way as $k_\star^j$. When the norms are the standard ones we refer to (21) as the COSSO estimator and in either case we denote by $(\hat{h}, \hat{\theta})$ a solution of (21). In Lin and Zhang (2006) it is proposed to find $(\hat{h}, \hat{\theta})$ by an alternating scheme. Adapting the argument to the present setting, this is motivated by the fact that for any fixed θ a solution over h of the functional in (21) has the form

$$\sum_{i \in \mathbb{N}_n} \alpha_i k_\star^\theta \left(x^i, x\right) + b$$

where $\alpha \in \mathbb{R}^n$, $b \in \mathbb{R}$ and finally

$$k_\star^\theta (x, y) = \sum_{j \in \mathbb{N}_m} \theta_j k_\star^j (x, y) \ .$$

The unknown α and b can be computed solving a linear system. In turn, for fixed h the optimization in terms of θ corresponds to the nonnegative garrote of Breiman (1995). The latter makes it possible to set some θ_j to zero, thus ensuring sparsity in the component functions. Different computational strategies were also proposed for closely related problems. One method is based on quadratically constrained quadratic programming (Lanckriet et al.

(2004a); Signoretto et al. (2008)). Semi-infinite programming has been considered for large-scale problems in Sonnenburg et al. (2006) and yet another computational approach has been recently studied in Rakotomamonjy et al. (2008).

Ensemble of Initial Kernels and Other Implementation Aspects

One relevant aspect concerning the practical implementation is the choice of the model space. Equivalently, this amounts to choose the ensemble of initial kernel functions $\left\{k^j \ : \ j \in \mathbb{N}_m\right\}$, each one encoding a specific feature from the input domain. Different choices are possible at this level. A popular formalism in nonparametric statistics is represented by the smoothing spline ANOVA framework where the features are represented by main effects or interactions between the covariates. In this case each component space is a subspace of a tensor product space arising from $i-$th order Sobolev spaces on the unit interval. In particular it is common to take $i = 2$, as done in Lin and Zhang (2006). The mathematical background as well as the closed form for the ensemble of kernel functions can be found e.g. in Wahba (1990), Chen (1993), Gu (2002), Berlinet and Thomas-Agnan (2004).

Alternatively one might consider tensor product of regularized Fourier expansions as in Subsection 4.3 (see also Chapter 6.5 and 6.6 of Vapnik (1995)). Also, it is possible to exploit the tensor product structure of the Gaussian RBF kernel (Steinwart et al. (2006)) to derive a similar construction. In this case each component space arises from a Gaussian RBF kernel defined as[1] $k^{\mathcal{A}}_{RBF}(x, y) = \exp\left(-\sigma^2 \sum_{j\in\mathcal{A}}(x_j - y_j)^2\right)$ where each $\mathcal{A} \subseteq \mathbb{N}_d$ is appropriately chosen to have a predefined set of main effects and/or interaction terms in the additive decomposition of the model. However, in contrast to the smoothing spline ANOVA kernels, the latter cases require to set one or more kernel parameters and this implies a more involved model selection strategy. Besides, in our experience the fitting corresponding to these initial kernels is not rewarded by any performance improvement in comparison with the one associated to smoothing spline ANOVA kernels. For these reasons we opt for the latter and chose $i = 2$ as in the common use. We consider pure additive models where each component function dependents upon a single distinct covariate. In the simple case without norm adaptation, the model is directly fitted via the COSSO estimator. In case of norm adaptation we first find the set of projectors as detailed in section 5. Successively we fit the model via the COSSO in association with the set of modified kernels. This gives rise to the ADACOSSO procedure as sketched in Table 3. See also the original paper Lin and Zhang (2006) for further details about the implementation of COSSO.

We remark here that the norm adaptation naturally integrates with the algorithmical approach used in the COSSO. The standard COSSO starts from an estimate corresponding to $\theta_j = 1$ for all $j \in \mathbb{N}_m$. At this point we perform what can be viewed as a *preconditioning* scheme at the function space level. Nevertheless we stress that the norm adaptation problem presented in Section 5 is of greater generality. It might be specialized to other loss functions and used in combination with other MKL algorithms. It copes with a problem which is not attached to the specific algorithmical implementation. Rather, it deals with the intrinsic nature of sparse recovery in function spaces.

1. Here it is assumed the same parameter is used for all the marginal spaces.

Table 3: Outline of the COSSO with data-dependent norms.

Pseudocode for the ADACOSSO	
Input:	Dataset $\mathcal{D}_n$; parameters $\lambda_0,\ \mu,\ \lambda$; set of *kernel functions* $\left\{k^j\ :\ j \in \mathbb{N}_m\right\}$; stopping criterion for the norm adaptation; COSSO stopping criterion.
Output:	Estimate $\hat{h}(x) = \sum_{i\in\mathbb{N}_n} \hat{\alpha}_i k^{\hat{\theta}}_{\mathcal{D}_n}\left(x^i, x\right) + \hat{b}$.
Algorithm:	1. Data-dependent norm adaptation: call the algorithm in Table 1 for the given input $\mathcal{D}_n$; parameters $\lambda_0,\ \mu$; set of *kernel functions* $\left\{k^j\ :\ j \in \mathbb{N}_m\right\}$; stopping criterion for the norm adaptation. 2. Initialization for COSSO: build the kernel matrix associated to the kernel function (17) where $\left\{k^j_{\mathcal{D}_n}\ :\ j \in \mathbb{N}_m\right\}$ is the set of data-dependent kernels found at the previous step. 3. Solve the linear system for α and b with the given value of λ_0. 4. Iterate until COSSO stopping criterion is met: (a) Garrote-step: for $\alpha,\ b$ fixed solve for θ with the given value of λ. (b) Build the kernel matrix associated to the weighted kernel function $k^\theta = \sum_{j\in\mathbb{N}_m} \theta_j k^j_{\mathcal{D}_n}$. (c) Solve the linear system for α and b with the given value of λ_0.

6.2 On Parameters Tuning

As for any other model fitting procedure an adequate choice of the parameters in (21) is important to achieve good results. In this Subsection we discuss possible strategies for tuning the parameters. As a general approach we suggest to proceed in a step-wise manner. First of all we follow Lin and Zhang (2006) and fix λ_0 at a value which gives an appropriate scaling of the problem. This is performed according to 5−fold cross validation keeping $\theta_j = 1$ for any $j \in \mathbb{N}_m$. We do this by using the kernel functions associated to the standard norm, that is, we equivalently minimize the 5−fold CV for the functional (18) when the norm parameter μ is set to zero. If we then perform norm adaptation we select μ keeping λ_0 fixed at this value. To our experience the choice of μ is not critical as long as the chosen value is not too high, where "high" depends on the specific dataset. In this case the resulting component selection is too greedy and some relevant components are shrunk to zero. A possible criterion at this level is the Generalized Cross Validation (GCV) of Golub et al. (1979). Let A be the smoothing matrix satisfying $\hat{y} = I_{\mathcal{D}_n}\hat{h} = Ay$ (Hastie and Tibshirani (1990); Wahba (1990)) and depending upon μ. Then we recall that

$$\mathrm{GCV} = \frac{\frac{1}{n}\|y - \hat{y}\|_n^2}{\left(\frac{1}{n}\operatorname{trace}(I - A)\right)^2}\ . \tag{22}$$

The empirical evidence shows however that this strategy sometimes lead to such an overly greedy situation. When the focus is on model selection, it is often recommended to use the Bayesian Information Criterion (BIC, Schwarz (1978)), see e.g. Zou et al. (2007). In this case we use

$$\text{BIC} = \log\left(\|y - \hat{y}\|_n^2\right) + \varsigma \log(n)/n \tag{23}$$

where trace(A) is employed as a proxy for the degrees of freedom ς. Our preliminary experience shows that adopting this as a criterion for the choice of μ, generally leads to better results in terms of sparse recovery. The same principle is used both in COSSO and ADACOSSO for the selection of a value for λ in (21), which is the last stage in our step-wise tuning procedure.

7. Experimental Results

Table 4: Results for the synthetic dataset with $m = 20$, $n = 50$, SNR $= 10$.

		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
Pr	COSSO	0.57(0.21)	0.54(0.17)	0.46(0.13)	0.39(0.12)	0.36(0.11)
	ADACOSSO	**0.75**(0.18)	**0.85**(0.17)	**0.85**(0.17)	**0.82**(0.17)	**0.78**(0.18)
Re	COSSO	0.98(0.07)	**0.99**(0.05)	**0.93**(0.14)	**0.86**(0.19)	**0.8**(0.22)
	ADACOSSO	**0.99**(0.03)	0.97(0.11)	0.8(0.17)	0.8(0.16)	0.74(0.18)
RSS	COSSO	5.74(2.24)	3.14(0.86)	2.79(0.68)	2.67(0.52)	2.53(0.42)
	ADACOSSO	**4.15**(1)	**2.47**(0.73)	**2.32**(0.53)	**2**(0.4)	**1.83**(0.32)

Table 5: Results for the synthetic dataset with $m = 20$, $n = 50$, SNR $= 3$.

		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
Pr	COSSO	0.53(0.27)	0.51(0.18)	0.42(0.14)	0.35(0.11)	0.32(0.11)
	ADACOSSO	**0.57**(0.17)	**0.71**(0.19)	**0.73**(0.21)	**0.68**(0.21)	**0.67**(0.23)
Re	COSSO	0.8(0.25)	0.86(0.2)	**0.83**(0.17)	**0.69**(0.21)	**0.62**(0.24)
	ADACOSSO	**0.95**(0.09)	**0.88**(0.16)	0.7(0.2)	0.62(0.2)	0.59(0.21)
RSS	COSSO	11.39(3.18)	5.78(1.5)	5.44(1.17)	5.35(1.02)	5.23(0.81)
	ADACOSSO	**8.19**(1.72)	**4.59**(0.92)	**4.36**(0.72)	**4.19**(0.62)	**4.02**(0.54)

Table 6: Results for the synthetic dataset with $m = 100$, $n = 100$, SNR $= 10$.

		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
Pr	COSSO	0.53(0.22)	0.44(0.15)	0.31(0.13)	0.26(0.08)	0.22(0.1)
	ADACOSSO	**0.74**(0.18)	**0.98**(0.06)	**0.83**(0.16)	**0.67**(0.19)	**0.54**(0.2)
Re	COSSO	0.83(0.22)	**0.9**(0.14)	0.92(0.14)	**0.78**(0.19)	0.62(0.21)
	ADACOSSO	**0.99**(0.05)	0.89(0.14)	**0.93**(0.12)	0.77(0.18)	**0.67**(0.21)
RSS	COSSO	7.87(1.24)	4.15(0.61)	3.24(0.55)	2.87(0.4)	2.7(0.32)
	ADACOSSO	5.54(1.14)	**2.74**(0.59)	**2.35**(0.39)	**2.32**(0.32)	**2.25**(0.29)

Table 7: Results for the synthetic dataset with $m = 100$, $n = 100$, SNR $= 3$.

		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
Pr	COSSO	0.54(0.24)	0.39(0.16)	0.25(0.12)	0.16(0.08)	0.13(0.07)
	ADACOSSO	**0.63**(0.18)	**0.96**(0.1)	**0.76**(0.2)	**0.56**(0.23)	**0.43**(0.24)
Re	COSSO	0.62(0.24)	0.79(0.19)	0.76(0.23)	0.57(0.23)	0.47(0.24)
	ADACOSSO	**0.96**(0.1)	**0.82**(0.17)	**0.77**(0.18)	**0.6**(0.22)	**0.48**(0.22)
RSS	COSSO	11.22(1.13)	6.21(0.71)	5.53(0.65)	5.2(0.54)	4.99(0.46)
	ADACOSSO	**8.76**(1.3)	**4.43**(0.68)	**4.21**(0.5)	**4.28**(0.44)	**4.28**(0.4)

Table 8: Results for the synthetic dataset with $m = 100$, $n = 200$, SNR $= 10$.

		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
Pr	COSSO	0.94(0.1)	0.85(0.15)	0.77(0.16)	0.67(0.18)	0.58(0.19)
	ADACOSSO	**0.94**(0.1)	**1**(0)	**0.99**(0.04)	**0.93**(0.12)	**0.87**(0.16)
Re	COSSO	**1**(0)	**1**(0)	**1**(0.02)	**0.91**(0.12)	**0.77**(0.15)
	ADACOSSO	1(0)	0.89(0.16)	0.95(0.1)	0.78(0.15)	0.68(0.16)
RSS	COSSO	**4.79**(0.54)	2.78(0.31)	2.26(0.23)	2.27(0.19)	2.3(0.17)
	ADACOSSO	5(0.9)	**2.7**(0.6)	**2.19**(0.24)	**2.25**(0.23)	**2.25**(0.19)

In this Section we compare the COSSO estimator with the ADACOSSO outlined in the previous section. For the implementation of COSSO we base our code on the MATLAB routines available from the authors of Lin and Zhang (2006).

7.1 Performance Indicators

For the analysis of input selection performances of the final models we resort to two popular measures, namely *Precision* and *Recall*. Given an oracle f, Precision (Pr) is the number of true positives divided by the total number of components selected, i.e.,

$$\mathrm{Pr} = \frac{\mathrm{card}\left(\mathcal{A}_{\hat{h}} \cap \mathcal{A}_f\right)}{\mathrm{card}\left(\mathcal{A}_{\hat{h}}\right)} \ .$$

Recall (Re) is defined as the number of true positives divided by the total number of components that are actually relevant, i.e.,

$$\mathrm{Re} = \frac{\mathrm{card}\left(\mathcal{A}_{\hat{h}} \cap \mathcal{A}_f\right)}{\mathrm{card}\left(\mathcal{A}_f\right)} \ .$$

$\mathrm{Pr} = 1$ is achieved if every component retrieved by the estimate was indeed relevant according to the oracle. On the other hand, $\mathrm{Re} = 1$ means that all relevant components of f were retrieved by $\hat{h}$ even though some irrelevant ones might have been included.

For assessing the prediction performances we consider the residual sum of squares

$$\mathrm{RSS} = \frac{1}{n_t} \left\| y_t - \hat{y}_t \right\|^2_{\mathbb{R}^{n_t}} \ ,$$

where y_t is the target vector associated to a new i.i.d. sample from the underlying distribution, $\hat{y}_t$ is the corresponding estimate and n_t is the sample size.

Table 9: Results for the synthetic dataset with $m = 100$, $n = 200$, SNR $= 3$.

		$t=0$	$t=1$	$t=2$	$t=3$	$t=4$
Pr	COSSO	0.83(0.13)	0.72(0.19)	0.54(0.14)	0.4(0.13)	0.33(0.14)
	ADACOSSO	**0.86**(0.13)	**1**(0.02)	**0.96**(0.1)	**0.87**(0.17)	**0.77**(0.24)
Re	COSSO	0.97(0.08)	**0.97**(0.08)	**0.98**(0.07)	**0.79**(0.17)	0.**65**(0.2)
	ADACOSSO	**0.99**(0.04)	0.81(0.18)	0.85(0.16)	0.69(0.17)	0.56(0.19)
RSS	COSSO	8.35(0.87)	4.59(0.46)	4.13(0.36)	4.27(0.29)	4.35(0.26)
	ADACOSSO	**7.74**(1.01)	**4.33**(0.63)	**3.88**(0.35)	**4.05**(0.27)	**4.18**(0.25)

7.2 Synthetic Datasets

We tested the performance of the two procedures on a number of synthetic examples. We use as building blocks for our generating models the same collection of univariate functions used in Lin and Zhang (2006):

$$g_1(t) = t \ , \qquad g_2(t) = (2t-1)^2 \ , \qquad g_3(t) = \frac{\sin(2\pi t)}{2 - sin(2\pi t)} \ ,$$
$$g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t) \ .$$

It is assumed that the sparsity pattern of a target oracle f in the model space (in the sense of Section 2) corresponds here to the relevant covariates according to the generating function.

We consider input data with covariance structure following the *Compound Symmetry* scheme, again used in Lin and Zhang (2006). For $j \in \mathbb{N}_m$, we generate $x_j = (w_j + tu)/(t+1)$, where w_j and u are both i.i.d. from a uniform distribution in $[0, 1]$ and u represents a factor common to all the covariates. In this way, $\text{corr}(x_j, x_k) = t^2/(t^2+1)$ for $j \neq k$ and we can use the parameter t to induce a certain degree of concurvity between the component functions in the model space. The additive generating model is taken to be

$$f(x) = 6g_1(x_1) + 6g_2(x_2) + 3g_3(x_3) + 3g_4(x_4)$$

and therefore $x_5, \ldots, x_m$ are redundant. For $i \in \mathbb{N}_n$, we draw observations from the target variable y according to

$$y_i = 6g_1(x_1^i) + 6g_2(x_2^i) + 3g_3(x_3^i) + 3g_4(x_4^i) + \epsilon_i \ ,$$

where the variance of the zero-mean gaussian variable ϵ is set so that the signal-to-noise ratio SNR $= \text{var}[f(x)]/\text{var}[\epsilon]$ is at a pre-specified value. We compare the estimated model with the original kernel functions, namely, $\sum_{i\in\mathbb{N}_n} \hat{\alpha}_i k^{\hat{\theta}}\left(x^i, x\right) + \hat{b}$, with the estimated model based on the set of adapted kernels, i.e., $\sum_{i\in\mathbb{N}_n} \hat{\tilde{\alpha}}_i k^{\hat{\tilde{\theta}}}_{\mathcal{D}_n}\left(x^i, x\right) + \hat{\tilde{b}}$. The comparison is done in terms of both selection and prediction performances. For the tuning of the parameters we follow the strategy in Subsection 6.2. In particular, for what concerns μ, we select the element of a vector of logarithmically spaced values that minimizes the BIC criterion. The same comment applies to λ. We do the comparison for increasing values of t at different regimes, namely for different values of m, n and SNR. We take $n_t = 1000$ when evaluating

the prediction performances of the estimated models on an independent test set. We run the simulations 100 times. The results are reported on Tables 4-9 in terms of means and standard deviations of the quantities introduced in the previous Subsection. We use boldface to highlight the best performance. The data-dependent norm adaptation generally results into better performances both in terms of input selection and prediction. In particular, as t increases keeping n fixed, the expected agreement between the sparsity pattern of the COSSO estimate with that of the oracle breaks down. The adaptation step helps in consistently selecting the right components. In low noise conditions perfect retrieval is almost always achieved as soon as n is big enough, no matter the degree of concurvity induced by t (Table 8). The presence of noise in the target substantially impact on the performances.

7.3 Real Examples

We applied the ADACOSSO and COSSO on a number of datasets collected from different repositories. The purpose is to illustrate the generalization capability of the adaptation step on real-life scenarios. The first case-study is represented by the NO2 dataset (NO2), in which the task is to study concentration of NO2 versus traffic volume and meteorological variables (7 predictors, 500 observations). Then we considered the Breast Cancer dataset (Breast) and we try to predict the recurrence time, training a model with a small number of patient records (32 features, 47 observations). Another real life example considered is the Census dataset (Census). The goal here is to predict the median price of houses in the region based on demographic composition and the state of housing market (16 variables, 14291 observations). For the Computer activity dataset (CompAct) the task is to estimate the fraction of time that CPUs run in user mode given a collection of computer systems activity measures (21 variables, 8192 records). Finally, we consider the Abalone dataset (Abalone) and we try to predict the age of abalones from physical measurements (4177 observations, 7 variables). All the input data are scaled to the unit interval, which is required to use the smoothing spline ANOVA kernels. We always consider 100 points for training and 300 for testing, with the only exception of Breast. In this case we take 32 points for training and 14 for testing. We run the experiments 200 times and average. Each time the datasets are bootstrapped from the given populations. We report the results in terms of prediction performance and model size (Table 10). We include the performance of the LASSO as baseline. We again use boldface for the best performances and we positively evaluate more parsimonious models. In all the cases the shrinkage parameter is determined via the BIC criterion (23). For linear models $\hat{h}(x) = \langle \hat{w}, x \rangle_{\mathbb{R}^d}$, estimated via the standard LASSO, we set $\varsigma = \mathrm{card}(\hat{w})$ (Zou et al. (2007)).

8. Conclusion

In this work we have elaborated on sparse recovery in function spaces. We have presented an oracle inequality relating $\hat{f}^{\tau}$ with an oracle f, mainly in connections with properties of the model space. In particular, an important role is played by the mutual coherence. The latter can be seen as a precise worst-case characterization of the phenomenon known in the literature as concurvity. Motivated by this oracle result, as a main contribution, we have proposed a data-dependent approach (the data-dependent norm adaptation), with the primary purpose of improving the sparse recovery. The idea basically consists of adapting the

Table 10: Performances on real examples.

	NO2		Breast		Census	
	RSS	size	RSS	size	RSS	size
LASSO	0.56(0.04)	**3.32**(0.5)	**1.06**(0.36)	**8.67**(1.56)	0.72(0.22)	7.66(2.2)
COSSO	0.49(0.05)	4.7(0.93)	2.14(0.82)	22.02(1.83)	0.72(0.24)	4.51(1.72)
ADACOSSO	**0.47**(0.03)	4.23(0.77)	1.51(0.57)	17.78(2.18)	**0.71**(0.23)	**3.93**(1.45)

	CompAct		Abalone	
	RSS	size	RSS	size
LASSO	0.47(0.3)	4.6(1.37)	0.55(0.12)	5.53(1.05)
COSSO	**0.08**(0.16)	3.45(0.99)	0.49(0.07)	4.22(1.2)
ADACOSSO	**0.08**(0.16)	**3.31**(0.81)	**0.48**(0.07)	**3.81**(1.13)

set of norms the component spaces are originally endowed with, to the in-sample dependence structure. To accomplish this task, we have presented an alternating algorithm that is derived from properties of RKHSs. By alternating between a regularized risk functional and a component-wise orthogonal approximation problem, we get what can be seen as an iterative data-dependent improvement on the original set of norms. This correspondingly amounts to the derivation of a a new ensemble of kernel functions, depending upon the input points in the dataset $\mathcal{D}_n$. Finally, we have proposed a modified version of the COSSO algorithm that exploits the new ensemble to perform sparse recovery.

Acknowledgments

The first author thanks Jan C. Willems and Carlo Savorgnan for fruitful discussions.

Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, CIF1, STRT1/08/023, 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.0302.07, G.0320.08, G.0558.08, G.0557.08, G.0588.09, G.0377.09; Research Communities ICCoS, ANMMM and MLDM; the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011), EU: ERNSI. IWT: PhD Grants, McKnow-E, Eureka-Flite+, SBO LeCoPro, SBO Climaqs.

Appendix A. Some Results on Variational Calculus

We recall in this Section some results from variational calculus, mainly adapted to the present setting from Ekeland and Temam (1976) and Ekeland and Turnbull (1983). To our purpose it is assumed that $\mathcal{V}$ is a Hilbert space endowed with inner product $\langle \cdot, \cdot \rangle$. Notice that this assumption results into simpler statements as the original ones. Indeed in this case subgradients of functionals can be considered as elements of the same space and we can avoid the use of duality pairing.

Definition A.1 *Let F be a functional $F : \mathcal{V} \to \mathbb{R}$. If $F(u)$ is finite, any $u_\star$ satisfying:*

$$\langle v - u, u_\star \rangle + F(u) \leq F(v), \quad \forall\ v \in \mathcal{V}$$

is called subgradient *of F at u.*

The set of all subgradients at u is called *subdifferential* at u and denoted by $\partial F(u)$. When $\partial F(u) = \emptyset$ then F is said not to be subdifferentiable at u. When it exists, the quantity:

$$F'(u; v) = \lim_{\lambda \to 0_+} \frac{F(u + \lambda v) - F(u)}{\lambda}$$

is called the directional derivative of F at u in the direction v. If there exists $u_\star$ such that $F'(u; v) = \langle v, u_\star \rangle$ then F is said to be *Gâteaux differentiable* at u and $F'(u) := F'(u; v)$ is called the Gâteaux-differential of F at u. Assume now F to be convex. If F is Gâteaux differentiable at u then it is subdifferentiable at the same point and the subgradient is unique: $\partial F(u) = \{F'(u)\}$. Conversely, if F is continuous at u and finite and has only one subgradient $u_\star$ then it is Gâteaux differentiable at u and $F'(u) = u_\star$. When the subdifferential is a singleton then with some abuse of notation we shall write $\partial F(u)$ to mean $F'(u)$.

To some extent the ordinary differential calculus can be extended into subdifferential calculus. In particular it holds that $\partial(\lambda F)(u) = \lambda \partial F(u)$ and for $F, G : \mathcal{V} \to \mathbb{R}$ we have that $\partial(F + G)(u) \supset \partial F(u) + \partial G(u)$. The inverse inclusion is not always true. If further F, G are convex and there exists $u \in \mathcal{V}$ where F and G are continuous and finite then $\partial(F + G)(u) = \partial F(u) + \partial G(u)$ (see again Ekeland and Temam (1976), Prop. 5.6). We shall also recall explicitly the following result.

Proposition A.1 (Ekeland and Temam (1976), I(Prop. 5.7)) *Let $\mathcal{V}, \mathcal{Y}$ be locally convex spaces. Denote by Λ a continuous linear operator $\Lambda : \mathcal{V} \to \mathcal{Y}$ with adjoint $\Lambda^\star$ and by F a convex lower semi-continuous functional $F : \mathcal{Y} \to \mathbb{R}$, continuous and finite at a point $\Lambda \bar{x}$. Then for any $x \in \mathcal{V}$ we have:*

$$\partial F \bullet \Lambda(x) = \Lambda^\star \left(\partial F(y)|_{y = \Lambda x} \right).$$

Here we are interested in the characterization of a specific problem of the type

$$\min_{v \in \mathcal{V}} F(v)$$

where F is convex. In particular, for F given by the sum of a differentiable functional and a (possibly) non differentiable one, the following holds.

Proposition A.2 (Ekeland and Temam (1976), II(Prop. 2.2)) *Assume that $\mathcal{C} \subseteq \mathcal{V}$ is a non-empty closed convex set and $F : \mathcal{C} \to \mathbb{R}$ is a lower semi-continuous convex function $F = F_1 + F_2$ where F_1 is Gâteaux-differentiable with differential F_1'. Then if $u \in \mathcal{C}$ the following three conditions are equivalent to each other:*

$$u = \arg\min_{v \in \mathcal{C}} F(v) \tag{24}$$

$$\langle F_1'(u), v - u \rangle + F_2(v) - F_2(u) \geq 0, \ \forall v \in \mathcal{C} \tag{25}$$

$$\langle F_1'(v), v - u \rangle + F_2(v) - F_2(u) \geq 0, \ \forall v \in \mathcal{C}. \tag{26}$$

We proceed by giving our definition of Nemitski functional adapted from De Vito et al. (2004) which in turns represents a minor generalization of the one in Ekeland and Turnbull (1983) (III.5.A).

Definition A.2 (De Vito et al. (2004), p. 1388) *Let $\mathcal{Z}$ be a locally compact second countable space, ρ a finite measure on $\mathcal{Z}$ and $W : \mathbb{R} \times \mathcal{Z} \to [0, +\infty[$ be a measurable function such that $W(\cdot, z)$ is convex for any $z \in \mathcal{Z}$. Then the integral map $I_W : L_2(\mathcal{Z}, \rho) \to \mathbb{R}$ defined as*

$$I_W(u) := \int_{\mathcal{Z}} W(u(z), z) d\rho(z)$$

is called Nemitski functional associated to W.

Notice that any closed subset of a Euclidean space is a locally compact second countable space. In particular, which is of interest here, $\mathcal{X} \times \mathcal{Y}$ introduced in Section 2 satisfies the assumption. In general, the Nemitski functional is convex and lower semi-continuous. Besides, we have the following result as part of Lemma 4 in De Vito et al. (2004). The latter is adapted from Ekeland and Turnbull (1983) (III.5, Prop. 1). We rewrite it to fit into the current setting.

Proposition A.3 *Let W be defined as above. If there are $b \in [0, +\infty]$ and $a : \mathcal{Z} \to [0, +\infty]$ such that*

1. *$W(w, z) \leq a(z) + b|w|^2, \ \forall (w, z) \in \mathbb{R} \times \mathcal{Z}$;*
2. *$\int_{\mathcal{Z}} a(z) d\rho(z) < +\infty$;*

then $I_w(u)$ is finite and continuous for any $u \in L_2(\mathcal{Z}, \rho)$.

We conclude by recalling the following result characterizing its subgradient. It is adapted from Ekeland and Turnbull (1983) (III.5, Prop. 3).

Proposition A.4 (De Vito et al. (2004), Prop. 15) *Assume there is an element $u_0 \in L_2(\mathcal{Z}, \rho)$ such that $\sup_{z \in \mathcal{Z}} |u_0(z)| < +\infty$ and $I_W(u_0) < +\infty$. Then for any $u \in L_2(\mathcal{Z}, \rho)$*

$$\partial I_W(u) = \{ s \in L_2(\mathcal{Z}, \rho) \ : \ s(z) \in \partial W(u(z), z) \ \rho - a.e. \} \ .$$

Appendix B. Proof of the Oracle Result

We begin this Section by deriving the closed form for the subgradients used in the proofs of the statements in Section 4. We refer the reader to Appendix A for an account of the results on variational calculus used in the following.

We start by restating C_n in a way that let us easily derive its subgradient. As a first step, we introduce the information operator $I_{\mathcal{D}_n} : \mathcal{H} \to \mathbb{R}^n$ defined as

$$I_{\mathcal{D}_n}(h) = \left(h\left(x^i\right) \ : \ (x^i, y_i) \in \mathcal{D}_n \right) \ .$$

Its adjoint is $I^\star_{\mathcal{D}_n}(c) = \sum c_i k\left(\cdot, x^i\right)$ where k is the kernel function (3). Define now $R_y(z) : \mathbb{R}^n \to \mathbb{R}$ as

$$R_y(z) = \frac{1}{n} \sum_{i \in \mathbb{R}^n} l\left(z_i, y_i\right) \ . \tag{27}$$

With these positions we can finally write

$$C_n(h) = R_y \bullet I_{\mathcal{D}_n}(h). \tag{28}$$

The subgradient of the cost functional C_n is now readily obtained via application of Proposition A.1 as:

$$\partial C_n(h) = I^{\star}_{\mathcal{D}_n}\left(\partial R_y(z)|_{z=I_{\mathcal{D}_n}(h)}\right) = \frac{1}{n}\sum_{i\in\mathbb{N}_n} \partial_u l(u,y_i)|_{u=(I_{\mathcal{D}_n}(h))_i}\, k\left(x^i,\cdot\right) \ . \tag{29}$$

Notice that, if the loss is differentiable in the first argument, as assumed in Section 2, $\partial_u l(u,y)$ stands for the ordinary derivative of l w.r.t. the first argument. In this case the subgradient of C_n at h is unique and thus $\partial C_n(h)$ stands for the Gâteaux differential of C_n at h. In particular for the case of quadratic loss $l(u,y) = \frac{1}{2}(u-y)^2$ the former gives

$$[\partial C_n(h)](x) = \frac{1}{n}\sum_{i\in\mathbb{N}_n} (h(x^i) - y_i)k\left(x^i,x\right) \ .$$

For the derivation of the subgradient of C we make use of results on integral convex functionals reported in Appendix A. We start by reformulating C in terms of the Nemitski functional of l. First of all notice that any $h^j \in \mathcal{H}^j$ can be considered as the function on the product domain $\mathcal{X}\times\mathcal{Y}$ s.t. $(x,y)\mapsto h^j(x)$ which, with some abuse, we still denote by h^j. As now

$$\int_{\mathcal{X}\times\mathcal{Y}} |h^j(x,y)|^2 d\rho(x,y) = \int_{\mathcal{X}} |h^j(x)|^2 d\xi(x) \leq +\infty \ ,$$

it follows that $h = \sum_{j\in\mathbb{N}_m} h^j$ can be included in $L_2(\mathcal{X}\times\mathcal{Y},\rho)$ via a linear, bounded (equivalently, continuos) operator $\mathrm{T}: \mathcal{H} \to L^2(\mathcal{X}\times\mathcal{Y},\rho)$. As above we still denote the image of h via T as h. We also write $k(x,\cdot,:) = \mathrm{T}(k_x)$ where $k(x,y,z) = k(x,y)$ for any $(y,z)\in\mathcal{X}\times\mathcal{Y}$. In a similar fashion, the loss function $l: \mathbb{R}\times\mathcal{Y}$ can be considered a function on $\mathbb{R}\times\mathcal{X}\times\mathcal{Y}$, still denoted by l. We can thus write C in terms of the Nemitski functional of l, $I_l: \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. We have

$$C(h) = I_l \bullet \mathrm{T}(h) = \int_{\mathcal{X}\times\mathcal{Y}} l(h(x,y),x,y)d\rho(x,y) \ .$$

Under the assumption on the loss, Proposition A.3 ensures that $C(h)$ is finite and continuous at any $h\in\mathcal{H}$. As for C_n, we can then apply Proposition A.1 to get $\partial C(h) = \mathrm{T}^{\star}\left(\partial I_l(u)|_{u=\mathrm{T}(h)}\right)$ where ∂I_l is given by Proposition A.4:

$$\partial I_l(u) = \left\{ s\in L_2(\mathcal{X}\times\mathcal{Y},\rho) \ : \ s(x,y)\in \partial_w l(w,x,y)|_{w=u(x,y)} \ \ \rho-a.e.\right\} \ .$$

Finally by the reproducing property and the definition of the adjoint of an operator we have:

$$\begin{aligned}
[\partial C(h)](s) &= \left\langle \mathrm{T}^{\star}\left(\partial I_l(u)|_{u=\mathrm{T}(h)}\right), k_s\right\rangle = \left\langle \partial I_l(u)|_{u=\mathrm{T}(h)}, \mathrm{T}(k_s)\right\rangle_{L^2(\mathcal{X}\times\mathcal{Y},\rho)} \\
&= \left\langle \partial_w l(w,\cdot,:)|_{w=h(\cdot,:)}, \mathrm{T}(k_s)\right\rangle_{L^2(\mathcal{X}\times\mathcal{Y},\rho)} = \int_{\mathcal{X}\times\mathcal{Y}} \partial_w l(w,x,y)|_{w=h(x,y)} k(s,x,y)d\rho(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \partial_w l(w,y)|_{w=h(x)} k(s,x)d\rho(x,y) \ .
\end{aligned}$$

At last, for the penalty we have $\partial J(h) = \sum_j \partial J_j(h)$ with

$$\partial J_j(h) = \begin{cases} \left\{ h^j / \|h^j\|_j \right\} & \text{if } h^j \neq 0 \\ \left\{ u_\star^j \in \mathcal{H}^j \ : \ \|u_\star^j\|_j \leq 1 \right\} & \text{if } h^j = 0 \end{cases} . \tag{30}$$

B.1 Proof of Proposition 4.2

Let $q = \text{card}(\mathcal{A}_f)$. By definition A.1 and equation (30) we have for any $j \in \mathcal{A}_f$:

$$\|f^j\|_j - \|h^j\|_j \leq \frac{1}{\|f^j\|_j} \langle f^j - h^j, f^j \rangle_j = \frac{1}{\|f^j\|_j} \langle f^j - h^j, D^j(f^j) \rangle_{L_2} \leq \frac{\|D^j(f)\|_{L_2}}{\|f^j\|_j} \|f^j - h^j\|_{L_2} . \tag{31}$$

By equation (11) and the Cauchy-Schwarz inequality:

$$\sum_{j \in \mathcal{A}_f} \left(\|f^j\|_j - \|h^j\|_j \right) \leq \sum_{j \in \mathcal{A}_f} \frac{\|D^j(f)\|_{L_2}}{\|f^j\|_j} \|f^j - h^j\|_{L_2} \leq$$

$$(\psi(\mathcal{A}_f))^{1/2} \left(\sum_{j \in \mathcal{A}_f} \|f^j - h^j\|_{L_2}^2 \right)^{1/2} . \tag{32}$$

For any $j \in \mathbb{N}_m$ pose now $g^j = f^j - h^j$. We now prove (12) by showing that:

1.
$$\left(\sum_{j \in \mathcal{A}_f} \|g^j\|_{L_2}^2 \right)^{1/2} (\kappa(\mathcal{A}_f))^{1/2} \leq \left\| \sum_{j \in \mathcal{A}_f} g^j \right\|_{L_2} \tag{33}$$

2.
$$\left\| \sum_{j \in \mathcal{A}_f} g^j \right\|_{L_2} \leq \left(1 - \delta^2(\mathcal{A}_f)\right)^{-1/2} \left\| \sum_{j \in \mathbb{N}_m} g^j \right\|_{L_2} . \tag{34}$$

If all the functions in $\{g^j \ : \ j \in \mathcal{A}_f\}$ are zero the inequality (33) is verified. Assume at least one function in $\{g^j \ : \ j \in \mathcal{A}_f\}$ is not zero. Then:

$$\frac{\left\| \sum_{j \in \mathcal{A}_f} g^j \right\|_{L_2}}{\left(\sum_{j \in \mathcal{A}_f} \|g^j\|_{L_2}^2 \right)^{1/2}} = \left\| \sum_{j \in \mathcal{A}_f} \frac{g^j}{\|g^j\|_{L_2}} \frac{\|g^j\|_{L_2}}{\left(\sum_{j \in \mathcal{A}_f} \|g^j\|_{L_2}^2 \right)^{1/2}} \right\|_{L_2} .$$

For $j \in \mathcal{A}_f$ pose now $\alpha_j = \frac{\|g^j\|_{L_2}}{\left(\sum_{j \in \mathcal{A}_f} \|g^j\|_{L_2}^2 \right)^{1/2}}$ and $h^j = \frac{g^j}{\|g^j\|_{L_2}}$. As $\|\alpha\|_{\mathbb{R}^q} = 1$ and $\{h^j \ : \ j \in \mathcal{A}_f\}$ is a set of functions having L_2 norm equal to 1, equation (33) now follows by Remark

4.1. A proof for the inequality (34) can be found in Koltchinskii (2009). We enclose it for completeness. Let $g_{\mathcal{A}_f} = \sum_{j\in\mathcal{A}_f} g^j \in \mathcal{H}_{\mathcal{A}_f},\ g_{\mathcal{A}_f^c} = \sum_{j\in\mathcal{A}_f^c} g^j \in \mathcal{H}_{\mathcal{A}_f^c}$. Then

$$\begin{aligned}\|g_{\mathcal{A}_f} + g_{\mathcal{A}_f^c}\|_{L_2}^2 &= \|g_{\mathcal{A}_f}\|_{L_2}^2 + \|g_{\mathcal{A}_f^2}\|_{L_2}^2 + 2\langle g_{\mathcal{A}_f}, g_{\mathcal{A}_f^c}\rangle_{L_2} \\ &= \left(1 - \frac{\langle g_{\mathcal{A}_f}, g_{\mathcal{A}_f^c}\rangle_{L_2}^2}{\|g_{\mathcal{A}_f}\|_{L_2}^2\|g_{\mathcal{A}_f^c}\|_{L_2}^2}\right)\|g_{\mathcal{A}_f}\|_{L_2}^2 + \left(\|g_{\mathcal{A}_f}\|_{L_2}\frac{\langle g_{\mathcal{A}_f}, g_{\mathcal{A}_f^c}\rangle_{L_2}}{\|g_{\mathcal{A}_f}\|_{L_2}\|g_{\mathcal{A}_f^c}\|_{L_2}} + \|g_{\mathcal{A}_f^c}\|_{L_2}\right)^2 \geq \\ &\left(1 - \frac{\langle g_{\mathcal{A}_f}, g_{\mathcal{A}_f^c}\rangle_{L_2}^2}{\|g_{\mathcal{A}_f}\|_{L_2}^2\|g_{\mathcal{A}_f^c}\|_{L_2}^2}\right)\|g_{\mathcal{A}_f}\|_{L_2}^2 \geq \left(1 - \delta^2\left(\mathcal{A}_f\right)\right)\|g_{\mathcal{A}_f}\|_{L_2}^2.\end{aligned}$$

Equation (12) now follows.

B.2 Proof of Proposition 4.3

By the Minkowski inequality:

$$\|D(\partial C_n(f))\|_{L_2} \leq \|D(\partial C_n(f)) - D(\partial C(f))\|_{L_2} + \|D(\partial C(f))\|_{L_2}\ . \tag{35}$$

The first member on the right hand side of (35) can be written as:

$$\begin{aligned}\|D(\partial C_n(f)) - D(\partial C(f))\|_{L_2} = \left\|\frac{1}{n}\sum_{i\in\mathbb{N}_n} D\left(v_f(x^i, y_i, \cdot)\right) - D\left(\int v_f(x, y, \cdot)d\rho(x, y)\right)\right\|_{L_2} = \\ \left\|\frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(D\left(v_f(x^i, y_i, \cdot)\right) - \left(\int D\left(v_f(x, y, \cdot)\right)d\rho(x, y)\right)\right)\right\|_{L_2}\ .\end{aligned} \tag{36}$$

By applying the Bennett inequality for vector-valued random variables (see Pinelis (1994) or e.g. Lemma 1 in Smale and Zhou (2007)) we can now upper-bound the second member on the right hand side of (35). Let

$$\sigma^2 := \mathbb{E}\Big[\ \|D(v_f(x, y, \cdot))\|_{L_2}^2\ \Big] = \int \|D(v_f(x, y, \cdot))\|_{L_2}^2\, d\rho(x, y)\ .$$

For any ϵ it holds that:

$$\begin{aligned}\text{Prob}\left\{\left\|\frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(D\left(v_f(x^i, y_i, \cdot)\right) - \left(\int D\left(v_f(x, y, \cdot)\right)d\rho(x, y)\right)\right)\right\|_{L_2} \geq \epsilon\right\} \leq \\ 2\exp\left(-\frac{n\epsilon}{2M}\log\left(1 + \frac{M\epsilon}{\sigma^2}\right)\right)\end{aligned} \tag{37}$$

or equivalently:

$$\begin{aligned}\text{Prob}\left\{\left\|\frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(D\left(v_f(x^i, y_i, \cdot)\right) - \left(\int D\left(v_f(x, y, \cdot)\right)d\rho(x, y)\right)\right)\right\|_{L_2} < \epsilon\right\} > \\ 1 - 2\exp\left(-\frac{n\epsilon}{2M}\log\left(1 + \frac{M\epsilon}{\sigma^2}\right)\right)\ .\end{aligned} \tag{38}$$

Since $\sigma^2 \leq M^2$ it holds that:

$$\text{Prob}\left\{\left\|\frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(D\left(v_f(x^i,y_i,\cdot)\right)-\left(\int D\left(v_f(x,y,\cdot)\right)d\rho(x,y)\right)\right)\right\|_{L_2}<\epsilon\right\}> \\ 1-2\exp\left(-\frac{n\epsilon}{2M}\log\left(1+\frac{\epsilon}{M}\right)\right)\ . \quad (39)$$

If we now take $\epsilon=\frac{M}{\nu}$ the former becomes:

$$\text{Prob}\left\{\left\|\frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(D\left(v_f(x^i,y_i,\cdot)\right)-\left(\int D\left(v_f(x,y,\cdot)\right)d\rho(x,y)\right)\right)\right\|_{L_2}<\frac{M}{\nu}\right\}> \\ 1-2\exp\left(-\frac{n}{2\nu}\log\left(\frac{1+\nu}{\nu}\right)\right)\ . \quad (40)$$

For the second member of (35) we have by application of Jensen's inequality:

$$\|D\left(\partial C(f)\right)\|_{L_2}=\left\|\int D\left(v_f(x,y,\cdot)\right)d\rho(x,y)\right\|_{L_2}\leq\int\|D\left(v_f(x,y,\cdot)\right)\|_{L_2}d\rho(x,y)\leq M\ .$$

It now follows from (35) and the probabilistic bound above that

$$\text{Prob}\left\{\|D\left(\partial C_n(f)\right)\|_{L_2}<\left(\frac{\nu+1}{\nu}\right)M\right\}>1-2\exp\left(-\frac{n}{2\nu}\log\left(\frac{1+\nu}{\nu}\right)\right)\ .$$

B.3 Proof of Theorem 4.4

The functional in (7) is convex and continuous and thus $\partial(C_n(h)+J(h))=\partial C_n(f)+\partial J(h)$. Furthermore under the assumption of the loss, C_n is Gâteaux differentiable. It then follows by definition of $\hat{h}^\tau$ and by Proposition A.2 that:

$$\tau J(\hat{h}^\tau)\leq\tau J(f)+\langle\partial C_n(f),f-\hat{h}^\tau\rangle\ .$$

Now we have:

$$\tau\sum_{j\notin\mathcal{A}_f}\|\hat{h}^{\tau,j}\|_j\leq\tau\sum_{j\in\mathcal{A}_f}\left(\|f^j\|_j-\|\hat{h}^{\tau,j}\|_j\right)+\langle\partial C_n(f),f-\hat{h}^\tau\rangle \quad (41)$$

and it now follows by the definition of the operator D and the Cauchy-Schwarz inequality that:

$$\tau\sum_{j\notin\mathcal{A}_f}\|\hat{h}^{\tau,j}\|_j\leq\tau\sum_{j\in\mathcal{A}_f}\left(\|f^j\|_j-\|\hat{h}^{\tau,j}\|_j\right)+\|D(\partial C_n(f))\|_{L_2}\|f-\hat{h}^\tau\|_{L_2}\ . \quad (42)$$

By Proposition 4.2 we can write:

$$\tau\sum_{j\notin\mathcal{A}_f}\|\hat{h}^{\tau,j}\|_j\leq\left(\tau\sqrt{\frac{\psi(\mathcal{A}_f)}{\kappa(\mathcal{A}_f)\left(1-\delta^2(\mathcal{A}_f)\right)}}+\|D(\partial C_n(f))\|_{L_2}\right)\|f-\hat{h}^\tau\|_{L_2}\ . \quad (43)$$

Finally, by Proposition 4.3 for any $\nu > 0$, $c > 0$ $\tau = cM$ we have, with probability at least $1 - 2\exp\left(-\frac{n}{2\nu}\log\left(\frac{1+\nu}{\nu}\right)\right)$:

$$\sum_{j \notin \mathcal{A}_f} \|\hat{h}^{\tau,j}\|_j \leq \left(\sqrt{\frac{\psi(\mathcal{A}_f)}{\kappa(\mathcal{A}_f)\left(1 - \delta^2(\mathcal{A}_f)\right)}} + \frac{1+\nu}{c\nu} \right) \|f - \hat{h}^{\tau}\|_{L_2} \ . \tag{44}$$

Equation (13) now follows from straightforward calculations.

References

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

F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. *ACM International Conference Proceeding Series*, 2004.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

L. Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37 (4):373–384, 1995.

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Annals of Statistics*, 17(2):453–510, 1989.

E.J. Candès. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15:257–325, 2006.

Z. Chen. Fitting Multivariate Regression Functions by Interaction Spline Models. *J. of the Royal Statistical Society. Series B (Methodological)*, 55(2):473–491, 1993.

J. De Leeuw, F.W. Young, and Y. Takane. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4): 471–503, 1976.

E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100 (5):2197–2202, 2003.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North-Holland Publishing Co., 1976.

I. Ekeland and T. Turnbull. *Infinite-dimensional optimization and convexity*. University of Chicago Press, 1983.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.

G.H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, pages 215–223, 1979.

C. Gu. Diagnostics for nonparametric regression models with additive terms. *Journal of the American Statistical Association*, 87(420):1051–1058, 1992.

C. Gu. *Smoothing spline ANOVA models*. Springer, 2002.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1990.

V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.*, 45(1):7–57, 2009.

V. Koltchinskii and M. Yuan. Sparse Recovery in Large Ensembles of Kernel Machines. COLT, 2008.

G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72, 2004a.

G.R.G. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordan, and W.S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004b.

Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.

L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. Technical report, 2008.

C.A. Micchelli and M. Pontil. Learning the Kernel Function via Regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

C.A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66(2):297–319, 2007.

I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Annals of Probability*, pages 1679–1706, 1994.

A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu. Simple MKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

F. Riesz and B. Sz.-Nagy. *Functional Analysis*. Frederick Ungar Publishing Co., New York, 1955.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, pages 461–464, 1978.

M. Signoretto, K. Pelckmans, and J.A.K. Suykens. Quadratically constrained quadratic programming for subspace selection in kernel regression estimation. In *Proceedings of the 18th international conference on Artificial Neural Networks, Part I*, pages 175–184. Springer, 2008.

V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, volume 22, pages 824–831, 2005.

S. Smale and D.X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

A.K. Smilde, R. Bro, and P. Geladi. *Multi-way analysis with applications in the chemical sciences*. Wiley, 2004.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large Scale Multiple Kernel Learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Verlag, 2008.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Trans. Inform. Theory*, 52:4635–4643, 2006.

G.W. Stewart. Collinearity and Least Squares Regression. *Statistical Science*, 2(1):68–84, 1987.

J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least squares support vector machines*. World Scientific, 2002.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5):2173–2192, 2007.