A two stage algorithm for kernel based partially linear modeling with orthogonality constraints

Tillmann Falck, Marco Signoretto, Johan A.K. Suykens, Bart De Moor

Abstract

It is shown that for existing kernel based partially linear modeling approaches, the parametric model part estimate depends on the kernel function. A novel orthogonality constraint ensures that this dependency is removed and that the estimate for the parametric part is as good as the ordinary least squares estimate. The results are presented in a optimization setting with primal and dual model representations as well as in a functional setting.

Index Terms

system identification, basis functions, kernel methods, SVMs, LS-SVMs, RKHSs, partially linear models, semi-parametric models

I. Introduction

Partially linear models [1] as shown in Figure 1 are a special case of semi-parametric models that combine a parametric model with a nonparametric one. In partially linear models the parametric part is linear in the parameters. Their usage is beneficial if a model (class) for some explanatory variables is known and a nonparametric model shall be used to improve the prediction performance.

Especially in system identification [2] for nonlinear systems, good parametric models can be hard to obtain. Yet the use of black box models [3] may have a higher computational cost and

T. Falck, M. Signoretto, J. Suykens and B. De Moor are with the SCD/SISTA group at the department of Electrical Engineering (ESAT) at the Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. E-mail: {tillmann.falck,marco.signoretto, johan.suykens,bart.demoor}@esat.kuleuven.be.

Research supported by Research Council KUL: PhD/postdoc grants, GOA/10/09 MaNet, CoE EF/05/006 OPTEC, several PhD/postdoc & fellow grants; Flemish Government: FWO G.0302.07 (SVM/Kernel), PhD/postdoc grants by FWO and IWT; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO); EU: ERNSI. Johan Suykens is a professor and Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium.

more importantly is mostly much harder to interpret. Also prior knowledge embedded in the choice of the parametric model is lost when switching to a nonparametric model which can yield a worse prediction performance and therefore defeat the purpose of using a black box model.

There is a wealth of literature for kernel based partially linear models [4], [5], [6], [7], [8], [9]. However in [4], [6], [7] it is assumed that the parametric and nonparametric function spaces are disjoint. Therefore it can be difficult to find a suitable nonparametric model for an existing and fixed parametric model. If there is an admissible combination, finite sample sizes give rise to problems similar to those when the assumptions are violated. Therefore even if the assumptions are satisfied, the estimate of the parametric model obtained with existing techniques can be substantially worse than the ordinary least squares (OLS) solution. In this paper we will address the implications of relaxing those assumptions and propose a modified estimation scheme to overcome these. The modification is based on an additional constraint that ensures that the estimates generated by the parametric and the nonparametric models are empirically orthogonal (uncorrelated). We also show how this relates to function spaces.

In this work Least Squares Support Vector Machines (LS-SVMs) [10] and Support Vector Machines (SVMs) [11] are used for the nonparametric part. SVMs and LS-SVMs are black box regression techniques based on convex optimization formulations, which allow to incorporate additional constraints straightforwardly [12]. Alternative kernel based regression techniques are e.g. splines [6] and regularization networks [13], which are formulated in Reproducing Kernel Hilbert Spaces (RKHSs).

The paper is organized as follows. In Section II we briefly review existing work on kernel based partially linear models, with an emphasis on the necessary assumptions. The effect of violating these assumptions is analyzed in III. We then propose a modification to overcome this effect and derive the corresponding kernel based model in the same section. In Section IV a separable estimation problem is presented and an equivalent kernel is derived. The next section derives equivalent results in a RKHS setting. The results based on LS-SVMs are then extended to different loss functions in the Section VI, where we integrate the idea into SVMs [11]. In Section VII we illustrate the effect of our modification on simulated as well as real world data.

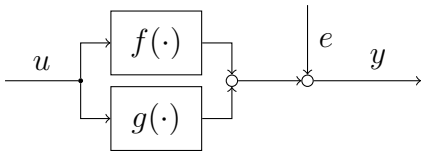


Fig. 1. Partially linear model with parametric part $f(\cdot)$ and kernel based part $g(\cdot)$.

II. Previous Work on Kernel Based Partially Linear Models

In the setting of Splines [6, Ch. 1] one assumes that two spaces $\mathcal{F}$ and $\mathcal{G}$ are given and that for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$ the function $f + g$ admits a unique decomposition. Under this condition one can form the direct sum of the two spaces $\mathcal{F} \oplus \mathcal{G} = \mathcal{H}$. The space $\mathcal{G}$ is a RKHS induced by a kernel K, while the space $\mathcal{F}$ can be expressed by a finite number of M basis functions $\psi_i : \mathbb{R}^D \to \mathbb{R}$, $i = 1, \ldots, M$. Given observational data $\mathcal{D} = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$ with $\boldsymbol{x}_t \in \mathbb{R}^D$ and $y_t \in \mathbb{R}$, a function $h \in \mathcal{H}$ can be estimated from

$$\min_{h \in \mathcal{H}} \quad \sum_{t=1}^{N} (y_t - h(\boldsymbol{x}_t))^2 + \lambda \| \mathscr{P}_{\mathcal{G}} h \|_{\mathcal{H}}^2 \tag{1}$$

where $\mathscr{P}_{\mathcal{G}}$ is the projector onto the reproducing kernel Hilbert Space (RKHS) $\mathcal{G}$ induced by K and $\lambda > 0$ a regularization constant. From a representer theorem [6] it follows that the solution to this problem can be expressed as

$$\hat{h}(\boldsymbol{z}) = \sum_{i=1}^{M} \vartheta_i \psi_i(\boldsymbol{z}) + \sum_{t=1}^{N} \alpha_t K(\boldsymbol{x}_t, \boldsymbol{z}) \tag{2}$$

with unknown coefficients $\vartheta_i, \alpha_t \in \mathbb{R}$. Using this finite dimensional representation the parameters can be obtained by solving the linear system [6, Eqs. (1.3.16) & (1.3.17)]

$$\begin{bmatrix} \boldsymbol{\Omega} + \lambda \boldsymbol{I} & \boldsymbol{\Psi}^T \\ \boldsymbol{\Psi} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\vartheta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}, \tag{3}$$

where $\Omega_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i, j = 1, \ldots, N$, $\boldsymbol{I}$ the N dimensional identity matrix, $\Psi_{ij} = \psi_i(\boldsymbol{x}_j)$ for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$, $\boldsymbol{\vartheta} = [\vartheta_1, \ldots, \vartheta_M]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^T$. While the work of Wahba [6] was derived for Spline kernels, there are extensions to arbitrary reproducing kernels [14].

III. IMPOSING ORTHOGONALITY CONSTRAINTS

This paper is concerned with the case when the parametric functions are (empirically) not orthogonal to the space induced by the kernel. In case one cannot form the direct sum of $\mathcal{F}$ and $\mathcal{G}$ and the estimation is still performed via (3) blindly without modification, certain drawbacks arise. This is due to a model ambiguity. Without imposing further restrictions on $\boldsymbol{\psi}$ and K some behavior may be modeled by one part as well as the other.

In this section we state a constrained optimization approach with primal and Lagrange dual model representations which straightforwardly admits to incorporate additional constraints [12]. In Section V the results are stated in a functional setting.

A. Partially linear model: primal and dual problem

We start with a partially linear model similar to the one in the previous section in a LS-SVM formulation as given in [7]

$$\begin{aligned} \min_{\boldsymbol{\vartheta},\boldsymbol{w},b,e_t} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^{N} e_t^2 \\ \text{subject to } & y_t = \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}) + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}) + b + e_t,\ t = 1,\ldots,N. \end{aligned} \tag{4}$$

Here $b \in \mathbb{R}$, $\boldsymbol{w} \in \mathbb{R}^{n_h}$ are model parameters of the nonparametric model and $\boldsymbol{\psi}(\boldsymbol{x}) = [\psi_1(\boldsymbol{x}), \ldots, \psi_M(\boldsymbol{x})]^T$. The feature map $\boldsymbol{\varphi} : \mathbb{R}^D \to \mathbb{R}^{n_h}$ is defined via its inner product $\boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{y}) = K(\boldsymbol{x},\boldsymbol{y})$. A popular choice for the positive semidefinite kernel K are RBF kernels $K_{\mathrm{RBF}}(\boldsymbol{x},\boldsymbol{y}) = \exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/\sigma^2)$ with $\sigma > 0$. Note that the feature maps associated with nonlinear kernels can be very high dimensional or even infinite dimensional as is the case of the RBF kernel. As the feature map is implicitly defined through the choice of a positive definite kernel is does not need to be known explicitly. For $b = 0$ the Lagrange dual is then equal to (3) with $\gamma = 1/\lambda$ where $\boldsymbol{\alpha}$ are Lagrange multipliers related to the constraints in (4).

B. Parametric estimates in the case of violated assumptions

The model ambiguity between the parametric and nonparametric parts can be seen from the estimates obtained for $\boldsymbol{\vartheta}$. To simplify the presentation we assume that the measurements y_t are zero mean. Then b in (4) is equal to zero and the LS-SVM estimate coincides with the Spline

one. Also assume that $\boldsymbol{\Psi}$ has full row rank and define $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_N)]$. Then the value for $\boldsymbol{\vartheta}$ obtained from (4) is

$$\hat{\boldsymbol{\vartheta}}_{\mathrm{PL}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}(\boldsymbol{y} - \boldsymbol{\Phi}^T\hat{\boldsymbol{w}}), \tag{5}$$

where $\hat{\boldsymbol{w}}$ is the estimate following from $(\boldsymbol{\Phi}\boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{\Phi}^T + \gamma^{-1}\boldsymbol{I})\boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{y}$, where $\boldsymbol{P}_{\Psi}^{\perp} = \boldsymbol{I} - \boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}$ is the projector onto the nullspace of $\boldsymbol{\Psi}$.

Under the assumptions used in previous works [6], [7] that the spaces $\mathcal{F}$ and $\mathcal{G}$ are orthogonal, the term $\boldsymbol{\Phi}\boldsymbol{\Psi}^T$ goes to zero as the number of samples N approaches infinity. Under the relaxed assumptions in [8] (nonparametric model is zero mean), [9] (no restrictions) this is not necessarily the case. For a finite amount of samples this does not hold even if the restrictive orthogonality assumption holds. Note that in the limit one has $\hat{\boldsymbol{\vartheta}}_{\mathrm{PL}} \to \hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$, i.e. the ordinary least squares (OLS) estimate.

Therefore in the case that $\boldsymbol{\Phi}\boldsymbol{\Psi}^T$ is empirically nonzero, the estimate $\hat{\boldsymbol{\vartheta}}_{\mathrm{PL}}$ depends not only on the data, but also on the kernel K and the regularization constant γ controlling the complexity of the nonparametric model. This is a manifestation of the ambiguity mentioned earlier.

Remark 1 (Ordinary least squares)**.** *Assume that the true underlying system generates data according to $y_t = \boldsymbol{\vartheta}_0^T\boldsymbol{\psi}(\boldsymbol{x}_t) + \varrho(\boldsymbol{x}_t) + \epsilon_t$, where $\varrho(\cdot)$ is the part of the underlying system that cannot be captured by the parametric function and ϵ_t is zero mean, Gaussian white noise with finite variance and independent of $\boldsymbol{x}_t$. Then the OLS estimate for $\boldsymbol{\vartheta}$ is $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = \boldsymbol{\vartheta}_0 + (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}(\boldsymbol{\epsilon} + \boldsymbol{\varrho})$, with $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$ and $\boldsymbol{\varrho} = [\varrho(\boldsymbol{x}_1), \ldots, \varrho(\boldsymbol{x}_N)]^T$. Note that the noise contribution goes to zero for increasing sample sizes. What remains is the contribution that depends only on the fraction of the true underlying system that is not captured by the parametric model.*

Remark 2 (Modeling of residuals)**.** *In view of Remark 1 it would seem beneficial to start with the OLS estimate $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}}$ and compute the residuals $r_t = y_t - \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t)$. However modeling these residuals with a nonparametric model gives rise to low prediction performances as shown in the experimental section.*

C. Imposing orthogonality

While a new model should have a comparable predictive performance as existing partially linear models, the goal of this paper is to ensure that the estimate of the parametric model given

by (5) is always as good as the OLS estimate. This should hold for finite sample sizes as well as when the orthogonality assumption is not satisfied. Therefore comparing (5) with the OLS estimate $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$ the term $(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\Phi}^T\hat{\boldsymbol{w}}$ has to be zero. This holds if and only if

$$\boldsymbol{\Psi}(\boldsymbol{\Phi}^T\hat{\boldsymbol{w}} + \mathbf{1}\hat{b}) = \mathbf{0} \tag{6}$$

holds. As standard SVM and LS-SVM models in regression and classification problems contain a bias term b it has been reintroduced here. Therefore this relation is imposed as an additional constraint in problem (4). This can be interpreted as (empirical) orthogonality of the parametric and nonparametric models. Another interpretation is that the nonparametric model is constrained such that its predictions are uncorrelated with the predictions of any function with the chosen parametrization. For a given data set $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$ the modified optimization problem is then

$$\begin{aligned} \min_{\boldsymbol{\vartheta},\boldsymbol{w},b,e_t} \ & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^N e_t^2 \\ \text{subject to} \quad & y_t = \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b + e_t, \quad t = 1,\ldots,N, \\ & \sum_{t=1}^N \psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b) = 0, \qquad i = 1,\ldots,M. \end{aligned} \tag{7}$$

Note that the basis functions $\boldsymbol{\psi}(\boldsymbol{x})$ need to be explicitly defined. On the other hand the feature map $\boldsymbol{\varphi}(\boldsymbol{x})$ is defined implicitly by a kernel function K at the dual level as is common in support vector machine methods.

D. Dual problem: model representation and estimation

As mentioned earlier in the case of the RBF kernel, the feature map $\boldsymbol{\varphi}$ is not only defined implicitly, but furthermore infinite dimensional. Thus it is in general not possible to directly solve the primal problem in (7). Therefore the problem is solved in the dual which is formalized in the following Lemma.

Lemma 1. *The solution to* (7) *is given by the linear system*

$$\begin{bmatrix} \boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{I} & \boldsymbol{\Omega}\boldsymbol{\Psi}^T & \mathbf{1} & \boldsymbol{\Psi}^T \\ \boldsymbol{\Psi}\boldsymbol{\Omega} & \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T & \boldsymbol{\Psi}\mathbf{1} & 0 \\ \mathbf{1}^T & \mathbf{1}^T\boldsymbol{\Psi}^T & 0 & 0 \\ \boldsymbol{\Psi} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ b \\ \boldsymbol{\vartheta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \tag{8}$$

The variables $\boldsymbol{\beta} \in \mathbb{R}^M$ are the Lagrange multipliers for the newly introduced orthogonality constraints in (7) *and can be used to express the dual predictive model as* $\hat{y} = \hat{h}(\boldsymbol{x}) = \sum_{t=1}^{N} \eta_t K(\boldsymbol{x}_t, \boldsymbol{x}) + \boldsymbol{\vartheta}^T \boldsymbol{\psi}(\boldsymbol{x}) + b$ *with* $\eta_t = \alpha_t + \boldsymbol{\beta}^T \boldsymbol{\psi}(\boldsymbol{x}_t)$.

Proof: The Lagrangian for (7) is $\mathscr{L}(\boldsymbol{w}, \boldsymbol{\vartheta}, b, e_t, \alpha_t, \beta_i) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma \sum_{t=1}^{N} e_t^2 - \sum_{t=1}^{N} \alpha_t (\boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_t) + \boldsymbol{\vartheta}^T \boldsymbol{\psi}(\boldsymbol{x}_t) + b + e_t - y_t) - \sum_{i=1}^{M} \beta_i \sum_{t=1}^{N} \psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_t) + b)$. The corresponding KKT conditions for optimality are $\mathbf{0} = \frac{\partial \mathscr{L}}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{t=1}^{N} (\alpha_t + \boldsymbol{\beta}^T \boldsymbol{\psi}(\boldsymbol{x}_t)) \boldsymbol{\varphi}(\boldsymbol{x}_t)$, $\mathbf{0} = \frac{\partial \mathscr{L}}{\partial \boldsymbol{\vartheta}} = -\sum_{t=1}^{N} \alpha_t \boldsymbol{\psi}(\boldsymbol{x}_t)$, $0 = \frac{\partial \mathscr{L}}{\partial b} = -(\sum_{t=1}^{N} \alpha_t + \boldsymbol{\beta}^T \psi(\boldsymbol{x}_t))$, $0 = \frac{\partial \mathscr{L}}{\partial e_t} = \gamma e_t - \alpha_t \; \forall t$, $0 = \frac{\partial \mathscr{L}}{\partial \alpha_t} = y_t - \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_t) - \boldsymbol{\vartheta}^T \boldsymbol{\psi}(\boldsymbol{x}_t) - b - e_t \; \forall t$ and $0 = \frac{\partial \mathscr{L}}{\partial \beta_i} = -\sum_{t=1}^{N} \psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_t) + b)$, $i = 1, \ldots, M$. Rewriting the conditions for $\boldsymbol{w}$ and e_t and substituting them into $\partial \mathscr{L} / \partial \alpha_t$ yields $y_t = \sum_{k=1}^{N} (\alpha_k + \boldsymbol{\beta}^T \boldsymbol{\psi}(\boldsymbol{x}_k)) K(\boldsymbol{x}_k, \boldsymbol{x}_t) + \boldsymbol{\vartheta}^T \boldsymbol{\psi}(\boldsymbol{x}_t) + b + \frac{1}{\gamma}\alpha_t$. Substitution of $\boldsymbol{w}$ into $\partial \mathscr{L} / \partial \beta_i$ yields $\sum_{t=1}^{N} \psi_i(\boldsymbol{x}_t) \sum_{k=1}^{N} (\alpha_k + \boldsymbol{\beta}^T \boldsymbol{\psi}(\boldsymbol{x}_k)) K(\boldsymbol{x}_k, \boldsymbol{x}_t) + b = 0$. Here, the inner products $\boldsymbol{\varphi}(\boldsymbol{x}_k)^T \boldsymbol{\varphi}(\boldsymbol{x}_t)$ were replaced by a positive definite kernel $K(\boldsymbol{x}_k, \boldsymbol{x}_t)$ using the kernel trick. The rewritten KKT conditions $\partial \mathscr{L} / \partial \alpha_t$ and $\partial \mathscr{L} / \partial \beta_i$ together with $\partial \mathscr{L} / \partial b$ and $\partial \mathscr{L} / \partial \boldsymbol{\vartheta}$ correspond to the dual system (8). ■

IV. Alternative estimation schemes and representations

A. Separation principle

Due to the orthogonality constraint imposed on the kernel based model the estimation problem can be decoupled.

Lemma 2. *Under the assumption that $\boldsymbol{\Psi}$ has full row rank, the estimation of the model can be performed in two stages as follows:*

1) *The parameter vector $\boldsymbol{\vartheta}$ can be estimated using ordinary least squares $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$.*
2) *The kernel based model can then be identified by solving the reduced linear system*

$$\begin{bmatrix} \boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{P}_{\Psi}^{\perp} & \mathbf{1} \\ \mathbf{1}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{r} \\ 0 \end{bmatrix} \tag{9}$$

with $\boldsymbol{r} = \boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{\Psi}^T\hat{\boldsymbol{\vartheta}}$ and as before $\boldsymbol{P}_{\Psi}^{\perp} = \boldsymbol{I} - \boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}$.

Proof: Let $\mathcal{R}_1$ to $\mathcal{R}_4$ denote the block rows of (8). Isolating $\boldsymbol{\Omega}\boldsymbol{\alpha}$ in $\mathcal{R}_1$ and substituting this into $\mathcal{R}_2$ yields $\boldsymbol{\Psi}\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\vartheta} + \gamma^{-1}\boldsymbol{\Psi}\boldsymbol{\alpha}$. Exploiting $\mathcal{R}_4$ we obtain $\boldsymbol{\vartheta} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$. Now consider a change of variables $\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\psi}^T\boldsymbol{\beta}$. Substitution of this relation into $\mathcal{R}_4$ we

obtain $\boldsymbol{\beta} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\eta}$. Stating $\mathcal{R}_3$ in terms of $\boldsymbol{\eta}$ yields $\mathbf{1}^T\boldsymbol{\eta} = 0$ and $\mathcal{R}_1$ can be rewritten as $(\boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{P}_{\Psi}^{\perp})\boldsymbol{\eta} + \mathbf{1}b = \boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{y}$. ■

Note that $\mathcal{R}_2$ in terms of $\boldsymbol{\eta}$ reads $\boldsymbol{\Psi}(\boldsymbol{\Omega}\boldsymbol{\eta} + \mathbf{1}b) = \mathbf{0}$. This nicely illustrates that the correlation between the kernel based model $\boldsymbol{\Omega}\boldsymbol{\eta} + \mathbf{1}b$ and the basis functions $\boldsymbol{\Psi}$ is zero on the estimation data.

B. Equivalent kernel

For simplicity of the presentation we will assume that the bias term b is zero in this subsection. This is no limitation as a static offset can be included in the parametric basis. In fact if this is done the orthogonality constraint will ensure that b is zero.

Lemma 3. *Under the assumption that $\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T$ is invertible, problem* (7) *can be solved equivalently as*

$$\begin{aligned} &\min_{\boldsymbol{w},e_t} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^{N} e_t^2 \\ &\text{subject to} \quad y_t = \boldsymbol{w}^T\boldsymbol{\varphi}_{\text{eq}}(\boldsymbol{x}_t) + e_t, \quad t = 1,\ldots,N, \end{aligned} \tag{10}$$

with an equivalent feature map $\boldsymbol{\varphi}_{\text{eq}}(\cdot) = \boldsymbol{P}_{\Phi\Psi}^{\perp}\boldsymbol{\varphi}(\cdot)$ where $\boldsymbol{P}_{\Phi\Psi}^{\perp}$ is defined as $\boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\Phi}^T$ with $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\varphi}(\boldsymbol{x}_N)]$.

This primal formulation can be solved using the dual linear system

$$\begin{bmatrix} \boldsymbol{\Omega}_{\text{eq}} + \gamma^{-1}\boldsymbol{I} & \boldsymbol{\Psi}^T \\ \boldsymbol{\Psi} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\vartheta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix},$$

where $(\boldsymbol{\Omega}_{\text{eq}})_{ij} = K_{\text{eq}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for $i,j = 1,\ldots,N$. The equivalent kernel is given by

$$K_{\text{eq}}(\boldsymbol{x},\boldsymbol{y}) = K(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{k}(\boldsymbol{x})^T\boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{k}(\boldsymbol{y}) \tag{11}$$

with $\boldsymbol{k}(\boldsymbol{z}) = [K(\boldsymbol{x}_1,\boldsymbol{z}),\ldots,K(\boldsymbol{x}_N,\boldsymbol{z})]^T$. This model can be evaluated at a new point $\boldsymbol{z}$ through $\hat{y} = \hat{h}(\boldsymbol{z}) = \sum_{t=1}^{N} K_{\text{eq}}(\boldsymbol{x}_t,\boldsymbol{z})\alpha_t + \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{z})$.

Proof: Substitution of the expansion for $\boldsymbol{w}$ obtained from the KKT conditions in the proof of Lemma 1 into the KKT condition for $\boldsymbol{\beta}$ gives $\boldsymbol{\Psi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{\alpha} + \boldsymbol{\Psi}\boldsymbol{\beta}) = 0$. Then, assuming invertibility, the Lagrange multipliers $\boldsymbol{\beta}$ can be expressed as $\boldsymbol{\beta} = -(\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\alpha}$. Using this solution in the expansion of $\boldsymbol{w}$ yields $\boldsymbol{w} = \boldsymbol{P}_{\Phi\Psi}^{\perp}\boldsymbol{\Phi}\boldsymbol{\alpha}$. Reading this columnwise for the columns of $\boldsymbol{\Phi}$ the equivalent feature map can be extracted. The equivalent kernel follows directly from $K_{\text{eq}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\varphi}_{\text{eq}}(\boldsymbol{x})^T\boldsymbol{\varphi}_{\text{eq}}(\boldsymbol{y}) = \boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{P}_{\Phi\Psi}^{\perp}\boldsymbol{\varphi}(\boldsymbol{y})$. ■

V. CHARACTERIZATION IN RKHSS

In this section we detail an alternative derivation based on the theory of RKHS's. Differently from [6], [14] we do not assume the existence of a direct sum decomposition between the spaces $\mathcal{F}$ and $\mathcal{G}$. Rather we construct a space by forming the direct sum between $\mathcal{F}$ and a certain suitable subspace $\mathcal{G}_{\mathrm{eq}}$ of $\mathcal{G}$.

More precisely the argument starts from an arbitrary RKHS $\mathcal{G}$ of real-valued functions on $\mathbb{R}^D$ and a finite dimensional space $\mathcal{F}$ spanned by the prescribed set of real-valued functions $\psi_i : \mathbb{R}^D \to \mathbb{R}$, $i = 1, \ldots, M$. We denote by K the reproducing kernel of $\mathcal{G}$ and by $\langle \cdot, \cdot \rangle_K$ the corresponding inner product. Furtheron we show that $\mathcal{G}_{\mathrm{eq}}$ is a RKHS and derive its reproducing kernel K_{eq}. As in the preceding sections the idea is based on an orthogonality criterion formulated upon the empirical sample. We begin by recalling the following result.

Lemma 4. *Kernel of a Closed Subspace [15, Theorem 11] Let $\mathcal{V}$ be a closed subspace of a RKHS $\mathcal{G}$. Then $\mathcal{V}$ is a RKHS and its kernel K' is given by*

$$K'(\boldsymbol{x}, \boldsymbol{y}) = \left[\mathscr{P}_{\mathcal{V}} K(\cdot, \boldsymbol{y})\right](\boldsymbol{x}) \tag{12}$$

where $\mathscr{P}_{\mathcal{V}}$ denotes the projection operator onto $\mathcal{V}$.

We now introduce the sampling operator $\mathscr{S} : \mathcal{L}_2 \to \mathbb{R}^N$, defined entry-wise as $(\mathscr{S})_t := f(\boldsymbol{x}_t)$ for $t = 1, \ldots, N$ and the linear operator $\mathscr{C} : \mathcal{L}_2 \to \mathbb{R}^M$ defined as $\mathscr{C} f = \boldsymbol{\Psi} \mathscr{S} f$ where $f \in \mathcal{L}_2$ where $\mathcal{L}_2$ is the space of square integrable functions[1]. It is not difficult to check that its adjoint operator $\mathscr{C}^* : \mathbb{R}^M \to \mathcal{G}$ is defined by $\mathscr{C}^* \boldsymbol{\beta} := \sum_{i=1}^{M} \sum_{t=1}^{N} \beta_i \psi_i(\boldsymbol{x}_t) K(\boldsymbol{x}_t, \cdot)$, where $\boldsymbol{\beta} \in \mathbb{R}^M$. Now $\mathrm{range}(\mathscr{C}^*)$ is a finite dimensional subspace of $\mathcal{G}$, $\mathrm{null}(\mathscr{C})$ is a closed subspace and

$$\mathcal{G} = \mathrm{range}(\mathscr{C}^*) \oplus \mathrm{null}(\mathscr{C}).$$

By Lemma 4 $\mathrm{null}(\mathscr{C})$ is a RKHS and its kernel function is solely given in terms of K and of $\mathscr{P}_{\mathrm{null}(\mathscr{C})}$ the projector onto the nullspace of $\mathscr{C}$. We have now that $\mathscr{P}_{\mathrm{null}(\mathscr{C})} : \mathcal{G} \to \mathcal{G}$ is given by (see e.g. [17, Theorem 1, Chapter 6.9])

$$\mathscr{P}_{\mathrm{null}(\mathscr{C})} g = \left(\mathscr{I}_{\mathrm{id}} - \mathscr{C}^* \left(\mathscr{C}\mathscr{C}^*\right)^{-1} \mathscr{C}\right) g = g - \sum_{t=1}^{N} \left(\boldsymbol{\Psi}^T \boldsymbol{\Omega}_{\Psi}^{-1} \boldsymbol{\Psi} \mathscr{S} g\right)_t K(\boldsymbol{x}_t, \cdot)$$

[1]We implicitly assume that $\mathcal{F}, \mathcal{G} \subset \mathcal{L}_2$ which, for $\mathcal{G}$, is true under mild conditions on the kernel K [16].

where $\boldsymbol{\Omega}_\Psi := \mathscr{C}\mathscr{C}^* = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T$ is assumed to be a full-rank matrix as in Lemma 3. We denote by $\mathscr{I}_{\mathrm{id}}$ the identity operator. By equation (12) we now have that the reproducing kernel of $\mathrm{null}(\mathscr{C})$, denoted by K_{eq}, is given by $K_{\mathrm{eq}}(\boldsymbol{x}, \boldsymbol{y}) = \left[\mathscr{P}_{\mathrm{null}(\mathscr{C})} K(\cdot, \boldsymbol{y})\right](\boldsymbol{x})$ which corresponds to (11). Denote now $\mathrm{null}(\mathscr{C})$ by $\mathcal{G}_{\mathrm{eq}}$. We have the following.

Proposition 1. *For any $f_1 \in \mathcal{F}$ and any $g_1 \in \mathcal{G}_{\mathrm{eq}}$ the function $h = f_1 + g_1$ admits a unique decomposition under the assumption that $\boldsymbol{\Psi}$ has full rank.*

Proof: Suppose that there are $f_2 \in \mathcal{F}$ and $g_2 \in \mathcal{G}$ with $f_1 \neq f_2$ such that $h = f_2 + g_2$. Note that f_1 and f_2 can be parametrized as $f_1(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\vartheta}$ and $f_2(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\vartheta}'$ respectively. As $f_1 \neq f_2$ also $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}'$. By assumption we need to have $(f_1 + g_1)(\boldsymbol{x}) = (f_2 + g_2)(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^D$. By construction we have $\mathscr{C} g_1 = \mathscr{C} g_2 = \mathbf{0}_M$. Therefore we have $\boldsymbol{\Psi}\mathscr{S} f_1 = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\vartheta} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\vartheta}' = \boldsymbol{\Psi}\mathscr{S} f_2$. As $\boldsymbol{\Psi}$ has full rank it follows $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}'$ which is a contradiction. ■

It now follows from Proposition 1 that we can write

$$\mathcal{H} := \mathcal{F} \oplus \mathcal{G}_{\mathrm{eq}}.$$

For any $h_1 = \boldsymbol{\vartheta}^T\boldsymbol{\psi} + g_1$ and $h_2 = \boldsymbol{\vartheta}'^T\boldsymbol{\psi} + g_2$ we can define the inner product

$$\langle h_1, h_2 \rangle := \boldsymbol{\vartheta}^T\boldsymbol{\vartheta}' + \langle g_1, g_2 \rangle_K$$

that turns $\mathcal{G}_{\mathrm{eq}}$ and $\mathcal{F}$ into orthogonal complements. The corresponding norm is $\|h\| := \sqrt{\langle h, h \rangle}$. Denoting by $\mathscr{P}_{\mathcal{G}_{\mathrm{eq}}}$ the projection onto $\mathcal{G}_{\mathrm{eq}}$ by the representer theorem, the solution of (1) is given by (2) where $\boldsymbol{\vartheta} \in \mathbb{R}^M$ and $\boldsymbol{\alpha} \in \mathbb{R}^N$ can be found in closed form by solving the system of linear equations (3). Since in the current setting we have that $\boldsymbol{\Psi}\boldsymbol{\Omega}_{\mathrm{eq}}\boldsymbol{\alpha} = \mathbf{0}_M$ similar arguments as in Lemma 2 can be used to establish that $\boldsymbol{\vartheta}$ is given by $\boldsymbol{\vartheta}_{\mathrm{OLS}}$. Once the parametric part has been computed one can also find the non-parametric part starting from the residuals $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{\Psi}^T\boldsymbol{\vartheta}_{\mathrm{OLS}}$ and solve $(\boldsymbol{\Omega}_{\mathrm{eq}} + \lambda \boldsymbol{I})\boldsymbol{\alpha} = \boldsymbol{r}$. The final model (for out-of sample prediction) is given by (2) in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\vartheta}$ and the equivalent kernel function K_{eq} that is given in terms of the original kernel K, the training data $\{\boldsymbol{x}_t\}_{t=1}^N$ and the parametric basis functions $\{\psi_i\}_{i=1}^N$.

VI. Extension to different loss functions

The outlined procedure can be extended to other loss functions. The ϵ-insensitive loss for example is defined as $l_\epsilon(x) = 0$ for$|x| \leq \epsilon$ and $l_\epsilon(x) = |x| - \epsilon$ otherwise. If this loss is used

instead of the L_2-loss as in (7) a SVM [11] is obtained. The primal formulation is

$$\begin{aligned} \min_{\boldsymbol{\vartheta},\boldsymbol{w},b,\xi_t} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{t=1}^{N}\xi_t \\ \text{subject to} \quad & |y_t - \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) - \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) - b| \leq \epsilon + \xi_t,\ \xi_t \geq 0, \quad t = 1,\ldots,N, \\ & \sum_{t=1}^{N}\psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b) = 0, \qquad\qquad i = 1,\ldots,M. \end{aligned} \tag{13}$$

The Lagrange dual can be obtained by solving the KKT system

$$\begin{aligned} \max_{\alpha_t,\beta_i} \quad & \boldsymbol{y}^T\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha} + \boldsymbol{\Psi}^T\boldsymbol{\beta})^T\boldsymbol{\Omega}(\boldsymbol{\alpha} + \boldsymbol{\Psi}^T\boldsymbol{\beta}) - \epsilon\mathbf{1}^T|\boldsymbol{\alpha}| \\ \text{subject to} \quad & -C \leq \alpha_t \leq C, \qquad\qquad t = 1,\ldots,N, \\ & \boldsymbol{\Psi}\boldsymbol{\alpha} = \mathbf{0}, \qquad \mathbf{1}^T(\boldsymbol{\alpha} + \boldsymbol{\Psi}^T\boldsymbol{\beta}) = 0. \end{aligned} \tag{14}$$

where $|\boldsymbol{\alpha}|$ is understood as the elementwise absolute value. The predictive model is identical to the one stated in Lemma 1. The primal variables b and ϑ_i can for example be obtained from a interior point solver and correspond to the dual variables of the equality constraints in (14).

Extensions to Huber's robust loss function and other convex loss functions can be derived in a similar fashion.

VII. Experiments

A. Experimental setup

In this section several modeling approaches are compared for simulated as well as for real life data. The following model structures using NARX regressors are considered

- purely parametric models (`LIN`),
- purely nonparametric black box models (`LS-SVM`, `SVM`),
- parametric model + nonlinear black box model for residuals (`RES (LS-SVM)`, cf. Remark 2),
- partially linear model without orthogonality constraints (`PL (LS-SVM)`, cf. Eq. (4)) and
- partially linear models with orthogonality constraints (`OPL (LS-SVM)`, cf. Eq. (7) and `OPL (SVM)`, cf. Eq. (13)).

For all experiments the parametric part is taken linear $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x}$ while the nonparametric contribution is modeled with a RBF kernel. Considering the results from [18] we note that the

assumption (direct sum decomposition) for classical partially linear models `PL` is satisfied as linear functions (and more generally polynomials) and the space induced by the Gaussian kernel can be combined by forming the direct sum of both. Therefore the proposed constraint would not be needed, but we show that in case of small but realistic sample sizes (> 1000 samples) we observe an improvement by using the constraint.

For the simulation data we report the distance of the estimated parametric model to the "true" one in addition to the root mean squared error (RMSE) on independent data. For the comparison of the linear submodels identified by the different models, a metric for ARMA systems proposed by Martin [19] is used. This metric compares the spectra of linear systems and as such is independent of their parametrization. For elaborate arguments why this metric is better in terms of system-theoretic properties than e.g. comparing AR coefficients please consult the reference. For two stable AR systems S_A, S_B with poles p_i^A, p_j^B, $i = 1, \ldots, P_A, j = 1, \ldots, P_B$ the metric can be computed as $d(S_A, S_B) = \ln(\prod_{i=1}^{P_A} \prod_{j=1}^{P_B} |1-p_i^A p_j^{B*}|^2 / (\prod_{i,j=1}^{P_A} (1-p_i^A p_j^{A*}) \prod_{i,j=1}^{P_B} (1-p_i^B p_j^{B*})))$ where x^* denotes the complex conjugate of x. For the generalization to stable ARX systems we refer to [19].

The tuning parameters of the nonlinear models (kernel parameters and regularization constant) are selected according to one-step ahead prediction performance on a validation set using grid search.

B. Mass-Spring-Damper system

To test the proposed method we consider a simple mass spring damper system with two masses and springs with cubic nonlinearities as depicted in Figure 2. The system is described by $m_1\ddot{s}_1 = -L_1 f_1(s_1)+L_2 f_2(s_2-s_1)-C_1\dot{s}_1+C_2(\dot{s}_2-\dot{s}_1)$, $m_2\ddot{s}_2 = -L_2 f_2(s_2-s_1)-C_2(\dot{s}_2-\dot{s}_1)+F_u$, $y = s_2$, $f_1(x) = x+0.02x^3$ and $f_2(x) = x+0.005x^3$. The states s_1 and s_2 are the displacement of mass 1 and mass 2 from rest respectively. The constants are chosen as $m_1 = 5\,\mathrm{kg}$, $m_2 = 0.1\,\mathrm{kg}$, $L_1 = 1.5\,\mathrm{N/m}$, $L_2 = 0.5\,\mathrm{N/m}$, $C_1 = 0.3\,\mathrm{Ns/m}$ and $C_2 = 0.05\,\mathrm{Ns/m}$. The output y is sampled with $4\,\mathrm{Hz}$ and the excitation force F_u is given by a Gaussian white noise process u_t. The equations are simulated using MATLAB's `ode45`, a zero order hold stage for the input signal and zero initial conditions. For each realization of the input signal and every amplitude value, 4000 samples are generated. The first 400 samples of each data set are discarded to exclude transient behavior. To compare measurements with different amounts of nonlinearity, excitation

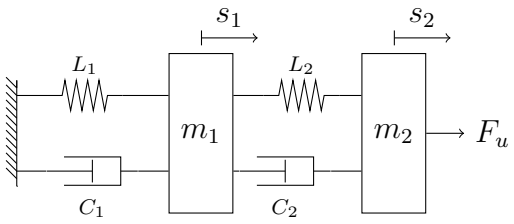


Fig. 2. Mass-spring-damper system with nonlinear springs.

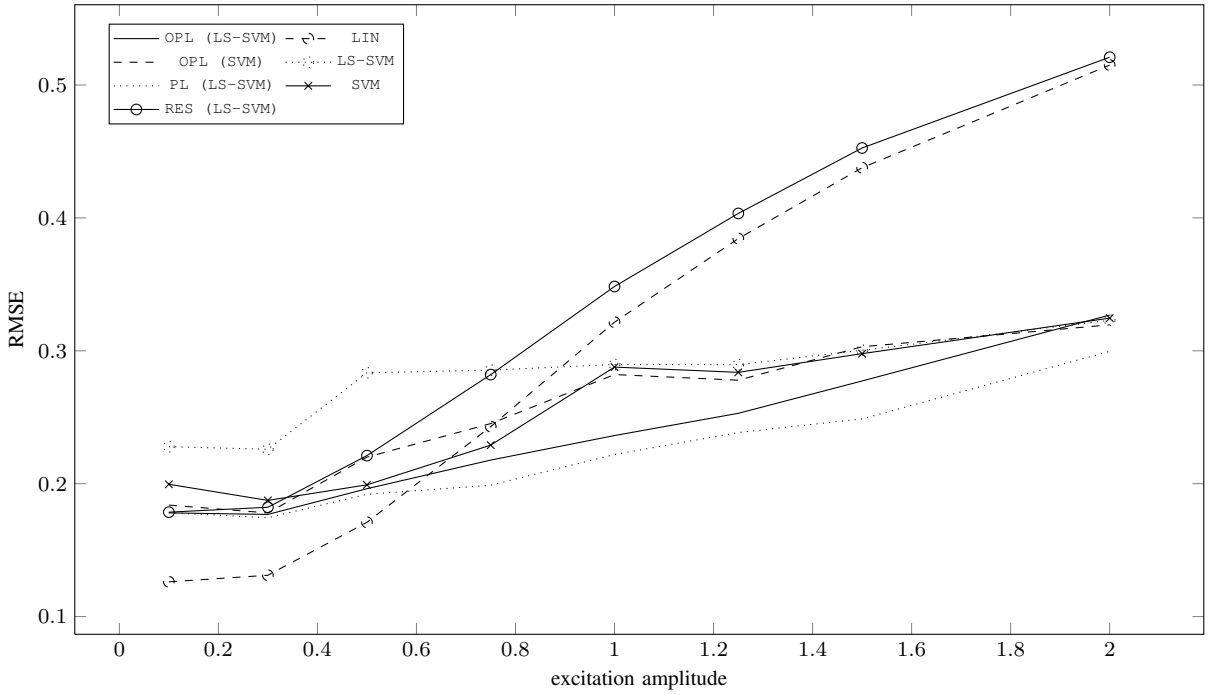


Fig. 3. Prediction performance of different model structures for the mass spring damper system on an independent test set as a function of excitation amplitude. The root mean squared error for simulated outputs averaged over ten input realizations is being reported.

amplitudes of $\{0.1, 0.3, 0.5, 0.75, 1, 1.25, 1.5, 2\}$ are tested. This accounts to different fractions of the system not being captured by the parametric (linear) model. For each amplitude the estimation is performed for ten realizations of the input signal. All data sets are split in three parts of equal size for model estimation, model validation and test.

Figure 3 depicts the simulation performance for different model structures. The model orders are as follows $p = q = 13$ for `LIN` and $p = q = 10$ for all other models, where the input variable

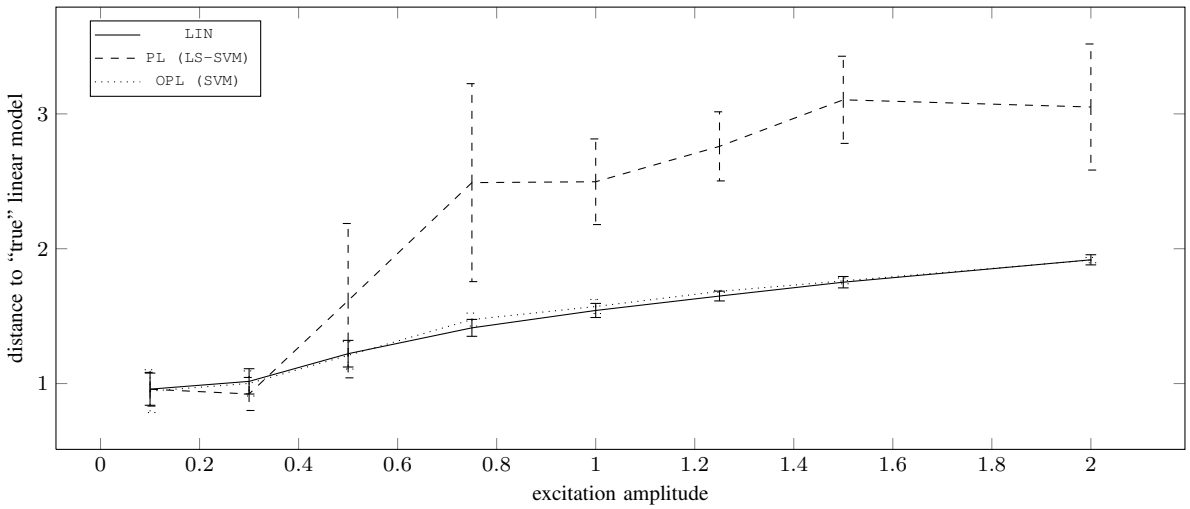


Fig. 4. Distance of the estimated parametric submodel to the "true" linear model for the mass-spring damper system for several model structures as a function of the excitation amplitude. The models `LIN`, `OPL (LS-SVM)` and `RES (LS-SVM)` have identical distances. Error bars indicate the standard deviation for ten realizations of the input signal. The comparison is based on the ARMA metric proposed in [19].

$\boldsymbol{x}_t$ has NARX structure with $\boldsymbol{x}_t = [y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-q}]^T$. For a fair comparison the input sequence u_t as well as the output sequence y_t are normalized to unit variance for each amplitude. As it can be expected the linear model is best for small amplitudes but degrades quickly as the amplitude increases. The nonlinear black box models `LS-SVM` and `SVM` are the best models for "large" amplitudes. Modeling the residuals is clearly inferior to all other approaches as it is never better than the linear model. Finally the partially linear structures give improved prediction performance for a number of amplitudes. Imposing orthogonality constraints yields slightly worse predictive performance than using a partially linear structure without constraints.

Setting $f_1(x) = f_2(x) = x$ a "true" linear system can be obtained. Using MATLAB's `c2d` a discrete time transfer function is computed from the continuous time system described above. The "true" linear system is of order four while the identified models are of orders thirteen and ten. Therefore the ARMA metric described earlier is used to compare the estimated with the "true" model. The result is shown in Figure 4. It can be seen that for small amplitudes the partially linear model is close to the linear estimate. Yet for larger amplitudes it quickly diverges

away from the linear estimate and not only has a larger distance from the "true" linear model but the estimates also have a larger variance. While both partially linear structures move away from the "true" linear model as the excitation amplitude increases, the models with orthogonality constraints degrade more gracefully than those without.

In conclusion, both partial linear structures yield better predictive performance than a parametric or nonparametric model alone. The unique advantage of using orthogonality constraints is that the estimate for parametric part has a much smaller variance and also is much closer to the "true" underlying model as seen in Fig. 4.

C. Wiener-Hammerstein benchmark data

To test on a real life data set, we consider the data of the Wiener-Hammerstein benchmark [20] at SYSID2009. As it just serves as a test problem, only a small subset of the 188,000 measured samples is taken. For model estimation and validation 2,000 samples each are used. The prediction performance is computed on 5,000 independent samples.

The results are shown in Table I, the reported values are for outputs normalized to unit variance. It can be seen that roughly the same behavior as with the simulated data is present. The partially linear models improve over the purely linear and purely nonlinear models. Using orthogonality constraints results in a slight degradation of prediction performance as observed in the previous section. In tests with small estimation sample sizes the prediction performance of of `OPL (LS-SVM)` is better than `PL (LS-SVM)` up to a sample size of about 200. Figure 5 shows transfer functions derived from the parametric model parts. The estimation is done for 20 subsamples of size 300. One can observe that the `OPL` structures are much closer to the reference spectrum given by the Best Linear Approximation than the `PL` model. Also their variability is much smaller.

VIII. Conclusions

Is was shown that the incorporation of orthogonality constraints into partially linear models is able to improve the parametric estimate for finite sample sizes. This is achieved by ensuring that the estimate is equal to the OLS estimate. Numerical examples illustrate that the proposed model structure is able to improve over a purely parametric or a purely nonparametric model. The parametric estimate obtained when considering the additional constraints is better in mean

TABLE I

PERFORMANCE OF DIFFERENT MODEL STRUCTURES ON A SUBSAMPLE OF THE WIENER-HAMMERSTEIN BENCHMARK DATASET. THE PERFORMANCE IS REPORTED AS ROOT MEAN SQUARE ERROR FOR SIMULATION VALUES.

model	training data	validation data	testing data
`LIN`	0.164	0.230	0.168
`LS-SVM`	0.067	0.183	0.135
`SVM`	0.053	0.139	0.079
`RES (LS-SVM)`	0.162	0.222	0.155
`PL (LS-SVM)`	0.037	0.133	0.068
`OPL (LS-SVM)`	0.091	0.166	0.094
`OPL (SVM)`	0.051	0.140	0.071

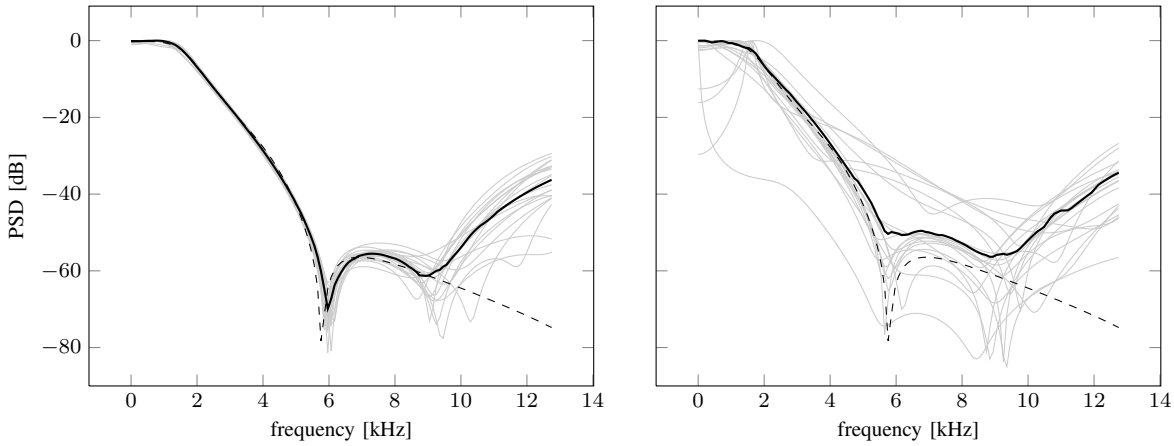


Fig. 5. Estimated transfer functions for the proposed model with orthogonality constraints (`OPL (LS-SVM)`, left) which improve over classical partially linear models (`PL (LS-SVM)`, right). The estimates are obtained for 20 different training samples of size of 300 taken out of the Wiener-Hammerstein benchmark dataset. Each light gray line corresponds to one realization, the thick black line indicates the median value while the dashed line gives the Best Linear Approximation [21] estimated on the full dataset as reference. The excitation signal was subject to low-pass filtering with a cut-off frequency of 10 kHz.

and much better in variance than existing techniques. This ensures that interpretabilty of the parametric model is not lost in a partially linear setting. Therefore in many cases where partially linear cases are applied inclusion of the proposed constraint is the better choice. For small sample sizes there is a small gain in predictive performance while otherwise there is a slight loss.

REFERENCES

[1] W. Härdle, H. Liang, and J. Gao, *Partially linear models*. Springer, 2000.

[2] L. Ljung, *System identification: Theory for the User*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1999.

[3] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, pp. 1691–1724, Dec. 1995.

[4] G. Wahba, "Partial spline models for the semiparametric estimation of functions of several variables," in *Statistical analysis of time series*. Tokyo: Institute of Mathemical Statistics, 1984, pp. 319–329.

[5] P. Speckman, "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 50, no. 3, pp. 413–436, 1988.

[6] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990.

[7] M. Espinoza, J. A. K. Suykens, and B. De Moor, "Kernel Based Partially Linear Models and Nonlinear Identification," *IEEE Transactions on Automatic Control*, vol. 50, pp. 1602–1606, Oct. 2005.

[8] Y.-F. Li, L.-J. Li, H.-Y. Su, and J. Chu, "Least Squares Support Vector Machine Based Partially Linear Model Identification," *Lecture Notes in Computer Science*, vol. 4113/2006, pp. 775–781, 2006.

[9] Y.-L. Xu and D.-R. Chen, "Partially-Linear Least-Squares Regularized Regression for System Identification," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2637–2641, 2009.

[10] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, 2002.

[11] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[12] J. A. K. Suykens, C. Alzate, and K. Pelckmans, "Primal and dual model representations in kernel-based learning," *Statistics Surveys*, vol. 4, pp. 148–183, Aug. 2010.

[13] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481 –1497, Sept. 1990.

[14] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, pp. 416–426, 2001.

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

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

[17] D. Luenberger, *Optimization by Vector Space Methods*. Wiley-IEEE, 1998.

[18] M. Ha Quang, G. Pillonetto, and A. Chiuso, "Nonlinear system identification via gaussian regression and mixtures of kernels," in *Proceedings of the 15th IFAC Symposium on System Identification*, Saint-Malo, France, Aug. 2009, pp. 528–533.

[19] R. J. Martin, "A Metric for ARMA Processes," *IEEE Transactions on Signal Processing*, vol. 48, no. 4, pp. 1164–1170, 2000.

[20] J. Schoukens, J. A. K. Suykens, and L. Ljung, "Wiener-Hammerstein benchmark," in *Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, 2009.

[21] J. Schoukens, J. G. Nemeth, P. Crama, Y. Rolain, and R. Pintelon, "Fast approximate identification of nonlinear systems," *Automatica*, vol. 39, pp. 1267–1274, 2003.

A two stage algorithm for kernel based partially linear modeling with orthogonality constraints

Tillmann Falck, Marco Signoretto, Johan A.K. Suykens,
Bart De Moor

***Abstract*—It is shown that for existing kernel based partially linear modeling approaches, the parametric model part estimate depends on the kernel function. A novel orthogonality constraint ensures that this dependency is removed and that the estimate for the parametric part is as good as the ordinary least squares estimate. The results are presented in a optimization setting with primal and dual model representations as well as in a functional setting.**

***Index Terms*—system identification, basis functions, kernel methods, SVMs, LS-SVMs, RKHSs, partially linear models, semi-parametric models**

I. Introduction

Partially linear models [1] as shown in Figure 1 are a special case of semi-parametric models that combine a parametric model with a nonparametric one. In partially linear models the parametric part is linear in the parameters. Their usage is beneficial if a model (class) for some explanatory variables is known and a nonparametric model shall be used to improve the prediction performance.

Especially in system identification [2] for nonlinear systems, good parametric models can be hard to obtain. Yet the use of black box models [3] may have a higher computational cost and more importantly is mostly much harder to interpret. Also prior knowledge embedded in the choice of the parametric model is lost when switching to a nonparametric model which can yield a worse prediction performance and therefore defeat the purpose of using a black box model.

There is a wealth of literature for kernel based partially linear models [4], [5], [6], [7], [8], [9]. However in [4], [6], [7] it is assumed that the parametric and nonparametric function spaces are disjoint. Therefore it can be difficult to find a suitable nonparametric model for an existing and fixed parametric model. If there is an admissible combination, finite sample sizes give rise to problems similar to those when the assumptions are violated. Therefore even if the assumptions are satisfied, the estimate of the parametric model obtained with existing techniques can be substantially worse than the ordinary least squares (OLS) solution. In this paper we will address the implications of relaxing those assumptions and propose a modified estimation scheme to overcome these. The modification is based on an additional constraint that ensures that the estimates generated by the parametric and the nonparametric models are empirically orthogonal (uncorrelated). We also show how this relates to function spaces.

In this work Least Squares Support Vector Machines (LS-SVMs) [10] and Support Vector Machines (SVMs) [11] are used for the nonparametric part. SVMs and LS-SVMs are black box regression techniques based on convex optimization formulations, which allow to incorporate additional constraints straightforwardly [12]. Alternative kernel based regression techniques are e.g. splines [6] and regularization networks [13], which are formulated in Reproducing Kernel Hilbert Spaces (RKHSs).

T. Falck, M. Signoretto, J. Suykens and B. De Moor are with the SCD/SISTA group at the department of Electrical Engineering (ESAT) at the Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. E-mail: {tillmann.falck,marco.signoretto, johan.suykens,bart.demoor}@esat.kuleuven.be.

Research supported by Research Council KUL: PhD/postdoc grants, GOA/10/09 MaNet, CoE EF/05/006 OPTEC, several PhD/postdoc & fellow grants; Flemish Government: FWO G.0302.07 (SVM/Kernel), PhD/postdoc grants by FWO and IWT; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO); EU: ERNSI. Johan Suykens is a professor and Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium.

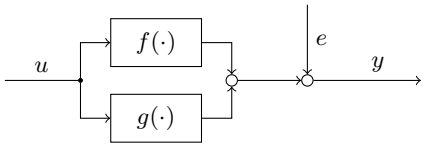


Fig. 1. Partially linear model with parametric part $f(\cdot)$ and kernel based part $g(\cdot)$.

The paper is organized as follows. In Section II we briefly review existing work on kernel based partially linear models, with an emphasis on the necessary assumptions. The effect of violating these assumptions is analyzed in III. We then propose a modification to overcome this effect and derive the corresponding kernel based model in the same section. In Section IV a separable estimation problem is presented and an equivalent kernel is derived. The next section derives equivalent results in a RKHS setting. The results based on LS-SVMs are then extended to different loss functions in the Section VI, where we integrate the idea into SVMs [11]. In Section VII we illustrate the effect of our modification on simulated as well as real world data.

II. Previous Work on Kernel Based Partially Linear Models

In the setting of Splines [6, Ch. 1] one assumes that two spaces $\mathcal{F}$ and $\mathcal{G}$ are given and that for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$ the function $f+g$ admits a unique decomposition. Under this condition one can form the direct sum of the two spaces $\mathcal{F} \oplus \mathcal{G} = \mathcal{H}$. The space $\mathcal{G}$ is a RKHS induced by a kernel K, while the space $\mathcal{F}$ can be expressed by a finite number of M basis functions $\psi_i : \mathbb{R}^D \to \mathbb{R}$, $i = 1, \ldots, M$. Given observational data $\mathcal{D} = \{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$ with $\boldsymbol{x}_t \in \mathbb{R}^D$ and $y_t \in \mathbb{R}$, a function $h \in \mathcal{H}$ can be estimated from

$$\min_{h \in \mathcal{H}} \quad \sum_{t=1}^{N} (y_t - h(\boldsymbol{x}_t))^2 + \lambda \|\mathscr{P}_{\mathcal{G}} h\|_{\mathcal{H}}^2 \tag{1}$$

where $\mathscr{P}_{\mathcal{G}}$ is the projector onto the reproducing kernel Hilbert Space (RKHS) $\mathcal{G}$ induced by K and $\lambda > 0$ a regularization constant. From a representer theorem [6] it follows that the solution to this problem can be expressed as

$$\hat{h}(\boldsymbol{z}) = \sum_{i=1}^{M} \vartheta_i \psi_i(\boldsymbol{z}) + \sum_{t=1}^{N} \alpha_t K(\boldsymbol{x}_t, \boldsymbol{z}) \tag{2}$$

with unknown coefficients $\vartheta_i, \alpha_t \in \mathbb{R}$. Using this finite dimensional representation the parameters can be obtained by solving the linear system [6, Eqs. (1.3.16) & (1.3.17)]

$$\begin{bmatrix} \boldsymbol{\Omega} + \lambda \boldsymbol{I} & \boldsymbol{\Psi}^T \\ \boldsymbol{\Psi} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\vartheta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}, \tag{3}$$

where $\Omega_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i, j = 1, \ldots, N$, $\boldsymbol{I}$ the N dimensional identity matrix, $\Psi_{ij} = \psi_i(\boldsymbol{x}_j)$ for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$, $\boldsymbol{\vartheta} = [\vartheta_1, \ldots, \vartheta_M]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_N]^T$. While the work of Wahba [6] was derived for Spline kernels, there are extensions to arbitrary reproducing kernels [14].

III. Imposing orthogonality constraints

This paper is concerned with the case when the parametric functions are (empirically) not orthogonal to the space induced by the kernel. In case one cannot form the direct sum of $\mathcal{F}$ and $\mathcal{G}$ and the estimation is still performed via (3) blindly without modification, certain drawbacks arise. This is due to a model ambiguity. Without

imposing further restrictions on ψ and K some behavior may be modeled by one part as well as the other.

In this section we state a constrained optimization approach with primal and Lagrange dual model representations which straightforwardly admits to incorporate additional constraints [12]. In Section V the results are stated in a functional setting.

A. Partially linear model: primal and dual problem

We start with a partially linear model similar to the one in the previous section in a LS-SVM formulation as given in [7]

$$\begin{aligned} \min_{\boldsymbol{\vartheta},\boldsymbol{w},b,e_t} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^{N} e_t^2 \\ \text{subject to } & y_t = \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}) + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}) + b + e_t,\ t = 1,\ldots,N. \end{aligned} \tag{4}$$

Here $b \in \mathbb{R}$, $\boldsymbol{w} \in \mathbb{R}^{n_h}$ are model parameters of the nonparametric model and $\boldsymbol{\psi}(\boldsymbol{x}) = [\psi_1(\boldsymbol{x}), \ldots, \psi_M(\boldsymbol{x})]^T$. The feature map $\boldsymbol{\varphi} : \mathbb{R}^D \to \mathbb{R}^{n_h}$ is defined via its inner product $\boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{y}) = K(\boldsymbol{x},\boldsymbol{y})$. A popular choice for the positive semidefinite kernel K are RBF kernels $K_{\mathrm{RBF}}(\boldsymbol{x},\boldsymbol{y}) = \exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/\sigma^2)$ with $\sigma > 0$. Note that the feature maps associated with nonlinear kernels can be very high dimensional or even infinite dimensional as is the case of the RBF kernel. As the feature map is implicitly defined through the choice of a positive definite kernel is does not need to be known explicitly. For $b = 0$ the Lagrange dual is then equal to (3) with $\gamma = 1/\lambda$ where $\boldsymbol{\alpha}$ are Lagrange multipliers related to the constraints in (4).

B. Parametric estimates in the case of violated assumptions

The model ambiguity between the parametric and nonparametric parts can be seen from the estimates obtained for $\boldsymbol{\vartheta}$. To simplify the presentation we assume that the measurements y_t are zero mean. Then b in (4) is equal to zero and the LS-SVM estimate coincides with the Spline one. Also assume that $\boldsymbol{\Psi}$ has full row rank and define $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_N)]$. Then the value for $\boldsymbol{\vartheta}$ obtained from (4) is

$$\hat{\boldsymbol{\vartheta}}_{\mathrm{PL}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}(\boldsymbol{y} - \boldsymbol{\Phi}^T\hat{\boldsymbol{w}}), \tag{5}$$

where $\hat{\boldsymbol{w}}$ is the estimate following from $(\boldsymbol{\Phi}\boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{\Phi}^T + \gamma^{-1}\boldsymbol{I})\boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{y}$, where $\boldsymbol{P}_{\Psi}^{\perp} = \boldsymbol{I} - \boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}$ is the projector onto the nullspace of $\boldsymbol{\Psi}$.

Under the assumptions used in previous works [6], [7] that the spaces $\mathcal{F}$ and $\mathcal{G}$ are orthogonal, the term $\boldsymbol{\Phi}\boldsymbol{\Psi}^T$ goes to zero as the number of samples N approaches infinity. Under the relaxed assumptions in [8] (nonparametric model is zero mean), [9] (no restrictions) this is not necessarily the case. For a finite amount of samples this does not hold even if the restrictive orthogonality assumption holds. Note that in the limit one has $\hat{\boldsymbol{\vartheta}}_{\mathrm{PL}} \to \hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$, i.e. the ordinary least squares (OLS) estimate.

Therefore in the case that $\boldsymbol{\Phi}\boldsymbol{\Psi}^T$ is empirically nonzero, the estimate $\hat{\boldsymbol{\vartheta}}_{\mathrm{PL}}$ depends not only on the data, but also on the kernel K and the regularization constant γ controlling the complexity of the nonparametric model. This is a manifestation of the ambiguity mentioned earlier.

Remark 1 (Ordinary least squares)**.** *Assume that the true underlying system generates data according to $y_t = \boldsymbol{\vartheta}_0^T\boldsymbol{\psi}(\boldsymbol{x}_t) + \varrho(\boldsymbol{x}_t) + \epsilon_t$, where $\varrho(\cdot)$ is the part of the underlying system that cannot be captured by the parametric function and ϵ_t is zero mean, Gaussian white noise with finite variance and independent of $\boldsymbol{x}_t$. Then the OLS estimate for $\boldsymbol{\vartheta}$ is $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = \boldsymbol{\vartheta}_0 + (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}(\boldsymbol{\epsilon} + \boldsymbol{\varrho})$, with $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$ and $\boldsymbol{\varrho} = [\varrho(\boldsymbol{x}_1), \ldots, \varrho(\boldsymbol{x}_N)]^T$. Note that the noise contribution goes to zero for increasing sample sizes. What remains is the contribution that depends only on the fraction of the true underlying system that is not captured by the parametric model.*

Remark 2 (Modeling of residuals)**.** *In view of Remark 1 it would seem beneficial to start with the OLS estimate $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}}$ and compute the residuals $r_t = y_t - \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t)$. However modeling these residuals with a nonparametric model gives rise to low prediction performances as shown in the experimental section.*

C. Imposing orthogonality

While a new model should have a comparable predictive performance as existing partially linear models, the goal of this paper is to ensure that the estimate of the parametric model given by (5) is always as good as the OLS estimate. This should hold for finite sample sizes as well as when the orthogonality assumption is not satisfied. Therefore comparing (5) with the OLS estimate $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$ the term $(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\Phi}^T\hat{\boldsymbol{w}}$ has to be zero. This holds if and only if

$$\boldsymbol{\Psi}(\boldsymbol{\Phi}^T\hat{\boldsymbol{w}} + \mathbf{1}\hat{b}) = \mathbf{0} \tag{6}$$

holds. As standard SVM and LS-SVM models in regression and classification problems contain a bias term b it has been reintroduced here. Therefore this relation is imposed as an additional constraint in problem (4). This can be interpreted as (empirical) orthogonality of the parametric and nonparametric models. Another interpretation is that the nonparametric model is constrained such that its predictions are uncorrelated with the predictions of any function with the chosen parametrization. For a given data set $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$ the modified optimization problem is then

$$\begin{aligned} \min_{\boldsymbol{\vartheta},\boldsymbol{w},b,e_t} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^{N} e_t^2 \\ \text{subject to} \quad & y_t = \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) + \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b + e_t, \quad t = 1,\ldots,N, \\ & \sum_{t=1}^{N}\psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b) = 0, \qquad i = 1,\ldots,M. \end{aligned} \tag{7}$$

Note that the basis functions $\boldsymbol{\psi}(\boldsymbol{x})$ need to be explicitly defined. On the other hand the feature map $\boldsymbol{\varphi}(\boldsymbol{x})$ is defined implicitly by a kernel function K at the dual level as is common in support vector machine methods.

D. Dual problem: model representation and estimation

As mentioned earlier in the case of the RBF kernel, the feature map $\boldsymbol{\varphi}$ is not only defined implicitly, but furthermore infinite dimensional. Thus it is in general not possible to directly solve the primal problem in (7). Therefore the problem is solved in the dual which is formalized in the following Lemma.

Lemma 1. *The solution to* (7) *is given by the linear system*

$$\begin{bmatrix} \boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{I} & \boldsymbol{\Omega}\boldsymbol{\Psi}^T & \mathbf{1} & \boldsymbol{\Psi}^T \\ \boldsymbol{\Psi}\boldsymbol{\Omega} & \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T & \boldsymbol{\Psi}\mathbf{1} & 0 \\ \mathbf{1}^T & \mathbf{1}^T\boldsymbol{\Psi}^T & 0 & 0 \\ \boldsymbol{\Psi} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ b \\ \boldsymbol{\vartheta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \\ 0 \\ \mathbf{0} \end{bmatrix}. \tag{8}$$

The variables $\boldsymbol{\beta} \in \mathbb{R}^M$ are the Lagrange multipliers for the newly introduced orthogonality constraints in (7) *and can be used to express the dual predictive model as $\hat{y} = \hat{h}(\boldsymbol{x}) = \sum_{t=1}^N \eta_t K(\boldsymbol{x}_t,\boldsymbol{x}) + \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}) + b$ with $\eta_t = \alpha_t + \boldsymbol{\beta}^T\boldsymbol{\psi}(\boldsymbol{x}_t)$.*

Proof: The Lagrangian for (7) is $\mathscr{L}(\boldsymbol{w},\boldsymbol{\vartheta},b,e_t,\alpha_t,\beta_i) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^N e_t^2 - \sum_{t=1}^N\alpha_t\ (\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) + b + e_t - y_t) - \sum_{i=1}^M\beta_i\sum_{t=1}^N\psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b)$. The corresponding KKT conditions for optimality are $\mathbf{0} = \frac{\partial\mathscr{L}}{\partial\boldsymbol{w}} = \boldsymbol{w} -$

$\sum_{t=1}^N(\alpha_t + \boldsymbol{\beta}^T\boldsymbol{\psi}(\boldsymbol{x}_t))\boldsymbol{\varphi}(\boldsymbol{x}_t)$, $\mathbf{0} = \frac{\partial \mathscr{L}}{\partial \boldsymbol{\vartheta}} = -\sum_{t=1}^N \alpha_t\boldsymbol{\psi}(\boldsymbol{x}_t)$, $0 = \frac{\partial \mathscr{L}}{\partial b} = -(\sum_{t=1}^N \alpha_t + \boldsymbol{\beta}^T\psi(\boldsymbol{x}_t))$, $0 = \frac{\partial \mathscr{L}}{\partial e_t} = \gamma e_t - \alpha_t\ \forall t$, $0 = \frac{\partial \mathscr{L}}{\partial \alpha_t} = y_t - \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) - \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) - b - e_t\ \forall t$ and $0 = \frac{\partial \mathscr{L}}{\partial \beta_i} = -\sum_{t=1}^N \psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b)$, $i = 1,\ldots,M$. Rewriting the conditions for $\boldsymbol{w}$ and e_t and substituting them into $\partial\mathscr{L}/\partial\alpha_t$ yields $y_t = \sum_{k=1}^N(\alpha_k + \boldsymbol{\beta}^T\boldsymbol{\psi}(\boldsymbol{x}_k))K(\boldsymbol{x}_k,\boldsymbol{x}_t) + \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) + b + \frac{1}{\gamma}\alpha_t$. Substitution of $\boldsymbol{w}$ into $\partial\mathscr{L}/\partial\beta_i$ yields $\sum_{t=1}^N\psi_i(\boldsymbol{x}_t)\sum_{k=1}^N(\alpha_k + \boldsymbol{\beta}^T\boldsymbol{\psi}(\boldsymbol{x}_k))K(\boldsymbol{x}_k,\boldsymbol{x}_t) + b = 0$. Here, the inner products $\boldsymbol{\varphi}(\boldsymbol{x}_k)^T\boldsymbol{\varphi}(\boldsymbol{x}_t)$ were replaced by a positive definite kernel $K(\boldsymbol{x}_k,\boldsymbol{x}_t)$ using the kernel trick. The rewritten KKT conditions $\partial\mathscr{L}/\partial\alpha_t$ and $\partial\mathscr{L}/\partial\beta_i$ together with $\partial\mathscr{L}/\partial b$ and $\partial\mathscr{L}/\partial\boldsymbol{\vartheta}$ correspond to the dual system (8). ∎

IV. Alternative estimation schemes and representations

A. Separation principle

Due to the orthogonality constraint imposed on the kernel based model the estimation problem can be decoupled.

Lemma 2. *Under the assumption that $\boldsymbol{\Psi}$ has full row rank, the estimation of the model can be performed in two stages as follows:*

1) *The parameter vector $\boldsymbol{\vartheta}$ can be estimated using ordinary least squares $\hat{\boldsymbol{\vartheta}}_{\mathrm{OLS}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$.*
2) *The kernel based model can then be identified by solving the reduced linear system*

$$\begin{bmatrix}\boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{P}_{\Psi}^{\perp} & \mathbf{1}\\ \mathbf{1}^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\eta}\\ b\end{bmatrix} = \begin{bmatrix}\boldsymbol{r}\\ 0\end{bmatrix} \tag{9}$$

with $\boldsymbol{r} = \boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{\Psi}^T\hat{\boldsymbol{\vartheta}}$ and as before $\boldsymbol{P}_{\Psi}^{\perp} = \boldsymbol{I} - \boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}$.

Proof: Let $\mathcal{R}_1$ to $\mathcal{R}_4$ denote the block rows of (8). Isolating $\boldsymbol{\Omega}\boldsymbol{\alpha}$ in $\mathcal{R}_1$ and substituting this into $\mathcal{R}_2$ yields $\boldsymbol{\Psi}\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\vartheta} + \gamma^{-1}\boldsymbol{\Psi}\boldsymbol{\alpha}$. Exploiting $\mathcal{R}_4$ we obtain $\boldsymbol{\vartheta} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{y}$. Now consider a change of variables $\boldsymbol{\eta} = \boldsymbol{\alpha} + \psi^T\boldsymbol{\beta}$. Substitution of this relation into $\mathcal{R}_4$ we obtain $\boldsymbol{\beta} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\eta}$. Stating $\mathcal{R}_3$ in terms of $\boldsymbol{\eta}$ yields $\mathbf{1}^T\boldsymbol{\eta} = 0$ and $\mathcal{R}_1$ can be rewritten as $(\boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{P}_{\Psi}^{\perp})\boldsymbol{\eta} + \mathbf{1}b = \boldsymbol{P}_{\Psi}^{\perp}\boldsymbol{y}$. ∎

Note that $\mathcal{R}_2$ in terms of $\boldsymbol{\eta}$ reads $\boldsymbol{\Psi}(\boldsymbol{\Omega}\boldsymbol{\eta} + \mathbf{1}b) = \mathbf{0}$. This nicely illustrates that the correlation between the kernel based model $\boldsymbol{\Omega}\boldsymbol{\eta} + \mathbf{1}b$ and the basis functions $\boldsymbol{\Psi}$ is zero on the estimation data.

B. Equivalent kernel

For simplicity of the presentation we will assume that the bias term b is zero in this subsection. This is no limitation as a static offset can be included in the parametric basis. In fact if this is done the orthogonality constraint will ensure that b is zero.

Lemma 3. *Under the assumption that $\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T$ is invertible, problem* (7) *can be solved equivalently as*

$$\begin{aligned}&\min_{\boldsymbol{w},e_t}\ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma\sum_{t=1}^N e_t^2\\ &\text{subject to}\quad y_t = \boldsymbol{w}^T\boldsymbol{\varphi}_{\mathrm{eq}}(\boldsymbol{x}_t) + e_t,\quad t = 1,\ldots,N,\end{aligned} \tag{10}$$

with an equivalent feature map $\boldsymbol{\varphi}_{\mathrm{eq}}(\cdot) = \boldsymbol{P}_{\Phi\Psi}^{\perp}\boldsymbol{\varphi}(\cdot)$ where $\boldsymbol{P}_{\Phi\Psi}^{\perp}$ is defined as $\boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{\Psi}^T\ (\boldsymbol{\Psi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\Phi}^T$ with $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\varphi}(\boldsymbol{x}_N)]$.

This primal formulation can be solved using the dual linear system

$$\begin{bmatrix}\boldsymbol{\Omega}_{\mathrm{eq}} + \gamma^{-1}\boldsymbol{I} & \boldsymbol{\Psi}^T\\ \boldsymbol{\Psi} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\vartheta}\end{bmatrix} = \begin{bmatrix}\boldsymbol{y}\\ \mathbf{0}\end{bmatrix},$$

where $(\boldsymbol{\Omega}_{\mathrm{eq}})_{ij} = K_{\mathrm{eq}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for $i,j = 1,\ldots,N$. The equivalent kernel is given by

$$K_{\mathrm{eq}}(\boldsymbol{x},\boldsymbol{y}) = K(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{k}(\boldsymbol{x})^T\boldsymbol{\Psi}^T(\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{k}(\boldsymbol{y}) \tag{11}$$

with $\boldsymbol{k}(\boldsymbol{z}) = [K(\boldsymbol{x}_1,\boldsymbol{z}),\ldots,K(\boldsymbol{x}_N,\boldsymbol{z})]^T$. This model can be evaluated at a new point $\boldsymbol{z}$ through $\hat{y} = \hat{h}(\boldsymbol{z}) = \sum_{t=1}^N K_{\mathrm{eq}}(\boldsymbol{x}_t,\boldsymbol{z})\alpha_t + \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{z})$.

Proof: Substitution of the expansion for $\boldsymbol{w}$ obtained from the KKT conditions in the proof of Lemma 1 into the KKT condition for $\boldsymbol{\beta}$ gives $\boldsymbol{\Psi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{\alpha} + \boldsymbol{\Psi}\boldsymbol{\beta}) = 0$. Then, assuming invertibility, the Lagrange multipliers $\boldsymbol{\beta}$ can be expressed as $\boldsymbol{\beta} = -(\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T)^{-1}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\alpha}$. Using this solution in the expansion of $\boldsymbol{w}$ yields $\boldsymbol{w} = \boldsymbol{P}_{\Phi\Psi}^{\perp}\boldsymbol{\Phi}\boldsymbol{\alpha}$. Reading this columnwise for the columns of $\boldsymbol{\Phi}$ the equivalent feature map can be extracted. The equivalent kernel follows directly from $K_{\mathrm{eq}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\varphi}_{\mathrm{eq}}(\boldsymbol{x})^T\boldsymbol{\varphi}_{\mathrm{eq}}(\boldsymbol{y}) = \boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{P}_{\Phi\Psi}^{\perp}\boldsymbol{\varphi}(\boldsymbol{y})$. ∎

V. Characterization in RKHSs

In this section we detail an alternative derivation based on the theory of RKHS's. Differently from [6], [14] we do not assume the existence of a direct sum decomposition between the spaces $\mathcal{F}$ and $\mathcal{G}$. Rather we construct a space by forming the direct sum between $\mathcal{F}$ and a certain suitable subspace $\mathcal{G}_{\mathrm{eq}}$ of $\mathcal{G}$.

More precisely the argument starts from an arbitrary RKHS $\mathcal{G}$ of real-valued functions on $\mathbb{R}^D$ and a finite dimensional space $\mathcal{F}$ spanned by the prescribed set of real-valued functions $\psi_i : \mathbb{R}^D \to \mathbb{R}$, $i = 1,\ldots,M$. We denote by K the reproducing kernel of $\mathcal{G}$ and by $\langle\cdot,\cdot\rangle_K$ the corresponding inner product. Furtheron we show that $\mathcal{G}_{\mathrm{eq}}$ is a RKHS and derive its reproducing kernel K_{eq}. As in the preceding sections the idea is based on an orthogonality criterion formulated upon the empirical sample. We begin by recalling the following result.

Lemma 4. *Kernel of a Closed Subspace [15, Theorem 11] Let $\mathcal{V}$ be a closed subspace of a RKHS $\mathcal{G}$. Then $\mathcal{V}$ is a RKHS and its kernel K' is given by*

$$K'(\boldsymbol{x},\boldsymbol{y}) = \left[\mathscr{P}_{\mathcal{V}}K(\cdot,\boldsymbol{y})\right](\boldsymbol{x}) \tag{12}$$

where $\mathscr{P}_{\mathcal{V}}$ denotes the projection operator onto $\mathcal{V}$.

We now introduce the sampling operator $\mathscr{S} : \mathcal{L}_2 \to \mathbb{R}^N$, defined entry-wise as $(\mathscr{S})_t := f(\boldsymbol{x}_t)$ for $t = 1,\ldots,N$ and the linear operator $\mathscr{C} : \mathcal{L}_2 \to \mathbb{R}^M$ defined as $\mathscr{C}f = \boldsymbol{\Psi}\mathscr{S}f$ where $f \in \mathcal{L}_2$ where $\mathcal{L}_2$ is the space of square integrable functions[1]. It is not difficult to check that its adjoint operator $\mathscr{C}^* : \mathbb{R}^M \to \mathcal{G}$ is defined by $\mathscr{C}^*\boldsymbol{\beta} := \sum_{i=1}^M\sum_{t=1}^N \beta_i\psi_i(\boldsymbol{x}_t)K(\boldsymbol{x}_t,\cdot)$, where $\boldsymbol{\beta} \in \mathbb{R}^M$. Now $\mathrm{range}(\mathscr{C}^*)$ is a finite dimensional subspace of $\mathcal{G}$, $\mathrm{null}(\mathscr{C})$ is a closed subspace and

$$\mathcal{G} = \mathrm{range}(\mathscr{C}^*) \oplus \mathrm{null}(\mathscr{C}).$$

By Lemma 4 $\mathrm{null}(\mathscr{C})$ is a RKHS and its kernel function is solely given in terms of K and of $\mathscr{P}_{\mathrm{null}(\mathscr{C})}$ the projector onto the nullspace of $\mathscr{C}$. We have now that $\mathscr{P}_{\mathrm{null}(\mathscr{C})} : \mathcal{G} \to \mathcal{G}$ is given by (see e.g. [17, Theorem 1, Chapter 6.9])

$$\mathscr{P}_{\mathrm{null}(\mathscr{C})}g = (\mathscr{I}_{\mathrm{id}} - \mathscr{C}^*\left(\mathscr{C}\mathscr{C}^*\right)^{-1}\mathscr{C})g = g - \sum_{t=1}^N\left(\boldsymbol{\Psi}^T\boldsymbol{\Omega}_{\Psi}^{-1}\boldsymbol{\Psi}\mathscr{S}g\right)_t K(\boldsymbol{x}_t,\cdot)$$

where $\boldsymbol{\Omega}_{\Psi} := \mathscr{C}\mathscr{C}^* = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}^T$ is assumed to be a full-rank matrix as in Lemma 3. We denote by $\mathscr{I}_{\mathrm{id}}$ the identity operator. By equation (12)

[1]We implicitly assume that $\mathcal{F},\mathcal{G} \subset \mathcal{L}_2$ which, for $\mathcal{G}$, is true under mild conditions on the kernel K [16].

we now have that the reproducing kernel of $\mathrm{null}(\mathscr{C})$, denoted by K_{eq}, is given by $K_{\mathrm{eq}}(\boldsymbol{x},\boldsymbol{y}) = \left[\mathscr{P}_{\mathrm{null}(\mathscr{C})}K(\cdot,\boldsymbol{y})\right](\boldsymbol{x})$ which corresponds to (11). Denote now $\mathrm{null}(\mathscr{C})$ by $\mathcal{G}_{\mathrm{eq}}$. We have the following.

Proposition 1. *For any $f_1 \in \mathcal{F}$ and any $g_1 \in \mathcal{G}_{\mathrm{eq}}$ the function $h = f_1 + g_1$ admits a unique decomposition under the assumption that $\boldsymbol{\Psi}$ has full rank.*

Proof: Suppose that there are $f_2 \in \mathcal{F}$ and $g_2 \in \mathcal{G}$ with $f_1 \neq f_2$ such that $h = f_2 + g_2$. Note that f_1 and f_2 can be parametrized as $f_1(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\vartheta}$ and $f_2(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\vartheta}'$ respectively. As $f_1 \neq f_2$ also $\boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}'$. By assumption we need to have $(f_1 + g_1)(\boldsymbol{x}) = (f_2 + g_2)(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^D$. By construction we have $\mathscr{C}g_1 = \mathscr{C}g_2 = \mathbf{0}_M$. Therefore we have $\boldsymbol{\Psi}\mathscr{S}f_1 = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\vartheta} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\boldsymbol{\vartheta}' = \boldsymbol{\Psi}\mathscr{S}f_2$. As $\boldsymbol{\Psi}$ has full rank it follows $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}'$ which is a contradiction. ∎

It now follows from Proposition 1 that we can write

$$\mathcal{H} := \mathcal{F} \oplus \mathcal{G}_{\mathrm{eq}}.$$

For any $h_1 = \boldsymbol{\vartheta}^T\boldsymbol{\psi} + g_1$ and $h_2 = \boldsymbol{\vartheta}'^T\boldsymbol{\psi} + g_2$ we can define the inner product

$$\langle h_1, h_2\rangle := \boldsymbol{\vartheta}^T\boldsymbol{\vartheta}' + \langle g_1, g_2\rangle_K$$

that turns $\mathcal{G}_{\mathrm{eq}}$ and $\mathcal{F}$ into orthogonal complements. The corresponding norm is $\|h\| := \sqrt{\langle h, h\rangle}$. Denoting by $\mathscr{P}_{\mathcal{G}_{\mathrm{eq}}}$ the projection onto $\mathcal{G}_{\mathrm{eq}}$ by the representer theorem, the solution of (1) is given by (2) where $\boldsymbol{\vartheta} \in \mathbb{R}^M$ and $\boldsymbol{\alpha} \in \mathbb{R}^N$ can be found in closed form by solving the system of linear equations (3). Since in the current setting we have that $\boldsymbol{\Psi}\boldsymbol{\Omega}_{\mathrm{eq}}\boldsymbol{\alpha} = \mathbf{0}_M$ similar arguments as in Lemma 2 can be used to establish that $\boldsymbol{\vartheta}$ is given by $\boldsymbol{\vartheta}_{\mathrm{OLS}}$. Once the parametric part has been computed one can also find the non-parametric part starting from the residuals $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{\Psi}^T\boldsymbol{\vartheta}_{\mathrm{OLS}}$ and solve $(\boldsymbol{\Omega}_{\mathrm{eq}} + \lambda\boldsymbol{I})\boldsymbol{\alpha} = \boldsymbol{r}$. The final model (for out-of sample prediction) is given by (2) in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\vartheta}$ and the equivalent kernel function K_{eq} that is given in terms of the original kernel K, the training data $\{\boldsymbol{x}_t\}_{t=1}^N$ and the parametric basis functions $\{\psi_i\}_{i=1}^N$.

VI. Extension to different loss functions

The outlined procedure can be extended to other loss functions. The ϵ-insensitive loss for example is defined as $l_\epsilon(x) = 0$ for$|x| \leq \epsilon$ and $l_\epsilon(x) = |x| - \epsilon$ otherwise. If this loss is used instead of the L_2-loss as in (7) a SVM [11] is obtained. The primal formulation is

$$\begin{aligned}
\min_{\boldsymbol{\vartheta},\boldsymbol{w},b,\xi_t} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{t=1}^{N}\xi_t && (13)\\
\text{subject to} \quad & |y_t - \boldsymbol{\vartheta}^T\boldsymbol{\psi}(\boldsymbol{x}_t) - \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) - b| \leq \epsilon + \xi_t,\ \xi_t \geq 0, \\
& \qquad\qquad t = 1,\ldots,N, \\
& \sum_{t=1}^{N}\psi_i(\boldsymbol{x}_t)(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + b) = 0, \qquad i = 1,\ldots,M.
\end{aligned}$$

The Lagrange dual can be obtained by solving the KKT system

$$\begin{aligned}
\max_{\alpha_t,\beta_i} \quad & \boldsymbol{y}^T\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha} + \boldsymbol{\Psi}^T\boldsymbol{\beta})^T\boldsymbol{\Omega}(\boldsymbol{\alpha} + \boldsymbol{\Psi}^T\boldsymbol{\beta}) - \epsilon\mathbf{1}^T|\boldsymbol{\alpha}| && (14)\\
\text{subject to} \quad & -C \leq \alpha_t \leq C, \qquad\qquad t = 1,\ldots,N, \\
& \boldsymbol{\Psi}\boldsymbol{\alpha} = \mathbf{0}, \qquad \mathbf{1}^T(\boldsymbol{\alpha} + \boldsymbol{\Psi}^T\boldsymbol{\beta}) = 0.
\end{aligned}$$

where $|\boldsymbol{\alpha}|$ is understood as the elementwise absolute value. The predictive model is identical to the one stated in Lemma 1. The primal variables b and ϑ_i can for example be obtained from a interior point solver and correspond to the dual variables of the equality constraints in (14).

Extensions to Huber's robust loss function and other convex loss functions can be derived in a similar fashion.

VII. Experiments

A. Experimental setup

In this section several modeling approaches are compared for simulated as well as for real life data. The following model structures using NARX regressors are considered

- purely parametric models (`LIN`),
- purely nonparametric black box models (`LS-SVM`, `SVM`),
- parametric model + nonlinear black box model for residuals (`RES (LS-SVM)`, cf. Remark 2),
- partially linear model without orthogonality constraints (`PL (LS-SVM)`, cf. Eq. (4)) and
- partially linear models with orthogonality constraints (`OPL (LS-SVM)`, cf. Eq. (7) and `OPL (SVM)`, cf. Eq. (13)).

For all experiments the parametric part is taken linear $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{x}$ while the nonparametric contribution is modeled with a RBF kernel. Considering the results from [18] we note that the assumption (direct sum decomposition) for classical partially linear models `PL` is satisfied as linear functions (and more generally polynomials) and the space induced by the Gaussian kernel can be combined by forming the direct sum of both. Therefore the proposed constraint would not be needed, but we show that in case of small but realistic sample sizes (> 1000 samples) we observe an improvement by using the constraint.

For the simulation data we report the distance of the estimated parametric model to the "true" one in addition to the root mean squared error (RMSE) on independent data. For the comparison of the linear submodels identified by the different models, a metric for ARMA systems proposed by Martin [19] is used. This metric compares the spectra of linear systems and as such is independent of their parametrization. For elaborate arguments why this metric is better in terms of system-theoretic properties than e.g. comparing AR coefficients please consult the reference. For two stable AR systems S_A, S_B with poles p_i^A, p_j^B, $i = 1,\ldots,P_A, j = 1,\ldots,P_B$ the metric can be computed as $d(S_A,S_B) = \ln(\prod_{i=1}^{P_A}\prod_{j=1}^{P_B}|1 - p_i^A p_j^{B^*}|^2/(\prod_{i,j=1}^{P_A}(1 - p_i^A p_j^{A^*})\prod_{i,j=1}^{P_B}(1 - p_i^B p_j^{B^*})))$ where x^* denotes the complex conjugate of x. For the generalization to stable ARX systems we refer to [19].

The tuning parameters of the nonlinear models (kernel parameters and regularization constant) are selected according to one-step ahead prediction performance on a validation set using grid search.

B. Mass-Spring-Damper system

To test the proposed method we consider a simple mass spring damper system with two masses and springs with cubic nonlinearities as depicted in Figure 2. The system is described by $m_1\ddot{s}_1 = -L_1 f_1(s_1) + L_2 f_2(s_2 - s_1) - C_1\dot{s}_1 + C_2(\dot{s}_2 - \dot{s}_1)$, $m_2\ddot{s}_2 = -L_2 f_2(s_2 - s_1) - C_2(\dot{s}_2 - \dot{s}_1) + F_u$, $y = s_2$, $f_1(x) = x + 0.02x^3$ and $f_2(x) = x + 0.005x^3$. The states s_1 and s_2 are the displacement of mass 1 and mass 2 from rest respectively. The constants are chosen as $m_1 = 5\,\mathrm{kg}$, $m_2 = 0.1\,\mathrm{kg}$, $L_1 = 1.5\,\mathrm{N/m}$, $L_2 = 0.5\,\mathrm{N/m}$, $C_1 = 0.3\,\mathrm{Ns/m}$ and $C_2 = 0.05\,\mathrm{Ns/m}$. The output y is sampled with $4\,\mathrm{Hz}$ and the excitation force F_u is given by a Gaussian white noise process u_t. The equations are simulated using MATLAB's `ode45`, a zero order hold stage for the input signal and zero initial conditions. For each realization of the input signal and every amplitude value, 4000 samples are generated. The first 400 samples of each data set are discarded to exclude transient behavior. To compare measurements with different amounts of nonlinearity, excitation amplitudes of $\{0.1, 0.3, 0.5, 0.75, 1, 1.25, 1.5, 2\}$ are tested. This accounts to different fractions of the system not being captured by the parametric (linear) model. For each amplitude the estimation is performed for

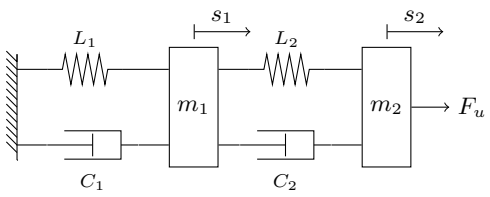


Fig. 2. Mass-spring-damper system with nonlinear springs.

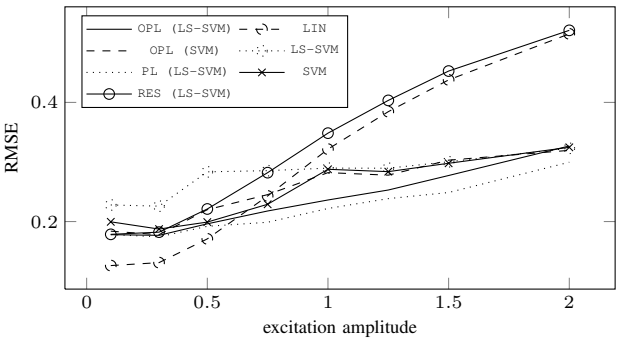


Fig. 3. Prediction performance of different model structures for the mass spring damper system on an independent test set as a function of excitation amplitude. The root mean squared error for simulated outputs averaged over ten input realizations is being reported.

ten realizations of the input signal. All data sets are split in three parts of equal size for model estimation, model validation and test.

Figure 3 depicts the simulation performance for different model structures. The model orders are as follows $p = q = 13$ for `LIN` and $p = q = 10$ for all other models, where the input variable $\boldsymbol{x}_t$ has NARX structure with $\boldsymbol{x}_t = [y_{t-1}, \ldots, y_{t-p}, u_{t-1}, \ldots, u_{t-q}]^T$. For a fair comparison the input sequence u_t as well as the output sequence y_t are normalized to unit variance for each amplitude. As it can be expected the linear model is best for small amplitudes but degrades quickly as the amplitude increases. The nonlinear black box models `LS-SVM` and `SVM` are the best models for "large" amplitudes. Modeling the residuals is clearly inferior to all other approaches as it is never better than the linear model. Finally the partially linear structures give improved prediction performance for a number of amplitudes. Imposing orthogonality constraints yields slightly worse predictive performance than using a partially linear structure without constraints.

Setting $f_1(x) = f_2(x) = x$ a "true" linear system can be obtained.

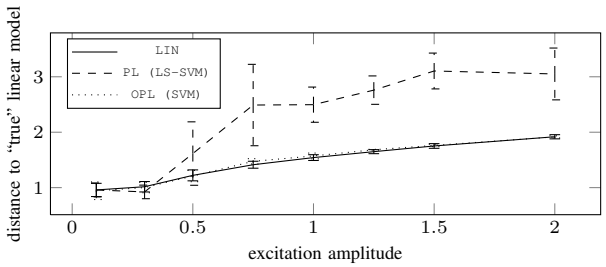


Fig. 4. Distance of the estimated parametric submodel to the "true" linear model for the mass-spring damper system for several model structures as a function of the excitation amplitude. The models `LIN`, `OPL (LS-SVM)` and `RES (LS-SVM)` have identical distances. Error bars indicate the standard deviation for ten realizations of the input signal. The comparison is based on the ARMA metric proposed in [19].

TABLE I
PERFORMANCE OF DIFFERENT MODEL STRUCTURES ON A SUBSAMPLE OF THE WIENER-HAMMERSTEIN BENCHMARK DATASET. THE PERFORMANCE IS REPORTED AS ROOT MEAN SQUARE ERROR FOR SIMULATION VALUES.

model	training data	validation data	testing data
`LIN`	0.164	0.230	0.168
`LS-SVM`	0.067	0.183	0.135
`SVM`	0.053	0.139	0.079
`RES (LS-SVM)`	0.162	0.222	0.155
`PL (LS-SVM)`	0.037	0.133	0.068
`OPL (LS-SVM)`	0.091	0.166	0.094
`OPL (SVM)`	0.051	0.140	0.071

Using MATLAB's `c2d` a discrete time transfer function is computed from the continuous time system described above. The "true" linear system is of order four while the identified models are of orders thirteen and ten. Therefore the ARMA metric described earlier is used to compare the estimated with the "true" model. The result is shown in Figure 4. It can be seen that for small amplitudes the partially linear model is close to the linear estimate. Yet for larger amplitudes it quickly diverges away from the linear estimate and not only has a larger distance from the "true" linear model but the estimates also have a larger variance. While both partially linear structures move away from the "true" linear model as the excitation amplitude increases, the models with orthogonality constraints degrade more gracefully than those without.

In conclusion, both partial linear structures yield better predictive performance than a parametric or nonparametric model alone. The unique advantage of using orthogonality constraints is that the estimate for parametric part has a much smaller variance and also is much closer to the "true" underlying model as seen in Fig. 4.

C. Wiener-Hammerstein benchmark data

To test on a real life data set, we consider the data of the Wiener-Hammerstein benchmark [20] at SYSID2009. As it just serves as a test problem, only a small subset of the 188,000 measured samples is taken. For model estimation and validation 2,000 samples each are used. The prediction performance is computed on 5,000 independent samples.

The results are shown in Table I, the reported values are for outputs normalized to unit variance. It can be seen that roughly the same behavior as with the simulated data is present. The partially linear models improve over the purely linear and purely nonlinear models. Using orthogonality constraints results in a slight degradation of prediction performance as observed in the previous section. In tests with small estimation sample sizes the prediction performance of of `OPL (LS-SVM)` is better than `PL (LS-SVM)` up to a sample size of about 200. Figure 5 shows transfer functions derived from the parametric model parts. The estimation is done for 20 subsamples of size 300. One can observe that the `OPL` structures are much closer to the reference spectrum given by the Best Linear Approximation than the `PL` model. Also their variability is much smaller.

VIII. CONCLUSIONS

Is was shown that the incorporation of orthogonality constraints into partially linear models is able to improve the parametric estimate for finite sample sizes. This is achieved by ensuring that the estimate is equal to the OLS estimate. Numerical examples illustrate that the proposed model structure is able to improve over a purely parametric or a purely nonparametric model. The parametric estimate obtained when considering the additional constraints is better in mean and

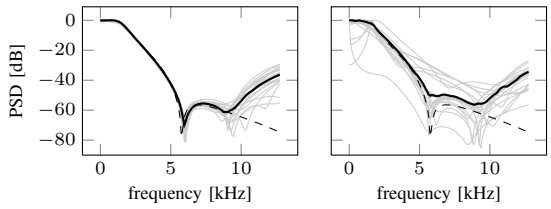


Fig. 5. Estimated transfer functions for the proposed model with orthogonality constraints (`OPL (LS-SVM)`, left) which improve over classical partially linear models (`PL (LS-SVM)`, right). The estimates are obtained for 20 different training samples of size of 300 taken out of the Wiener-Hammerstein benchmark dataset. Each light gray line corresponds to one realization, the thick black line indicates the median value while the dashed line gives the Best Linear Approximation [21] estimated on the full dataset as reference. The excitation signal was subject to low-pass filtering with a cut-off frequency of 10 kHz.

much better in variance than existing techniques. This ensures that interpretabilty of the parametric model is not lost in a partially linear setting. Therefore in many cases where partially linear cases are applied inclusion of the proposed constraint is the better choice. For small sample sizes there is a small gain in predictive performance while otherwise there is a slight loss.

REFERENCES

[1] W. Härdle, H. Liang, and J. Gao, *Partially linear models*. Springer, 2000.

[2] L. Ljung, *System identification: Theory for the User*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1999.

[3] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, pp. 1691–1724, Dec. 1995.

[4] G. Wahba, "Partial spline models for the semiparametric estimation of functions of several variables," in *Statistical analysis of time series*. Tokyo: Institute of Mathemical Statistics, 1984, pp. 319–329.

[5] P. Speckman, "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 50, no. 3, pp. 413–436, 1988.

[6] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990.

[7] M. Espinoza, J. A. K. Suykens, and B. De Moor, "Kernel Based Partially Linear Models and Nonlinear Identification," *IEEE Transactions on Automatic Control*, vol. 50, pp. 1602–1606, Oct. 2005.

[8] Y.-F. Li, L.-J. Li, H.-Y. Su, and J. Chu, "Least Squares Support Vector Machine Based Partially Linear Model Identification," *Lecture Notes in Computer Science*, vol. 4113/2006, pp. 775–781, 2006.

[9] Y.-L. Xu and D.-R. Chen, "Partially-Linear Least-Squares Regularized Regression for System Identification," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2637–2641, 2009.

[10] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, 2002.

[11] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[12] J. A. K. Suykens, C. Alzate, and K. Pelckmans, "Primal and dual model representations in kernel-based learning," *Statistics Surveys*, vol. 4, pp. 148–183, Aug. 2010.

[13] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481 –1497, Sept. 1990.

[14] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, pp. 416–426, 2001.

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

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

[17] D. Luenberger, *Optimization by Vector Space Methods*. Wiley-IEEE, 1998.

[18] M. Ha Quang, G. Pillonetto, and A. Chiuso, "Nonlinear system identification via gaussian regression and mixtures of kernels," in *Proceedings of the 15th IFAC Symposium on System Identification*, Saint-Malo, France, Aug. 2009, pp. 528–533.

[19] R. J. Martin, "A Metric for ARMA Processes," *IEEE Transactions on Signal Processing*, vol. 48, no. 4, pp. 1164–1170, 2000.

[20] J. Schoukens, J. A. K. Suykens, and L. Ljung, "Wiener-Hammerstein benchmark," in *Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009)*, Saint-Malo, France, 2009.

[21] J. Schoukens, J. G. Nemeth, P. Crama, Y. Rolain, and R. Pintelon, "Fast approximate identification of nonlinear systems," *Automatica*, vol. 39, pp. 1267–1274, 2003.