Functional ANOVA Models: Convex-concave approach and concurvity analysis

Marco Signoretto **Kristiaan Pelckmans** **Johan A.K. Suykens**

K.U. Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Abstract

This paper aims at bridging a gap between functional ANOVA modeling and recent advances in machine learning and kernel-based models. Functional ANOVA on the one hand extends linear ANOVA techniques to nonlinear multivariate models as smoothing splines, aiming to provide interpretability to an estimate and handling the curse of dimensionality. Multiple kernel learning (MKL) on the other hand is well studied in a context of machine learning, Support Vector Machines (SVMs) and kernel-based modeling. Its goal is to combine different sources of information into a powerful predictive model. This paper points to conceptual and computational relations of both frameworks: a first result is the integration of RBF kernels into a formal tensor product space suitable for functional ANOVA models. From a computational perspective, COSSO as proposed in nonparametric statistics is found to fit a framework of convex-concave programming. A main theme is that the occurrence of *concurvity* obstructs the interpretation of the models studied in both frameworks.

1 Introduction

Functional ANOVA models have emerged as a class of structured models useful to capture nonlinear relations in the data, while still providing insight in the model and dealing appropriately with the curse of dimensionality. The general principle behind functional ANOVA models is to approximate the underlying multivariate functional relation by an additive expansion, where components are functions on subsets of covariates that account for both main effects and interactions. Given a multivariate random variable $X = (X_1, \ldots, X_d) \subseteq \mathbb{R}^d$ the decomposition is:

$$f(X) = f_\emptyset + \sum_i f_{\{i\}}(X_i) + \sum_{i>j} f_{\{i,j\}}(X_i, X_j) + \ldots \quad (1)$$

where the components in the expansion are mutually orthogonal in the approximating space. A special case in this framework is represented by the popular additive models $f_\emptyset + \sum_i f_{\{i\}}(X_i)$ discussed in [1], which provide an extension of the standard linear regression model while retaining the non interaction assumption. The tensor product formalism of RKHS provides a convenient way of dealing with the more general case and let one construct approximating spaces which structure can be exploited for the design of algorithms. In particular, a theoretically sound estimator for this type of models was studied under the name of COSSO (Component Selection and Smoothing Operator) in [2] in the framework of smoothing spline ANOVA models. The approach has the advantage of performing at the same time model fitting and model selection, in the sense that it leads to a solution $\hat{f}$ which has a restricted number of non-zero functional terms.

On a different track, recent research in the machine learning community focused on the Multiple Kernel Learning (MKL) framework where one seeks for an optimal combination of kernels for a given task. This subject has attracted a lot of attention and it has been investigated both from a functional viewpoint [3],[4] and from the point of view of optimization [5]. Although conceptually different the two perspectives converge when the selection is tailored to the data since in both cases it corresponds to learn a sparse combination of kernel matrices arising from the observed sample. This framework is appealing especially to combining different data sources. However it is by no means clear how the arising model (1) should be interpreted. One question to ask, for example, is whether the main effect and interactions captured are really representative of the underlying functional relationship or should be regarded suspiciously.

In this work we further elaborate on the connection between the two frameworks. Section 2 begins by briefly sketching the two main paradigms for learning with structured functional models in RKHS's and then present the tensor product device which constitutes a standard tool for constructing spaces of multivariate functions. While traditionally this formalism has been used in combination with Sobolev spaces (Subsection 2.1) we show in Subsection 2.4 how the tensor product space arising from the popular RBF Gaussian kernel constitutes an alternative for the construction of functional ANOVA models. In Section 3 we then recall different characterizations of the phenomenon of *concurvity* which represents the analogous of multicollinearity for standard linear regression and discuss the problem in connection to the model spaces presented. Section 4 considers the optimization problem corresponding to the COSSO objective functional for performing at the same time model fitting and model selection. While in [2] an ad-hoc alternating algorithm was proposed, we illustrate how the solution corresponds to the saddle point of a convex-concave optimization problem in a Euclidean space. We then conclude with Section 5 by presenting some experimental results comparing the COSSO to the convex-concave problem for the estimation of ANOVA models of Gaussian RBF kernel.

2 Structured Models in RKHS's

2.1 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 recall that for a discrete finite set $\mathcal{A}$ its *power set*, denoted by $\mathcal{P}(\mathcal{A})$ or simply with $\mathcal{P}$ when there is no confusion, is the set of all subsets of $\mathcal{A}$, so that e.g. $\mathcal{P}(\mathbb{N}_2) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$. In this letter we deal with product domains $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ (also denoted by $\times_{i \in \mathbb{N}_d} \mathcal{X}_i$) where each set in the Cartesian product is a subset of $\mathbb{R}$ having non-empty interior. We use x to indicate an arbitrary element $(x_j \ : \ j \in \mathbb{N}_d) \in \mathcal{X}$ and we write x_j^i to mean the $j-$th component of $x^i \in \mathcal{X}$. Throughout this work we consider vector spaces of functions defined on $\mathcal{X}$ and taking their values in $\mathbb{R}$. We denote them by $\mathcal{H}(\mathcal{X})$ or sometimes simply by $\mathcal{H}$ for brevity. When $\mathcal{H}(\mathcal{X})$ is endowed with the structure of a Hilbert space defined according to an inner product $\langle \cdot, \cdot \rangle$ we shall denote it by $(\mathcal{H}(\mathcal{X}), \langle \cdot, \cdot \rangle)$. We recall that $(\mathcal{H}(\mathcal{X}), \langle \cdot, \cdot \rangle)$ is a reproducing kernel Hilbert space (RKHS) if for all $x \in \mathcal{X}$ the evaluation functional $L_x : \mathcal{H}(\mathcal{X}) \to \mathbb{R}$ defined by $L_x f = f(x)$ is continuous and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *reproducing kernel* if we have $k(x, \cdot) \in \mathcal{H}(\mathcal{X})$ for any $x \in \mathcal{X}$ and $f(x) = \langle f, k(x, \cdot) \rangle$.

Let now $\mathcal{Y} \subseteq \mathbb{R}$ and $p_{XY} = p_{Y|X} p_X$ be the joint probability density function governing a random vector (X, Y) where $p_{Y|X}$ denotes the *supervisor response*. Further let $\mathcal{D} = \{(X^i, Y^i), \ i \in \mathbb{N}_n\} \subset \mathcal{X} \times \mathcal{Y}$ be an empirical sample drawn i.i.d. from p_{XY}. Then in great generality the problem of learning amounts at choosing in a predefined set of functions the one that minimize the *expected risk* $\int L(f(x), y) p_{XY} dxdy$ where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function suitable for the task at hand. Specifically for regression problems a commonly used loss function is the least square error $L(y, f(x)) = (y - f(x))^2$ and then the problem of learning boils down to the approximation of the *regression function* $\int y p_{Y|x} dy$.This is sometimes referred as *random design* as opposed to *fixed design*. This latter setting is studied in nonparametric regression by smoothing splines [6], [7], where an explicit experimental set-up is assumed and only the outputs are random. Indeed in this latter case one is concern with recovering the true values of a function f at given sites $\{f(x^i) \ : \ (x^i, Y^i) \in \mathcal{D}\}$ from a collection of noisy observations at the same locations $\{Y^i = f(x^i) + \epsilon^i \ : \ (x^i, Y^i) \in \mathcal{D}\}$, where $\{\epsilon^i, \ i \in \mathbb{N}_n\}$ is a set of i.i.d. zero-mean random variables.

In both cases a widely used approach consists in minimizing in a RKHS a regularization functional

$$\frac{1}{n} \sum_{i \in \mathbb{N}_n} L(y^i, f(x^i)) + \mu J(f) \tag{2}$$

where $J : \mathcal{H} \to \mathbb{R}$ is a penalty functional and μ is a positive parameter. When dealing with this problem it is often desirable to use *structured* approximating space $\mathcal{H}(\mathcal{X}) = \oplus_{i \in \mathcal{I}} \mathcal{F}_i(\mathcal{X})$ where $\mathcal{I}$ is a suitable index set and $\{\mathcal{F}_i(\mathcal{X}), \ i \in \mathcal{I}\}$ is a family of subspaces mutually orthogonal w.r.t. $\langle \cdot, \cdot \rangle$. In this way the penalty J can be designed accordingly so that the underlying nature of the functional relationship can be uncovered. This setting is very general: for example functions in the subspaces might refer to different resolutions as in wavelets or might be functions of subsets of variables when dealing with product domains. In this last case the structure allows to circumvent the curse of dimensionality and to select those variables which are more relevant for the task at hand. This is specifically the case we are about to deal with.

2.2 The Tensor Product Formalism and Functional ANOVA Models

The tensor product represents a classical devise for the contruction of functions on a product domain $\mathcal{X}$ [6]. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and consider the Hilbert spaces of functions $(\mathcal{W}_1(\mathcal{X}_1), \langle \cdot, \cdot \rangle_1)$, $(\mathcal{W}_2(\mathcal{X}_2), \langle \cdot, \cdot \rangle_2)$. Let for $i \in \mathbb{N}_2$ and an appropriate index set $\mathcal{I}_i$, $\mathcal{B}_i = \{f_{ij} \ : \ j \in \mathcal{I}_i\}$ denotes a basis for $\mathcal{W}_i(\mathcal{X}_i)$. We define the $2-$fold

tensor product space $(\mathcal{W}_1 \otimes \mathcal{W}_2(\mathcal{X}), \langle \cdot, \cdot \rangle)$ as the completition of the vector space generated by the functions

$$\begin{array}{rccc} f \otimes g & : & \mathcal{X} & \longrightarrow & \mathbb{R} \\ & & x & \longmapsto & f(x_1)g(x_2) \end{array}$$

for any $f \in \mathcal{B}_1$ and $g \in \mathcal{B}_2$, w.r.t. the valid inner product defined for $(f_1, f_2) \in \mathcal{B}_1 \times \mathcal{B}_1$ and $(g_1, g_2) \in \mathcal{B}_2 \times \mathcal{B}_2$ by $\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle = \langle f_1, f_2 \rangle_1 \langle g_1, g_2 \rangle_2$. If further $\mathcal{W}_1(\mathcal{X}_1)$, $\mathcal{W}_2(\mathcal{X}_2)$ are RKHS's with r.k.'s k_1 and k_2 then also the tensor product space is a RKHS with r.k. [8]:

$$\begin{array}{rccc} k_1 \otimes k_2 & : & \mathcal{X} \times \mathcal{X} & \longrightarrow & \mathbb{R} \\ & & (x, y) & \longmapsto & k_1(x_1, y_1)k_2(x_2, y_2). \end{array} \tag{3}$$

A central aspect of the tensor product approach is represented by the following nice property. Let $\{1\}$ denotes the set of constant functions on the appropriate domain and suppose $\mathcal{W}_1 = \{1\} \oplus \mathcal{W}_1^{(1)}$, $\mathcal{W}_2 = \{1\} \oplus \mathcal{W}_2^{(1)}$. Then it can be seen that:

$$\mathcal{W}_1 \otimes \mathcal{W}_2 = \Big(\{1\} \otimes \{1\}\Big) \oplus \Big(\mathcal{W}_1^{(1)} \otimes \{1\}\Big) \oplus \Big(\{1\} \otimes \mathcal{W}_2^{(1)}\Big) \oplus \Big(\mathcal{W}_1^{(1)} \otimes \mathcal{W}_2^{(1)}\Big).$$

Both the definition and the former property of $2-$fold tensor product naturally generalizes to the $d-$fold tensor product of a finite family of *factor spaces* $\{(\mathcal{W}_i(\mathcal{X}_i), \langle \cdot, \cdot \rangle_i) \ : \ i \in \mathbb{N}_d\}$. In this case, assuming each factor space admits a decomposition as for the $2-$fold case, the resulting tensor product space admits a decomposition into 2^d subspaces. Specifically, if for any $u \in \mathcal{P}(\mathbb{N}_d) \setminus \emptyset$ we define $\mathcal{W}_u(\mathcal{X}) = \otimes_{i \in u} \mathcal{W}_i^{(1)}(\mathcal{X}_i)$ and we let $\mathcal{W}_\emptyset(\mathcal{X})$ be the set of constant functions, then we can write

$$\otimes_{i \in \mathbb{N}_d} \mathcal{W}_i(\mathcal{X}_i) = \oplus_{u \in \mathcal{P}(\mathbb{N}_d)} \mathcal{W}_u(\mathcal{X}).$$

Accordingly, any $f \in \otimes_{i \in \mathbb{N}_d} \mathcal{W}_i(\mathcal{X}_i)$ can be written in a unique way as sum of functions in the subspaces: $f = \sum_{u \in \mathcal{P}} f_u$. If the factor spaces $\{\mathcal{W}_i \ : \ i \in \mathbb{N}_d\}$ are RKHS's so are their corresponding subspaces $\{\mathcal{W}_i^{(1)} \ : \ i \in \mathbb{N}_d\}$ and for any $u \in \mathcal{P}$ the r.k. of $\mathcal{W}_u(\mathcal{X})$ is given by the natural generalization of (3).

2.3 Sobolev Spaces on the Unit Interval

For a positive integer l denote by $f^{(l)}$ the $l-$th order derivative of a function f. The $m-$th order Sobolev space $\mathcal{W}([0,1])$ is then defined as [8]:

$$\mathcal{F}_m([0,1]) = \{f : [0,1] \to \mathbb{R}, \text{ with } f, f^{(1)}, \ldots, f^{(m-1)} \text{ absolutely continuous}, \ f^{(m)} \in \mathcal{L}^2([0,1])\}.$$

A classical implementation of the tensor product formalism assumes $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d = [0,1]^d$ and for any $i \in \mathbb{N}_d$ the factor $\mathcal{W}_i(\mathcal{X}_i)$ corresponding to the second order Sobolev space $\mathcal{F}_2([0,1])$. When endowed with the inner product:

$$\langle f, g \rangle_i := \int_0^1 f \int_0^1 g + \int_0^1 f' \int_0^1 g' + \int_0^1 f'' g''$$

$\mathcal{W}_i$ can be decomposed as $\mathcal{W}_i = \{1\} \oplus \mathcal{W}_i^{(1)}$ where any function f in $\mathcal{W}_i^{(1)}$ satisfies the boundary condition $\int_{\mathcal{X}_i} f = 0$. Furthermore $\mathcal{W}_i$ admits a r.k. k_i which can be given in terms of Bernoulli polynomials [6],[7] and correspondingly for any $u \in \mathcal{P}$ $\mathcal{W}_u(\mathcal{X})$ is a RKHS with r.k. given in terms of $\{k_i \ : \ i \in u\}$. The following result arises immediately by the definition of $\mathcal{W}_u(\mathcal{X})$ for $u \in \mathcal{P}$ and by the boundary condition satisfied by the functions in the factor spaces $\{\mathcal{W}_i(\mathcal{X}) \ : \ i \in u\}$.

Proposition 2.1 (Variance decomposition)**.** *Let $X = (X_1, \ldots, X_d)$ be a vector of independent random variables uniformly distributed in $[0,1]$. Then for $\mathcal{A} \subseteq \mathcal{P}(\mathbb{N}_d)$ and any $f = \sum_{u \in \mathcal{A}} f_u \in \oplus_{u \in \mathcal{A}} \mathcal{W}_u(\mathcal{X})$, the random variables $\{f_u(X) \ : \ u \in \mathcal{A}\}$ are mutually uncorrelated and we can write*

$$\mathrm{var}[f(X)] = \sum_{u \in \mathcal{A}} \mathrm{var}[f_u(X)].$$

For random design this gives a statistical characterization of the additive decomposition, thus motivating the name functional ANOVA in connection to the conventional ANOVA framework.

2.4 Functional ANOVA Models with Gaussian RBF Kernels

Radial basis functions (RBF) have been recognized as a powerful tool in approximation. For a product domain $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d \subset \mathbb{R}^d$ endowed with the Euclidean norm $\| \cdot \|_2$ they can all be defined as $k(x, y) = \psi(\|x - y\|_2)$ for certain functions $\psi : \mathbb{R}^+ \to \mathbb{R}$ [9]. In machine learning the most popular is certainly the real-valued Gaussian kernel defined for a width parameter $\sigma > 0$ as:

$$\begin{array}{rccc} k & : & \mathcal{X} \times \mathcal{X} & \longrightarrow & \mathbb{R} \\ & & (x, y) & \longmapsto & \exp(-\sigma^2 \|x - y\|_2^2) \end{array} \tag{4}$$

which has been shown to be *universal* in the sense that any reasonable function on $\mathcal{X}$ can be approximated arbitrarily well by a function in the RKHS associated with (4) [10] . An explicit description of such a space was given in [11] in terms of the RKHS of a complex-valued analogous of (4) defined on $\mathbb{C}^d \times \mathbb{C}^d$. Recall that, for an index set $\mathcal{I}$, the space of real-valued square-summable sequences $l_2(\mathcal{I})$ is $l_2(\mathcal{I}) = \{(a_i \in \mathbb{R} \ : \ i \in \mathcal{I}) \text{ s.t. } \|a\|_{l_2}^2 = \sum_{i \in \mathcal{I}} a_i^2 \leq \infty\}$ which is an Hilbert space with $\langle a, b \rangle_{l_2} = \sum_{i \in \mathcal{I}} a_i b_i$ and let $\mathbb{N}_0 =$

$\{0\}\cup\mathbb{N}$. Then it follows from the arguments in [11] that for any $j \in \mathbb{N}_d$ the RKHS $(\mathcal{W}_j(\mathcal{X}_j), \langle\cdot,\cdot\rangle_j)$ associated with the univariate RBF Gaussian kernel $k : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$, $k(x_j, y_j) = \exp(-\sigma^2(x_j - y_j)^2)$ has orthonormal basis $\left\{ e_n(x_j) = \sqrt{\frac{(2\sigma^2)^n}{n!}} x_j^n e^{-\sigma^2 x_j^2} \ : \ x_j \in \mathcal{X}_j,\ n \in \mathbb{N}_0 \right\}$ and we have $\mathcal{W}_j(\mathcal{X}_j) = \{f \ : \ f = \sum_{i\in\mathbb{N}_0} a_i e_i \text{ s.t. } a \in l_2(\mathbb{N}_0)\}$ and for any pairs $(a, b) \in l_2(\mathbb{N}_0) \times l_2(\mathbb{N}_0)$ and corresponding functions $(f, g) \in \mathcal{W}_j(\mathcal{X}_j) \times \mathcal{W}_j(\mathcal{X}_j)$ $\langle f, g\rangle = \langle a, b\rangle_{l_2}$. Furthermore by Corollary 3.9 in [11] any non identically zero constant function is not in $\mathcal{W}_j(\mathcal{X}_j)$. The set $\{1\}$ can thus be adjoined and following the argument in Subsection 2.2 we can formally state a tensor product decomposition:

$$\otimes_{i\in\mathbb{N}_d} (\mathcal{W}_i(\mathcal{X}_i) \oplus \{1\}) = \oplus_{u\in\mathcal{P}(\mathbb{N}_d)} \mathcal{W}_u(\mathcal{X})$$

where for any $u \in \mathcal{P}$ $\mathcal{W}_u(\mathcal{X})$ is a RKHS with r.k. $k_u(x, y) = \exp\left(-\sigma^2 \sum_{j\in u}(x_j - y_j)^2\right)$. The construction remains valid if factor spaces are associated with Gaussian kernels with different width parameters. However for the applications this is rather unpractical as it results in a more involved tuning procedure.

3 The Problem of Concurvity

In the context of multivariate linear regression the problem of multicollinearity, namely the linear dependency between predictor variables, has been extensively studied both by numerical analysts and statisticians [12], and is associated with inflationary effects on standard errors of regression coefficients. The term *concurvity* has been used to describe the multicollinearity of nonlinear transforms of predictors arising in the context of additive models [13]. In the following subsections we assume an additive model with components indexed by $\mathcal{A} \subseteq \mathcal{P}(\mathbb{N}_d)$ and for $u \in \mathcal{A}$ we denote by f_u the component defined on $\mathcal{X}$ which only depends on input variables indexed by elements of u.

3.1 Prospective Concurvity

From a population perspective, given an arbitrary set of functions $\mathcal{W}(\mathcal{X})$ and a $\mathcal{X}-$valued random vector $X = (X_1, \ldots, X_d)$, concurvity was formalized in [13]. It corresponds to the existence of a family of functions $\{f_u \ : \ u \in \mathcal{A}\}$ in $\mathcal{W}(\mathcal{X})$ and a certain set $\{\lambda_u \in \mathbb{R}\backslash\{0\} \ : \ u \in \mathcal{A}\}$ such that $\sum_{u\in\mathcal{A}} \lambda_u f_u(X) = 0$. In a similar spirit *prospective concurvity* was used in [14] to indicate a property of an arbitrary space of functions, this time with reference to their restriction at fixed design points $\{x^i \ : \ (x^i, Y^i) \in \mathcal{D}\} \subset \mathcal{X}$. Specifically it is used to indicate the existence of a family of functions in the set such that the vectors in $\{f_u^x = (f_u(x^i) \ : \ (x^i, Y^i) \in \mathcal{D}),\ u \in \mathcal{A}\}$ are collinear. In this case concurvity corresponds to the fact that the decomposition might be not well defined at the design sites in the sense that once we have an estimate $\hat{f} = \sum_{u\in\mathcal{A}} \hat{f}_u$ computed on the base of the fixed points and the corresponding outputs, $\sum_{u\in\mathcal{A}} \hat{f}_u^x$ might be equivalent to $\sum_{u\in\mathcal{A}} \lambda_u \hat{f}_u^x$ for nontrivial coefficients $\{\lambda_u \neq 1 \ : \ u \in \mathcal{A}\}$ so that we could have used the estimate $\hat{g} = \sum_{u\in\mathcal{A}} \lambda_u \hat{f}_u$ instead, at least if $J(\hat{g}) \leq J(\hat{f})$.

3.2 Model Diagnostic: Observed Concurvity

An additional characterization of the problem directly pertains the fitted model and leads to the definition of diagnostic tools. *Observed concurvity* is defined as the collinearity of $\{\hat{f}_u^x = (\hat{f}_u(x^i) \ : \ (x^i, Y^i) \in \mathcal{D}),\ u \in \mathcal{A}\}$ i.e. of vectors representing the restriction of the components of the fitted model $\hat{f} = \sum_{u\in\mathcal{A}} \hat{f}_u$ w.r.t. the points $\{x^i \ : \ (x^i, Y^i) \in \mathcal{D}\} \subset \mathcal{X}$. Originally proposed in a fixed design setting [14] it can also considered in random design with input sites corresponding to i.i.d. random variables. Once defined $Y^x = (Y^i \ : \ (X^i, Y^i) \in \mathcal{D})$ we may write:

$$Y^x = \sum_{u\in\mathcal{A}} \hat{f}_u^x + \epsilon \tag{5}$$

where ϵ denotes the error vector. In real life application an analysis based on the retrospective linear model (5) can be conducted via the collinearity indices [12] and can be used to assess the validity of the estimated function $\hat{f}$. Let $A^\dagger$ denotes the pseudoinverse of a matrix A. For $\mathcal{A} = \{u_1, \ldots, u_m\}$ and $F = (f_{u_1}^x, \ldots, f_{u_m}^x)$ the $j-$th collinearity index of F is defined as $r_j = \|F_j\|_2 \|F_j^\dagger\|_2$, where F_j is the j-th column of F and $F_j^\dagger$ is the $j-$th row of the corresponding pseudo inverse. The index r_j is always greater or equal than one with equality if and only if F_j is orthogonal to the other columns of F. With i.i.d. input data orthogonality then asymptotically translates into statistical uncorrelation and if the indices are all close to one then the model is able to isolate main effects and interaction terms.

3.3 Concurvity and Tensor Product Spaces

We stress here a straightforward consequence of Proposition 2.1. If the input vectors are i.i.d. from a uniform distribution in $[0, 1]^d$ and the tensor product of Subsection 2.1 is used as the approximating space, then $\{\hat{f}_u^x \ : \ u \in \mathcal{A}\}$ must be asymptotically uncorrelated and, since components in the additive decomposition are not *confounded* with other terms, the effect captured by each one can be regarded as a reliable representation of the corresponding main effect or interaction with undisputed benefit for the interpretation of the model. Unfortunately it seems a result

of this kind cannot be derived for the tensor product decomposition with RBF Gaussian kernels and thus the construction of Subsection 2.4 can be regarded as "merely" functional. In any case how to deal and especially avoid concurvity in the general non uniform case seems to remain an open issue. In real-life applications concurvity arises especially when the input data exhibit a strong association. Indeed in this case the fitted components are also generally associated and the collinearity indices together with the simple cross-correlation can be used to detect if this occurs, as we shall illustrate in a simple toy example in Subsection 5.1. Should this happen the interpretation of the recovered structure has to be conducted very carefully. However even in this case the additive structure of the model can still be useful to circumvent the curse of dimensionality and to access the relative importance of the single components.

4 Structural Learning from Empirical Data

We discuss at this point model selection and fitting in the framework of functional ANOVA models via convex-concave optimization. Our goal is to approximate the regression function by an ANOVA model on the base of $\mathcal{D}$. Let $(\mathcal{W}(\mathcal{X}), \langle\cdot,\cdot\rangle)$ denotes a space of structured models and more specifically, for $\mathcal{A} \subseteq \mathcal{P}\setminus\emptyset$, satisfying $\mathcal{W}(\mathcal{X}) = \mathcal{W}_\emptyset(\mathcal{X}) \oplus (\oplus_{u\in\mathcal{A}}\mathcal{W}_u(\mathcal{X}))$ in the sense of Section 2. Furthermore for any $u \in \mathcal{A}$ let P^u denotes the orthogonal projection onto $\mathcal{W}_u(\mathcal{X})$ and let $\|\cdot\|$ be the norm corresponding to $\langle\cdot,\cdot\rangle$.

4.1 The COSSO Functional for Structural Learning

Different approaches can be pursued for the estimation of the model and many of them amount at minimizing a functional like (2). Here we consider the minimization in $\mathcal{W}(\mathcal{X})$ of the regularized functional with pseudo-norm penalty $J(f) = \sum_{u\in\mathcal{A}} \|P^u f\|$:

$$A_\tau(f) = \frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(y^i - f(x^i)\right)^2 + \tau J(f) \qquad (6)$$

which was originally studied in [2] under the name of COmponent Selection and Smoothing Operator (COSSO) in connection to smoothing splines ANOVA models. For a given value of the shrinkage parameter $\tau > 0$, this approach has the advantage of performing at the same time model fitting and model selection, in the sense that it leads to a solution $\hat{f} = \sum_{u\in\mathcal{A}}\hat{f}_u$ which has a restricted number of non-zero components. In turn, the amount of sparsity can be trimmed by an appropriate setting of the shrinkage parameter. The minimization of (6) represents a generalization of the optimization problem in the LASSO [15] to which it was shown to correspond when the learning process takes place in a space of affine functions. In [2] it was proved that a solution of (6) exists and that for

$$M_{\lambda_0}(\theta, f) = \frac{1}{n}\sum_{i\in\mathbb{N}_n}\left(y^i - f(x^i)\right)^2 + \lambda_0\sum_{u\in\mathcal{A}}\theta_u^{-1}\|P^u f\|^2 \qquad (7)$$

the functional

$$B_{\lambda,\lambda_0}(\theta, f) = M_{\lambda_0}(\theta, f) + \lambda\sum_{u\in\mathcal{A}}\theta_u \qquad (8)$$

is equivalent to (6) in the following sense.

Lemma 4.1 ([2], Lemma 2)**.** *If for arbitrary $\tau > 0$ $\hat{f}$ minimizes $A_\tau(f)$ in $\mathcal{F}(\mathcal{X})$ then for any pair (λ, λ_0) s.t. $\lambda\lambda_0 = \tau^2/4$, once set, for any $u \in \mathcal{A}$, $\hat{\theta}_u = \lambda^{-1/2}{\lambda_0}^{1/2}\|P^u\hat{f}\|$, $(\hat{\theta}, \hat{f})$ minimizes $B_{\lambda,\lambda_0}(\theta, f)$ in $\mathbb{R}_+^{|\mathcal{A}|}\times\mathcal{F}(\mathcal{X})$ where $\mathbb{R}_+^{|\mathcal{A}|} := \{x \in \mathbb{R}^{|\mathcal{A}|} : x_i \geq 0,\ i \in \mathbb{N}_{|\mathcal{A}|}\}$. On the other hand, if for arbitrary (λ, λ_0) a pair $(\hat{\theta}, \hat{f})$ minimizes $B_{\lambda,\lambda_0}(\theta, f)$ in $\mathbb{R}_+^{|\mathcal{A}|}\times\mathcal{F}(\mathcal{X})$ then $\hat{f}$ minimizes $A_\tau(f)$ in $\mathcal{F}(\mathcal{X})$ for any τ s.t. $\tau = 2\lambda^{1/2}{\lambda_0}^{1/2}$.*

This result can be exploited to provide effective algorithms for computing a minimizer of the original functional. While in [2] it was proposed to find a solution of (8) by alternating between the optimization of θ and that of f, here we illustrate a different approach inspired by the multiple kernel learning framework to which the functional (6) is strictly related [4].

4.2 Convex-concave Formulation

For appropriate choice of $\rho > 0$ the minimization of $B_{\lambda,\lambda_0}(\theta, f)$ in $\mathbb{R}_+^{|\mathcal{A}|}\times\mathcal{F}(\mathcal{X})$ is equivalent to the minimization of $M_{\lambda_0}(\theta, f)$ in $\Theta\times\mathcal{F}(\mathcal{X})$ where $\Theta = \{\theta \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_{u\in\mathcal{A}}\theta_u \leq \rho\}$. For any fixed $\theta \in \Theta$ the minimizer of $M_{\lambda_0}(\theta, f)$ in $\mathcal{F}(\mathcal{X})$ takes the form $\hat{f}(\cdot) = \sum_{i\in\mathbb{N}_n}\hat{\alpha}_i k_\theta(x_i, \cdot) + \hat{b}$ where $k_\theta = \sum_{u\in\mathcal{A}}\hat{\theta}_u k_u$ and $\hat{\alpha}$ and $\hat{b}$ are the solution of the system of linear equations [6]:

$$\left(\sum_{u\in\mathcal{A}}\theta_u K^{(u)} + \lambda_0 n I\right)\alpha + 1b = y \qquad (9)$$

$$1^\top\alpha = 0 \qquad (10)$$

where for any $u \in \mathcal{A}$, $K^{(u)} = (k_u(x_i, x_j) : (i,j) \in \mathbb{N}_n\times\mathbb{N}_n)$ and I, 1 denote respectively the identity matrix and the vector of ones of proper dimensions. In turn (9) corresponds to the K.K.T. optimality conditions of the constrained maximization problem:

$$\begin{array}{ll}\max\limits_{\alpha\in\mathbb{R}^N} & \alpha^\top y - \frac{1}{2}\alpha^\top\left(\sum_{u\in\mathcal{A}}\theta_u K^{(u)} + \lambda_0 n I\right)\alpha \\ \text{s.t.} & 1^\top\alpha = 0\ .\end{array} \qquad (11)$$

Indeed, denoted respectively by b the dual variable of the equality constraint and by $L(\alpha, b)$ the Lagrangian for the former maximization problem, then (9) corresponds to $\partial L(\alpha, b)/d\alpha = 0$ and (10) is the qualification for the constraint. For appropriate choice of κ the solution of (8) can now be computed via :

$$\begin{aligned} \min_{\theta \in \mathbb{R}_+^{|\mathcal{A}|}} \max_{\alpha \in \mathbb{R}^n} \quad & G_{\lambda_0,\kappa}(\theta, \alpha) := \alpha^\top y - \tfrac{1}{2}\alpha^\top \textstyle\sum_{u \in \mathcal{A}} \theta_u K^{(u)} \alpha \\ & -1/2\lambda_0 n \|\alpha\|_2^2 + \kappa \|\theta\|_1 \\ \text{s.t.} \quad & 1^\top \alpha = 0 \end{aligned} \tag{12}$$

which can be shown to be a convex-concave problem [16] with basically the same arguments as in [17]. As a consequence, any solution $(\hat{\theta},\ \hat{\alpha})$ satisfies for every feasible point the saddle-point property:

$$G_{\lambda_0,\kappa}(\hat{\theta}, \alpha) \leq G_{\lambda_0,\kappa}(\hat{\theta}, \hat{\alpha}) \leq G_{\lambda_0,\kappa}(\theta, \hat{\alpha})$$

and can be computed with general interior point solvers by resorting to an equivalent convex formulation in the class of second order cone programming (SOCP) problems [18]. A more direct approach we have been using corresponds to a log barrier algorithm which directly exploit the convex-concave structure via an appropriate modification of an infeasible-start Newton method [16].

4.3 Relative Importance and Convex-concave Structure

The intrinsic convex-concave nature of the problem leads to the following nice interpretation of the model fitted solving (12). It can be shown via optimality conditions that the set of selected components $\hat{\mathcal{A}} = \{u\ :\ \hat{\theta}_u > 0\}$ is such that

$$\hat{\mathcal{A}} = \arg\max_{u \in \mathcal{A}} \hat{\alpha}^\top K^{(u)} \hat{\alpha}$$

and that $\max_{u \in \mathcal{A}} \hat{\alpha}^\top K^{(u)} \hat{\alpha} = 2\kappa$. For any u in the index set $\hat{\mathcal{A}}$ of selected components, the norm is

$$\|P^u \hat{f}\| = \hat{\theta}_u \sqrt{\hat{\alpha}^\top K^{(u)} \hat{\alpha}}. \tag{13}$$

This norm can be used as a measure of relevance. Since the right hand side corresponds to $\hat{\theta}_u \sqrt{2\kappa}$ for any $u \in \hat{\mathcal{A}}$, $\hat{\theta}_u$ can be used to rank the components. This relates to the LASSO [15] and extends the common practice in linear modelling, where the absolute value of the elements in the parameter vector is often taken as an indicator of relative importance.

4.4 Parameters Invariance

As follows from Lemma 4.1 the problem has essentially one parameter so that, when searching for solution with different degrees of sparsity, either λ or λ_0 can be kept fixed in the objective (8). This property carries on in the optimization problem (12). To see this set $\alpha > 0$ and let $\tilde{\alpha} = a\alpha$ and $\tilde{\theta} = 1/a\theta$. Then $\sum_{i \in \mathbb{N}_n} \alpha_i \left(\sum_{u \in \mathcal{A}} \theta_u k_u(x_i, \cdot)\right) + b = \sum_{i \in \mathbb{N}_n} \tilde{\alpha}_i \left(\sum_{u \in \mathcal{A}} \tilde{\theta}_u k_u(x_i, \cdot)\right) + b$ that is, $(\tilde{\alpha}, \tilde{\theta}, b)$ and (α, θ, b) lead to the same function. Now with the position $\epsilon = y - 1/2 \sum_{u \in \mathcal{A}} \theta_u K^{(u)} \alpha$ we have:

$$\begin{aligned} G_{\lambda_0,\kappa}(\theta, \alpha) = \alpha^\top \epsilon - \frac{\lambda_0 n}{2} \|\alpha_2\|_2^2 + \kappa \|\theta\|_1 = \frac{1}{a} \tilde{\alpha}^\top \epsilon - \\ \frac{\lambda_0 n}{2a^2} \|\tilde{\alpha}\|_2^2 + \kappa\, a \|\tilde{\theta}\|_1 = a G_{\lambda_0/a, \kappa a^2}(\tilde{\theta}, \tilde{\alpha}). \end{aligned} \tag{14}$$

The following proposition in then a straightforward consequence.

Proposition 4.2. *Let $(\hat{\theta}, \hat{\alpha})$ satisfies the saddle point property for a pair (λ_0, κ). Then $(\tilde{\theta}, \tilde{\alpha})$ such that $\tilde{\alpha} = a\,\hat{\alpha}$ and $\tilde{\theta} = 1/a\,\hat{\theta}$ satisfies the property for $(\lambda_0/a, \kappa a^2)$ and $\tilde{f}_{(\lambda_0/a, \kappa a^2)} = \hat{f}_{(\lambda_0, \kappa)}$.*

This practically means that the problem can be reparametrized according to a unique parameter (e.g. $\kappa\lambda_0^2$) and for each value of it a corresponding pair (λ_0, κ) can be chosen so that the algorithm is numerically well conditioned. In practice we observed a speed-up of convergence of a factor 5 for appropriately chosen pairs. By Lemma 4.1 we also have for any $u \in \mathcal{A}$ $\hat{\theta}_u = \lambda^{-1/2} {\lambda_0}^{1/2} \|P^u \hat{f}\| = \lambda^{-1/2} {\lambda_0}^{1/2} \hat{\theta}_u \sqrt{\hat{\alpha}^\top K^{(u)} \hat{\alpha}}$ and thus $\lambda_0 = \lambda/(2\kappa)$ which gives the relationship between λ in (8) and κ in (12).

5 Experimental Results

5.1 Toy Problem

We illustrate here the problem of (observed) concurvity and its detection with a simple 2-D toy problem. We begin by considering 400 points drawn at random from the uniform distribution in $[0, 1]^2$ and generate the output measurements starting from univariate building blocks:

$$\begin{aligned} & g_1(x) = \sin(2\pi x) \quad g_2(x) = \exp(x) \quad g_3(x) = \cos(4\pi x) \\ & Y^i = g_1(X_1^i) + 1.4 g_2(X_2^i) + g_3(X_1^i - X_2^i) + 3 \end{aligned} \tag{15}$$

where $\{\epsilon^i\ :\ i \in \mathbb{N}_{400}\}$ are i.i.d. zero-mean Gaussian random variables with variance chosen so that the signal-to-noise ratio is 2.77. We fit the models by solving (12) where an appropriate value for the pair (λ_0, κ) is chosen via 5−fold CV by searching in a one dimensional parameter space (cf. Subsection 4.4). We consider the full interaction models (with the notation above $\mathcal{A} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$) for the tensor product space presented in Section 2. As expected the model doesn't suffer from concurvity in this uniform case as

shown by the collinearity indices $\{r_i : i \in \mathbb{N}_3\}$ of Subsection 3.2 and the correlation coefficients reported in the upper panel of table 1. In comparison the ANOVA model with the RBF kernels already in this case exhibits association among the fitted components, as illustrated in table 2.

Table 1: Concurvity in the uniform case and with dependand data: collinearity indices and correlation coefficients.

	uniform		
	$f_{\{1\}}$	$f_{\{2\}}$	$f_{\{1,2\}}$
k_i	1	1	1
$f_{\{1\}}$	1	0.01	0.02
$f_{\{2\}}$	0.01	1	0.01
$f_{\{1,2\}}$	0.02	0.01	1

	non-uniform		
	$f_{\{1\}}$	$f_{\{2\}}$	$f_{\{1,2\}}$
k_i	1.65	1.48	1.87
$f_{\{1\}}$	1	0.66	-0.88
$f_{\{2\}}$	0.66	1	-0.93
$f_{\{1,2\}}$	-0.88	-0.93	1

Table 2: Concurvity in the uniform case for the ANOVA model with RBF kernels.

	RBF uniform		
	$f_{\{1\}}$	$f_{\{2\}}$	$f_{\{1,2\}}$
k_i	1.38	1.21	1.55
$f_{\{1\}}$	1	0.03	0.64
$f_{\{2\}}$	0.03	1	0.43
$f_{\{1,2\}}$	0.64	0.43	1

In order to illustrate the effect of the association among the input data on the fitted components, we then generate dependent points in the unit-box drawn from a Gaussian copula with uniform marginals and correlation matrix $\begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$. We leave the rest of the set-up unchanged. Concurvity now affects the model: especially the main effects $f_{\{1\}}$, $f_{\{2\}}$ are both confounded with the interaction term, as detected by the high correlation coefficient in the lower panel of table 1. In figure 1 we plot the underlying surface along with the noisy measurements and the contour lines of the input pdf for the dependent case (1(a)); the components of estimated main effects with the data generated via the Gaussian copula, as an illustration of concurvity (1(b)); the estimated main effects in the uniform case which recover the generating functions (1(c)). In this case we don't recover the mechanism generating the data, cf eq. (15).

5.2 Real Life Data-set

In this section we consider the application of the illustrated procedures on the Concrete Compressive Strength dataset publicly available at the UCI repository. The task is that of predicting the compressive strength of high-performance concrete given the content of different compounds in the mixture. Whenever

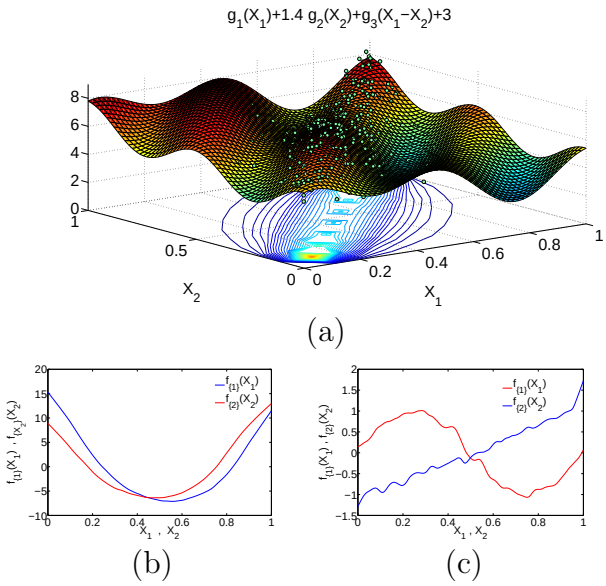

Figure 1: 1(a) The underlying function with the noisy data and the input pdf for the non uniform case. 1(b) Concurvity arises when input data have strong association as can be clearly visualized in this case by plotting the main effects on the same axis. 1(c) The fitted main effects for the uniform case, again plotted on the same axis, capture in this case the shape of the underlying components (15).

the tensor product of Sobolev spaces is used we rescale the input data to the unit interval. For each procedure tested the observations in the datasets 400 points are used for training and rest for testing. The tuning parameters in each procedure are selected by 5−fold cross validation. We then use the full training set to estimate $\hat{f}$ with the optimal set of parameter and evaluate the performance on the test set. The whole procedure is repeated 5 times. We use this procedure to compare a) the model estimated by the publicly available MATLAB code for the algorithm presented in [2] (COSSO) b) a log-barrier algorithm for the solution of the min-max problem (12) with the two-way interaction model in the tensor decomposition of Subsection 2.1 (MINMAX-SS) c) the former procedure used in combination with the two-way interaction model with RBF's (MINMAX-RBF) d) a least-square SVM [19] for unstructured learning with a unique RBF kernel (LS-SVM). The average MSE together with the standard deviation is reported in table 3. The different algorithms show comparable results except from the RBF ANOVA model which performs worse than the others. The LS-SVM unstructured model shows already good prediction performances. However it does not allow to capture the relative importance of the components. This can be done e.g. by comparing the

Table 3: Prediction performances.

Components	avg MSE (std dev)
COSSO	0.14 (0.01)
MINMAX-SS	0.13 (0.01)
MINMAX-RBF	0.2 (0.02)
LS-SVM	0.14 (.02)

optimal weights $\{\hat{\theta}_u \ : \ u \in \hat{\mathcal{A}}\}$ which corresponds to the most prominent components w.r.t. the norm of the function space (cf. Subsection 4.3). The reader is referred to the UCI repository for the meaning of the variables of which we preserved the index number. The first 5 components ranked according to this criterion are reported in table 4 which presents a) the associated (scaled) $\hat{\theta}_u$ b) the correlation matrix showing the cross-association c) the collinearity indices illustrating how much each f_u^x can be explained by all the rest of components in the model $\{f_s^x \ : \ s \in \hat{\mathcal{A}} \setminus u\}$ (cf the retrospective linear model (5)).

Table 4: Model Diagnostic

	$\hat{\theta}_u$	$\{8\}$	$\{7\}$	$\{1,2\}$	$\{3,6\}$	$\{7,8\}$
k_u		3.83	6.4	2.59	2.19	2.95
u						
$\{8\}$	1	1	0.17	-0.15	0	0.17
$\{7\}$	0.57	0.17	1	-0.43	-0.24	0.1
$\{1,2\}$	0.34	-0.15	-0.43	1	0.39	-0.1
$\{3,6\}$	0.28	0	-0.24	0.39	1	-0.18
$\{7,8\}$	0.26	0.17	0.1	-0.1	-0.19	1

Acknowledgements

Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM, several PhD/postdoc - fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), 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; FP7-HD-MPC (Collaborative Project STREP-grantnr. 223854) Contract Research: AMINAL Other: Helmholtz: viCERP

References

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

[2] Y. Lin and H.H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5), 2006.

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

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

[5] 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.

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

[7] C. Gu. *Smoothing spline ANOVA models*. Series in Statistics, 2002.

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

[9] R. Schaback and H. Wendland. Characterization and construction of radial basis functions. *Multivariate Approximation and Applications*, 2001.

[10] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93, 2002.

[11] 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.

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

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

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

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

[16] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[18] M. Signoretto, K. Pelckmans, and J.A.K. Suykens. Quadratically constrained quadratic programming for subspace selection in kernel regression estimation. 2008. accepted for publication in Artificial Neural Networks - ICANN.

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