A Perturbation Analysis using Second Order Cone Programming for Robust Kernel Based Regression

Tillmann Falck, Marcelo Espinoza, Johan A. K. Suykens, Bart De Moor
K.U. Leuven, ESAT-SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.
{tillmann.falck,johan.suykens}@esat.kuleuven.be

***Abstract*—The effects of perturbations on the regression variables in nonlinear black box modeling are analysed using kernel based techniques. Starting from a linear regression problem in the primal space a robust primal formulation is obtained in form of a Second Order Cone Program (SOCP). The underlying worst case assumption corresponds to an additional regularization term in the primal, regularizing the subspace spanned by the first derivatives of the learned nonlinear model. This information is transferred from the primal domain into a dual formulation. The equivalent least squares problem is derived where the assumption is incorporated into a modified kernel matrix. One-step ahead prediction rules directly arise from the dual models and explicitly incorporate the imposed assumptions. The results are applied to study the influence of different inputs and different kernel choices on the prediction performance.**

I. Introduction

The estimation of a black-box model in order to produce precise forecasts starting from a set of observations is a common practice in system identification [1]. According to the nature of the problem different model structures can be used like autoregressive (ARX), output-error (OE) or errors-in-variables. In nonlinear system identification [2], [3] kernel based estimation techniques, such as Support Vector Machines (SVMs) [4], Least Squares Support Vector Machines (LS-SVMs) [5], [6] and Splines [7] have shown to be capable nonlinear black-box regression methods. Imposing a model structure in a nonlinear kernel setting has been tried for example in [8] for a NOE model, yielding a nonconvex recurrent formulation or in [9] for partially linear models.

In this work we are going to consider the case that the regressors are subject to an unknown but bounded perturbation. Except from boundedness we impose no further restrictions on this perturbation. For linear systems this is analyzed in [10] by using Second Order Cone Programs (SOCPs) and SemiDefinite Programs (SDPs) yielding robust linear models. A similar result is obtained in [11] using mostly algebraic relations. Both works relate the robust setting which uses a bound on the perturbation to classical regularization schemes for LS and to Total Least Squares (TLS) [12]. Extensions for nonlinear systems appear first in [13] where an iterative algorithm for nonlinear in parameters models is developed. In [14], [15] robust parametric nonlinear models are identified using SOCPs. For linear in the parameters models the results are exact except for a first order approximation of the nonlinear basis functions, for models that are nonlinear in the parameters upper bounds are derived. The first order approximation will also be used in a similar fashion later in this work. All these works are either restricted to linear or parametric models.

Support Vector techniques allow the estimation of non-parametric nonlinear models by solving convex problems. In particular the solution of LS-SVMs is given as a system of linear equations. The main result of this work is formulated as a SOCP. In the primal domain a linear model is estimated in the so-called feature space. By deriving the dual it is possible to replace all occurrences of the usually unknown feature map φ by that of a chosen positive definite kernel function K. By modifying the primal problem prior knowledge can be incorporated into the identified model in the dual domain. Robust results based on SOCPs exist for example in [16] in a probabilistic context. Results for a deterministic setting are derived in [17]. Although robust kernel based models are identified they either not formulated in way that allows to use it for prediction, or omit the some power of the identified model by switching to a parametric formulation. A result for the related TLS in a SVM context can be found in [18]. The major properties of the results derived in here are nonparametric, for nonlinear modeling, in a deterministic framework, convex, having explicit expressions for prediction, and incorporating the primal-dual knowledge into the model.

This paper is organized as follows. The problem setting is outlined in Section II. In Section III the underlying approximations are stated and a robust primal problem is derived by stating it as a SOCP. Then the dual for the SOCP is derived and expressed in terms of the kernel function. The next result is the prediction equation that makes explicit use not only of the identified parameters but of the model as well. In Section IV the SOCP is recasted into a least squares problem for computational efficiency. The prediction equation for this solution is obtained as well. In Section V we give numerical results for the robust versions of the developed kernel based model. Special attention is paid to a perturbation analysis of LS-SVM. Finally the conclusions are given in Section VI.

II. Problem Statement

Given a system with one output y_k and M inputs $u_k^{(1)}, \ldots, u_k^{(M)}$ a regression vector for an ARX model can be defined as $\boldsymbol{x}_k = [y_{k-1}, \ldots, y_{k-p}, u_k^{(1)}, \ldots, u_{k-q_1}^{(1)},$

$\ldots, u_k^{(M)}, \ldots, u_{k-q_M}^{(M)}]^T$, where $y_k, u_k^{(1)}, \ldots, u_k^{(M)} \in \mathbb{R}$, $p \geq 1$ and $q_1, \ldots, q_M \geq 0$. For given values $b, \boldsymbol{w}, \boldsymbol{\varphi}$ a nonlinear predictive model with predicted output $\hat{y}$ is given by

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}) + b. \tag{1}$$

The nonlinearity is modeled by the mapping $\boldsymbol{\varphi} : \mathbb{R}^n \to \mathbb{R}^{n_h}$ where $n = p + \sum_{m=1}^{M}(q_m + 1)$ and the dimension of the image space is typically $n_h \gg n$ and may even become infinite dimensional. In kernel based techniques the mapping $\boldsymbol{\varphi}$ is usually not known explicitly but only its inner products which are equal to a chosen kernel function.

To identify the model on a given dataset $\{\boldsymbol{x}_k, y_k\}_{k=1}^N$ a cost function is defined in the primal space [5]

$$\min_{\boldsymbol{w},b,\epsilon_k} \lambda \left\|\boldsymbol{w}\right\|^2 + \left\|\boldsymbol{\epsilon}\right\|^2 \text{ subject to } y_k = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k, \quad k = 1, \ldots, N. \tag{2}$$

This is a regularized linear regression in the feature space, in this form also known as Least Squares Support Vector Machine Regression. The squared l_2 norms will be replaced by l_2 for most of this paper. The regression error ϵ_k is assumed to be zero mean, i.i.d. and to have bounded variance. Computing the Lagrange dual of this problem, one KKT condition yields an expansion of $\boldsymbol{w}$ in terms of the dual variables $\boldsymbol{\alpha}$

$$\boldsymbol{w} = \sum_{k=1}^{N} \alpha_k \boldsymbol{\varphi}(\boldsymbol{x}_k). \tag{3}$$

This expansion arises naturally when squared l_2 norms are used for the regularization term, for unsquared norms it is not directly available. Using this expansion the solution of a linear system in the dual yields the optimal values for $\boldsymbol{\alpha}$ and b. The expansion in Eq. (3) can be substituted into the initial predictive equation (1) yielding a one-step ahead predictor for the model in terms of $\boldsymbol{\alpha}$, b and inner products of the feature map, which can be evaluated with the kernel function.

Instead of the standard model in (1) we will consider a modified version that allows perturbations on the regression vector $\boldsymbol{x}_k$

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x} + \boldsymbol{\delta}) + b. \tag{4}$$

The disturbance $\boldsymbol{\delta}$ is assumed to be bounded $\left\|\boldsymbol{\delta}_k\right\| \leq \varrho$ and uncorrelated with the error $\mathrm{E}\left\{\epsilon_k \boldsymbol{\delta}_k\right\} = \mathbf{0}$. Unless otherwise stated all norms are considered to be l_2-norms.

III. Robust kernel regression in SOCP formulation

A. A robust primal formulation

Under the assumption that the disturbances of the regressors $\boldsymbol{\delta}_k$ are sufficiently small, their influence can be reasonably well approximated by a first order Taylor series expansion

$$\varphi_j(\boldsymbol{x}_k + \boldsymbol{\delta}_k) \simeq \varphi_j(\boldsymbol{x}_k) + \left[\boldsymbol{\nabla}^T \varphi_j(\boldsymbol{x})\right]_{\boldsymbol{x}=\boldsymbol{x}_k} \boldsymbol{\delta}_k \tag{5}$$

for all $j = 1, \ldots, n_h$ where $\boldsymbol{\nabla} = \left[\frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_n}\right]^T$. During the remainder of the paper $\boldsymbol{\Phi}$ will be used as a shorthand notation for the $n_h \times N$ matrix $\left[\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_N)\right]$, $\boldsymbol{\Phi}'_k$ for the $n_h \times n$ matrix $\left[\left.\frac{\partial}{\partial x_1}\boldsymbol{\varphi}(\boldsymbol{x})\right|_{\boldsymbol{x}_k} \cdots \left.\frac{\partial}{\partial x_n}\boldsymbol{\varphi}(\boldsymbol{x})\right|_{\boldsymbol{x}_k}\right]$ and $\boldsymbol{\Phi}'$ for the $n_h \times (N \cdot n)$ matrix $\left[\boldsymbol{\Phi}'_1, \ldots, \boldsymbol{\Phi}'_N\right]$. The formulation corresponding to Problem (2), using the model (4) that takes the perturbations into account and applying the approximation in (5) is

$$\min_{\boldsymbol{w},b,\epsilon_k,\boldsymbol{\delta}_k} \quad \lambda\|\boldsymbol{w}\| + \|\boldsymbol{\epsilon}\| \tag{6a}$$

subject to

$$y_k = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_k) + \boldsymbol{w}^T \boldsymbol{\Phi}'_k \boldsymbol{\delta}_k + b + \epsilon_k, \quad k = 1, \ldots, N. \tag{6b}$$

Note that we have dropped the squares, thus the solution is now given in terms of an SOCP instead of a linear system. In SOCP form (6) can be recasted into a convex problem by taking a worst case assumption for the perturbations $\|\boldsymbol{\delta}_k\| \leq \varrho$. Based on [10] the result is formalized in the following Lemma.

Lemma 1: Problem (6) is bounded from above by the convex problem

$$\min_{\boldsymbol{w},b,\epsilon_k} \quad \sup_{\|\boldsymbol{\delta}_k\| \leq \varrho} \quad \lambda\|\boldsymbol{w}\| + \|\boldsymbol{\epsilon}\| \tag{7a}$$

subject to

$$y_k = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_k) + \boldsymbol{w}^T \boldsymbol{\Phi}'_k \boldsymbol{\delta}_k + b + \epsilon_k, \quad k = 1, \ldots, N. \tag{7b}$$

The solution of this worst case approximation is equivalent to adding an additional regularization term to the objective function.

$$\min_{\boldsymbol{w},b,\epsilon_k} \quad \lambda \left\|\boldsymbol{w}\right\| + \left\|\boldsymbol{\epsilon}\right\| + \varrho \left\|\boldsymbol{\Phi}'^T \boldsymbol{w}\right\| \tag{8a}$$

subject to

$$y_k = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k, \quad k = 1, \ldots, N. \tag{8b}$$

Proof: To start Problem (7) has to rewritten. Therefore we define the $(N \cdot n) \times N$ matrix $\boldsymbol{\Delta} = \mathrm{blockdiag}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N)$. One obtains

$$\min_{\boldsymbol{w},b,\epsilon} \quad \lambda \left\|\boldsymbol{w}\right\| + \epsilon$$

subject to

$$\epsilon = \sup_{\|\boldsymbol{\delta}_k\| \leq \varrho} \left\|\boldsymbol{y}^T - \boldsymbol{w}^T \boldsymbol{\Phi} - \boldsymbol{w}^T \boldsymbol{\Phi}' \boldsymbol{\Delta} - b\mathbf{1}^T\right\|.$$

The supremum can be computed explicitly. The derivation thereof is adapted from [10]. First compute an upper bound for the supremum

$$\begin{aligned}
&\sup_{\|\boldsymbol{\Delta}\| \leq \varrho} \left\|\boldsymbol{\Phi}^T \boldsymbol{w} + b\mathbf{1} - \boldsymbol{y} + \boldsymbol{\Delta}^T \boldsymbol{\Phi}'^T \boldsymbol{w}\right\| \\
&\quad \leq \left\|\boldsymbol{\Phi}^T \boldsymbol{w} + b\mathbf{1} - \boldsymbol{y}\right\| + \sup_{\|\boldsymbol{\Delta}\| \leq \varrho} \left\|\boldsymbol{\Delta}^T \boldsymbol{\Phi}'^T \boldsymbol{w}\right\| \\
&\quad \leq \left\|\boldsymbol{\Phi}^T \boldsymbol{w} + b\mathbf{1} - \boldsymbol{y}\right\| + \sup_{\|\boldsymbol{\Delta}\| \leq \varrho} \|\boldsymbol{\Delta}\| \left\|\boldsymbol{\Phi}'^T \boldsymbol{w}\right\| \\
&\quad \leq \left\|\boldsymbol{\Phi}^T \boldsymbol{w} + b\mathbf{1} - \boldsymbol{y}\right\| + \varrho \left\|\boldsymbol{\Phi}'^T \boldsymbol{w}\right\|.
\end{aligned}$$

Substitution shows that this upper bound is exact for $\boldsymbol{\Delta} = \varrho \boldsymbol{u}\boldsymbol{v}^T$ with

$$\boldsymbol{u} = \frac{\boldsymbol{\Phi}^T\boldsymbol{w} + b\mathbf{1} - \boldsymbol{y}}{\left\|\boldsymbol{\Phi}^T\boldsymbol{w} + b\mathbf{1} - \boldsymbol{y}\right\|}, \quad \boldsymbol{v} = \frac{\boldsymbol{\Phi}'^T\boldsymbol{w}}{\left\|\boldsymbol{\Phi}'^T\boldsymbol{w}\right\|}.$$

The matrix norm is assumed to be the maximum singular value norm. As $\boldsymbol{\Delta}^T\boldsymbol{\Delta} = \operatorname{diag}(\boldsymbol{\delta}_1^T\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N^T\boldsymbol{\delta}_N)$, the norm $\|\boldsymbol{\Delta}\| = \sigma_{\max}(\boldsymbol{\Delta}) = \max_k \|\boldsymbol{\delta}_k\|$. Combining both yields

$$\begin{aligned}\sup_{\|\boldsymbol{\delta}_k\| \leq \varrho} &\left\|\boldsymbol{y} - \boldsymbol{\Phi}^T\boldsymbol{w} - \boldsymbol{\Delta}^T\boldsymbol{\Phi}'^T\boldsymbol{w} - b\mathbf{1}\right\| \\ &= \left\|\boldsymbol{\Phi}^T\boldsymbol{w} + b\mathbf{1} - \boldsymbol{y}\right\| + \varrho\left\|\boldsymbol{\Phi}'^T\boldsymbol{w}\right\|\end{aligned}$$

which concludes the proof. ∎

Remark 1: Instead of the largest singular value norm as matrix norm the Frobenius norm could be used. This would correspond to a different assumption on the perturbations $\boldsymbol{\delta}_k$ namely $\sum_{k=1}^N \|\boldsymbol{\delta_k}\|^2 \leq \varrho^2$.

Remark 2: In case it is known a priori that one or more of the regression variables are not perturbed, the regressors can be partitioned as $\boldsymbol{x} = \left[\boldsymbol{x}_{\mathcal{C}}^T\, \boldsymbol{x}_{\mathcal{D}}^T\right]^T$, where in $\boldsymbol{x}_{\mathcal{C}}$ are all regressors without a perturbation and in $\boldsymbol{x}_{\mathcal{D}}$ all perturbed ones. Then the first order approximation in Eq. (5) becomes

$$\varphi_j\left(\begin{bmatrix}\boldsymbol{x}_{\mathcal{C},k}\\ \boldsymbol{x}_{\mathcal{D},k}\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{\delta}_k\end{bmatrix}\right) \simeq \varphi_j(\boldsymbol{x}_k) + \boldsymbol{\nabla}_{\boldsymbol{x}_{\mathcal{D}}}^T\varphi_j(\boldsymbol{x})\Big|_{\boldsymbol{x}_k}\boldsymbol{\delta}_k$$

where the gradient is with respect to $\boldsymbol{x}_{\mathcal{D}}$ only.

Assume that some a priori information about $\boldsymbol{\delta}_k$ is given in form of a strictly positive definite matrix D (for a example a correlation structure). This knowledge can be included in the norm bound $\|\boldsymbol{D}\boldsymbol{\delta}_k\| \leq \varrho$ and then transferred into the primal problem by a change of variables. Let $\boldsymbol{\delta}_k = \boldsymbol{D}^{-1}\bar{\boldsymbol{\delta}}_k$ then an equivalent feature map embedding this information is $\bar{\boldsymbol{\Phi}}_k' = \boldsymbol{\Phi}_k'\boldsymbol{D}^{-1}$.

B. Kernelizing via the dual

Problem (8) is in the primal from. As such it cannot be solved directly as in most cases the feature map $\boldsymbol{\varphi}$ itself is unknown. Instead a positive definite kernel function is given. According to Mercer's theorem [4] any positive definite function $K(\boldsymbol{x}, \boldsymbol{y})$ allows the expansion $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\varphi}^T(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{y})$ in terms of some basis $\boldsymbol{\varphi}$, often called the "kernel trick". Thereby replacing inner product of a high dimensional mapping with itself by a scalar positive definite function. The kernel function can be evaluated on the given dataset and the results collected in a Gram matrix. This matrix is defined as $\Omega_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for all $i, j = 1, \ldots, N$. Instead of computing derivatives on the feature map they can be equivalently computed on the kernel. Define $\boldsymbol{\varphi}^T(\boldsymbol{x}_i)\boldsymbol{\Phi}_j' = \nabla_{\boldsymbol{y}}^T K(\boldsymbol{x}, \boldsymbol{y})\big|_{\boldsymbol{x}=\boldsymbol{x}_i, \boldsymbol{y}=\boldsymbol{x}_j} = \boldsymbol{\Omega}_{ij}'$ and $\boldsymbol{\Phi}_i'^T\boldsymbol{\Phi}_j' = \nabla_{\boldsymbol{x}}\nabla_{\boldsymbol{y}}^T K(\boldsymbol{x}, \boldsymbol{y})\big|_{\boldsymbol{x}=\boldsymbol{x}_i, \boldsymbol{y}=\boldsymbol{x}_j} = \boldsymbol{\Omega}_{ij}''$. The derivatives can be brought to the front because they are independent of the argument of the first feature map. The $N \times (N \cdot n)$ block matrix $\boldsymbol{\Omega}'$ combines the individual $1 \times n$ submatrices $\boldsymbol{\Omega}_{ij}'$. The matrix $\boldsymbol{\Omega}''$ is also block structured and of dimension $(N \cdot n) \times (N \cdot n)$ and collects the $n \times n$ elements $\boldsymbol{\Omega}_{ij}''$.

Using this formalism the dual of Problem (8) can be derived in a kernel based form. The result is stated in the following lemma

Lemma 2: The Lagrange dual of Problem (8) is

$$\max_{\boldsymbol{\alpha}, \boldsymbol{v}} \quad \sum_{k=1}^N \alpha_k y_k \tag{10a}$$

subject to

$$\mathbf{1}^T\boldsymbol{\alpha} = 0 \tag{10b}$$

$$\|\boldsymbol{\alpha}\| \leq 1,\ \|\boldsymbol{v}\| \leq 1,\ \left\|\boldsymbol{G}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{v}\end{bmatrix}\right\| \leq \lambda \tag{10c}$$

where $\boldsymbol{G}$ is the Cholesky factor of the Cholesky decomposition of the matrix

$$\tilde{\boldsymbol{\Omega}} = \begin{bmatrix}\boldsymbol{\Omega} & -\varrho\boldsymbol{\Omega}'^T\\ -\varrho\boldsymbol{\Omega}' & \varrho^2\boldsymbol{\Omega}''\end{bmatrix}$$

into $\tilde{\boldsymbol{\Omega}} = \boldsymbol{G}^T\boldsymbol{G}$.

Proof: A trick borrowed from [16] allows to rewrite norms by introducing slack variables $\|\boldsymbol{x}\| = \max_{\|\boldsymbol{c}\| \leq 1} \boldsymbol{c}^T\boldsymbol{x}$. Applying this technique the Lagrangian for Problem (8) can be written as

$$\begin{aligned}\mathscr{L} = \lambda\boldsymbol{u}^T\boldsymbol{w} + \varrho\boldsymbol{v}^T\boldsymbol{\Phi}'^T\boldsymbol{w} - \boldsymbol{a}^T\boldsymbol{\epsilon} \\ - \sum_{k=1}^N \alpha_k\left(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k - y_k\right).\end{aligned}$$

The constraints on the dual variables, that are introduced as slacks, are

$$\|\boldsymbol{u}\|, \|\boldsymbol{v}\|, \|\boldsymbol{a}\| \leq 1.$$

Taking the conditions for optimality yields

$$\text{KKT}\begin{cases}\dfrac{\partial\mathscr{L}}{\partial b} = 0 & \Rightarrow \quad \mathbf{1}^T\boldsymbol{\alpha} = 0\\ \dfrac{\partial\mathscr{L}}{\partial\epsilon_k} = 0 & \Rightarrow \quad a_k = \alpha_k\\ \dfrac{\partial\mathscr{L}}{\partial\boldsymbol{w}} = \mathbf{0} & \Rightarrow \quad \lambda\boldsymbol{u} + \varrho\boldsymbol{\Phi}'^T = \boldsymbol{\Phi}\boldsymbol{\alpha}\end{cases} \tag{11}$$

Backsubstitution into the Lagrangian results in the dual optimization problem

$$\max_{\boldsymbol{\alpha}, \boldsymbol{u}, \boldsymbol{v}} \quad \sum_{k=1}^N \alpha_k y_k \tag{12a}$$

subject to

$$\mathbf{1}^T\boldsymbol{\alpha} = 0 \tag{12b}$$

$$\|\boldsymbol{\alpha}\| \leq 1,\ \|\boldsymbol{u}\| \leq 1,\ \|\boldsymbol{v}\| \leq 1 \tag{12c}$$

$$\lambda\boldsymbol{u} + \varrho\boldsymbol{\Phi}'^T\boldsymbol{v} = \boldsymbol{\Phi}\alpha. \tag{12d}$$

This problem still has references to the feature map itself, thus we have to apply the kernel trick to rewrite it only in

terms of the kernel function. The constraint (12d) can to be substituted into the squared constraint (12c). This yields

$$\begin{aligned} 0 \leq \|\boldsymbol{u}\|^2 &= \frac{1}{\lambda^2}\left\|\boldsymbol{\Phi}\alpha - \varrho\boldsymbol{\Phi}'^T\boldsymbol{v}\right\|^2 \\ &= \frac{1}{\lambda^2}\left(\boldsymbol{\alpha}^T\boldsymbol{\Phi}^T - \varrho\boldsymbol{v}^T\boldsymbol{\Phi}'^T\right)\left(\boldsymbol{\Phi}\boldsymbol{\alpha} - \varrho\boldsymbol{\Phi}'\boldsymbol{v}\right) \\ &= \frac{1}{\lambda^2}\left(\boldsymbol{\alpha}^T\boldsymbol{\Omega}\boldsymbol{\alpha} - 2\varrho\boldsymbol{v}^T\boldsymbol{\Omega}'^T\boldsymbol{\alpha} + \varrho^2\boldsymbol{v}^T\boldsymbol{\Omega}''\boldsymbol{v}\right) \\ &= \frac{1}{\lambda^2}\begin{bmatrix}\boldsymbol{\alpha}^T\boldsymbol{v}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega} & -\varrho\boldsymbol{\Omega}'^T \\ -\varrho\boldsymbol{\Omega}' & \varrho^2\boldsymbol{\Omega}''\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{v}\end{bmatrix}. \end{aligned}$$

The kernel matrix of this quadratic form is $\tilde{\boldsymbol{\Omega}}$ and it is positive semidefinite by construction as it corresponds to a squared norm. Using the Cholesky decomposition of $\tilde{\boldsymbol{\Omega}}$ the proof can be completed. ∎

Remark 3: The kernel function can be any positive definite function. Commonly used are a Gaussian RBF kernel $K_G(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{1}{\sigma^2}\|x-y\|^2\right)$ and the polynomial kernel $K_P(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y}+c)^d$ of degree d with $c \geq 0$. The feature map corresponding to the Gaussian kernel is infinite dimensional. The polynomial kernel corresponds to a feature map containing all monomials with order up to $x_i^d y_i^d$.

Remark 4: The problem at hand has several free parameters. The primal problem has the regularization parameter λ and the bound on the errors ϱ. Depending on the choice of the kernel, it may have one or more parameters that have to be chosen. The training procedure involves the selection of these hyperparameters, which can be done e.g. by cross-validation, Bayesian techniques [19] or others.

C. Deriving a predictive equation

To be able to use the identified model to make predictions, a prediction equation, like Eq. (1) with the expansion in Eq. (3), has to be deduced. The model equation remains the same, but unlike in the simple case, the KKT conditions (11) do not allow an expansion in terms of $\boldsymbol{w}$. Using a different Lagrangian the expansion can be derived except for the length of $\boldsymbol{w}$. Yet the length can be computed by a solving a small auxiliary linear system. The following lemma states the final prediction equation for the SOCP problem.

Lemma 3: Given the optimal solutions $\boldsymbol{\alpha}^*$ and $\boldsymbol{v}^*$ for the dual variables in (10), the predictive equation for (8) is

$$\hat{y}(\boldsymbol{z}) = \frac{L_w^*}{\lambda}\sum_{k=1}^{N}\alpha_k^* K(\boldsymbol{z},\boldsymbol{x}_k) + b^* - \varrho\frac{L_w^*}{\lambda}\sum_{k=1}^{N}\boldsymbol{K}'(\boldsymbol{z},\boldsymbol{x}_k)\boldsymbol{v}_k^* \tag{13}$$

with $\boldsymbol{K}'(\boldsymbol{x}_0,\boldsymbol{y}_0) = \boldsymbol{\nabla}_{\boldsymbol{y}}^T K(\boldsymbol{x},\boldsymbol{y})\Big|_{\boldsymbol{x}_0,\boldsymbol{y}_0}$, $\boldsymbol{v}^* = \left[\boldsymbol{v}_1^{*T},\ldots,\boldsymbol{v}_N^{*T}\right]^T$ and $\boldsymbol{v}_k^* \in \mathbb{R}^n$. The values of L_w^*, $L_\epsilon^*, b^* \in \mathbb{R}$ are the solutions of the system

$$\boldsymbol{y} = \frac{L_w}{\lambda}\left(\boldsymbol{\Omega}\boldsymbol{\alpha}^* - \varrho\boldsymbol{\Omega}'\boldsymbol{v}^*\right) + b\boldsymbol{1} + L_\epsilon\boldsymbol{\alpha}^*. \tag{14}$$

Proof: Using an alternative form of the Lagrangian for Problem (8)

$$\begin{aligned}\mathscr{L} = \lambda\|\boldsymbol{w}\| + \|\boldsymbol{\epsilon}\| + \varrho\boldsymbol{v}^T\boldsymbol{\Phi}'^T\boldsymbol{w} \\ - \sum_{k=1}^{N}\alpha_k\left(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k - y_k\right)\end{aligned}$$

the KKT conditions for $\boldsymbol{w}$ and $\boldsymbol{\epsilon}$ can be recomputed.

$$\text{KKT}\begin{cases}\dfrac{\partial\mathscr{L}}{\partial\boldsymbol{w}} = \boldsymbol{0} \quad\Rightarrow\quad \lambda\dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \varrho\boldsymbol{\Phi}'\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\alpha} \\ \dfrac{\partial\mathscr{L}}{\partial\boldsymbol{\epsilon}} = \boldsymbol{0} \quad\Rightarrow\quad \dfrac{\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|} = \boldsymbol{\alpha}.\end{cases} \tag{15}$$

This yields an expansion of $\boldsymbol{w}$ in terms of the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{v}_k$ for which the optimal values are already known. Yet the expansion contains the lengths $L_w = \|\boldsymbol{w}\|$ and $L_\epsilon = \|\boldsymbol{\epsilon}\|$ as free variables which have to be determined. Substituting the KKT conditions (15) for the optimal values $\boldsymbol{\alpha}^*$ and $\boldsymbol{v}^*$ into Eq. (8b) yields Eq. (14). The one-step ahead predictor (13) is thus the combination of the model equation (1), the expansion in (15) and the length $L_{\boldsymbol{w}}^*$ of $\boldsymbol{w}$ given by the solution of (14). ∎

IV. Recasting into a least squares problem

The objective function in (2) contained squared l_2-norms which were dropped later on to be able to compute the worst case solution. Thus the least squares problem was transferred into a SOCP. For computational efficiency the squares can be reintroduced into (8). After this transition the regularization parameters are different. The interpretation of ϱ as the bound on the perturbations is lost. The results of the modification are summarized in the following lemma.

Lemma 4: For some parameters $\tilde{\lambda}$ and $\tilde{\varrho}$ different from λ and ϱ the original problem is equivalent in terms of the solution to the following least squares problem

$$\min_{\boldsymbol{w},b,\epsilon_k} \quad \frac{1}{2}\tilde{\lambda}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\tilde{\varrho}\boldsymbol{w}^T\boldsymbol{\Phi}'\boldsymbol{\Phi}'^T\boldsymbol{w} + \frac{1}{2}\sum_{k=1}^{N}\epsilon_k^2$$

subject to

$$y_k = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k, \qquad k = 1,\ldots,N.$$

The dual of this problem corresponds to the system of linear equations

$$\begin{bmatrix}\frac{1}{\tilde{\lambda}}\breve{\boldsymbol{\Omega}} + \boldsymbol{I}_N & \boldsymbol{1} \\ \boldsymbol{1}^T & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ b\end{bmatrix} = \begin{bmatrix}\boldsymbol{y} \\ 0\end{bmatrix}$$

where a modified kernel matrix $\breve{\boldsymbol{\Omega}}$ is defined as $\breve{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \boldsymbol{\Omega}'\left(\frac{\tilde{\lambda}}{\tilde{\varrho}}\boldsymbol{I}_{(N\cdot n)} + \boldsymbol{\Omega}''\right)^{-1}\boldsymbol{\Omega}'^T$.

Proof:

$$\begin{aligned}\mathscr{L} = \frac{1}{2}\tilde{\lambda}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\tilde{\varrho}\boldsymbol{w}^T\boldsymbol{\Phi}'\boldsymbol{\Phi}'^T\boldsymbol{w} + \frac{1}{2}\sum_{k=1}^{N}\epsilon_k^2 \\ - \sum_{k=1}^{N}\alpha_k(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k - y_k).\end{aligned}$$

Now the conditions for optimality can be computed

$$\text{KKT} \begin{cases} \dfrac{\partial \mathscr{L}}{\partial b} = 0 & \Rightarrow \quad \displaystyle\sum_{k=1}^{N} \alpha_k = 0 \\ \dfrac{\partial \mathscr{L}}{\partial \epsilon_k} = 0 & \Rightarrow \quad \epsilon_k = \alpha_k \\ \dfrac{\partial \mathscr{L}}{\partial \boldsymbol{w}} = \boldsymbol{0} & \Rightarrow \quad \tilde{\lambda}\boldsymbol{w} + \tilde{\varrho}\boldsymbol{\Phi}'\boldsymbol{\Phi}'^T\boldsymbol{w} = \displaystyle\sum_{k=1}^{N} \alpha_k \boldsymbol{\varphi}(\boldsymbol{x}_k) \\ \dfrac{\partial \mathscr{L}}{\partial \alpha_k} = 0 & \Rightarrow \quad y_k = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_k) + b + \epsilon_k. \end{cases}$$

Applying the matrix inversion lemma to the constraint $\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}} = \boldsymbol{0}$ yields

$$\tilde{\lambda}\boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{\alpha} - \boldsymbol{\Phi}'\left(\frac{\tilde{\lambda}}{\tilde{\varrho}}\boldsymbol{I}_{(N\cdot n)} + \boldsymbol{\Phi}'^T\boldsymbol{\Phi}'\right)^{-1}\boldsymbol{\Phi}'^T\boldsymbol{\Phi}\boldsymbol{\alpha} \quad (17a)$$

$$= \boldsymbol{\Phi}\boldsymbol{\alpha} - \boldsymbol{\Phi}'\left(\frac{\tilde{\lambda}}{\tilde{\varrho}}\boldsymbol{I}_{(N\cdot n)} + \boldsymbol{\Omega}''\right)^{-1}\boldsymbol{\Omega}'^T\boldsymbol{\alpha}. \quad (17b)$$

Substituting back into Eq. (8b)

$$\boldsymbol{y} = \frac{1}{\tilde{\lambda}}\boldsymbol{\Omega}\boldsymbol{\alpha} - \frac{1}{\tilde{\lambda}}\boldsymbol{\Omega}'\left(\frac{\tilde{\lambda}}{\tilde{\varrho}}\boldsymbol{I}_{(N\cdot n)} + \boldsymbol{\Omega}''\right)^{-1}\boldsymbol{\Omega}'^T\boldsymbol{\alpha} + b\boldsymbol{1} + \boldsymbol{\alpha}.$$

This concludes the derivation. ■

Remark 5: For out of sample extensions a prediction equation has to be derived. Therefore the value for $\boldsymbol{w}$ in Eq. (17) is substituted into the prediction model (1) which yields

$$\hat{y}(\boldsymbol{z}) = \frac{1}{\tilde{\lambda}}\sum_{k=1}^{N}\alpha_k K(\boldsymbol{x}_k, \boldsymbol{z}) + b - \frac{1}{\tilde{\lambda}}\left[\boldsymbol{K}'(\boldsymbol{z}, \boldsymbol{x_1}), \ldots, \boldsymbol{K}'(\boldsymbol{z}, \boldsymbol{x_N})\right]\left(\frac{\tilde{\lambda}}{\tilde{\varrho}}\boldsymbol{I}_{(N\cdot n)} + \boldsymbol{\Omega}''\right)^{-1}\boldsymbol{\Omega}'^T\boldsymbol{\alpha}$$

V. Examples

In this section we illustrate two possible applications of the presented formulations. The SOCP form can be used to analyse the robustness of an identified model, the influence of specific regressors or of choices of the kernel function. For this purpose a regular system ($\varrho = 0$) is identified and in an afterwards analysis the influence of a varying ϱ is studied. The second application is the identification of a robust model, for this ϱ is seen as a hyperparameter that has to be selected in the same fashion as λ and the kernel parameters. In this case performance is more important than the interpretation of ϱ as such the Least Squares formulation can be used for increased computational efficiency.

All simulations are conducted in MATLAB using CVX as a modeler for convex problems [20], [21]. The solver called by CVX to solve the SOCPs is SDPT3 [22], [23].

Unless stated otherwise all examples consist of a training and a validation set with 50 samples each and a test set with 100 samples. All samples are drawn from a uniform distribution on $[-3, 3]$, and the outputs are subject to an additive Gaussian white noise with $\sigma = 0.2\sqrt{\mathrm{Var}(y_k)}$. Unless otherwise stated a Gaussian RBF kernel will be used throughout. All parameters are chosen according to the validation set. The following toy problems are used to illustrate the possible applications

1) Static sinc *function:* $f(x_1, x_2, x_3) = \sum_{j=1}^{3}\frac{\sin(x_j)}{x_j}$.

2) Static polynomial: $g(x_1, x_2, x_3) = x_1^4 + 2x_2^3 - 5x_3 - 2x_1x_2 + 2$.

3) NARX model: $\hat{y}_k = h(y_{k-1}, y_{k-2}, u_k, \ldots, u_{k-3}) = u_k^2 + u_{k-1}\sin(u_{k-2} + u_{k-3}) + y_{k-1}^2\,\mathrm{sinc}(y_{k-2})$. For the NARX model the training and the validation set are of size 250 samples. The size of $\tilde{\boldsymbol{\Omega}}$ in this case is 2000×2000.

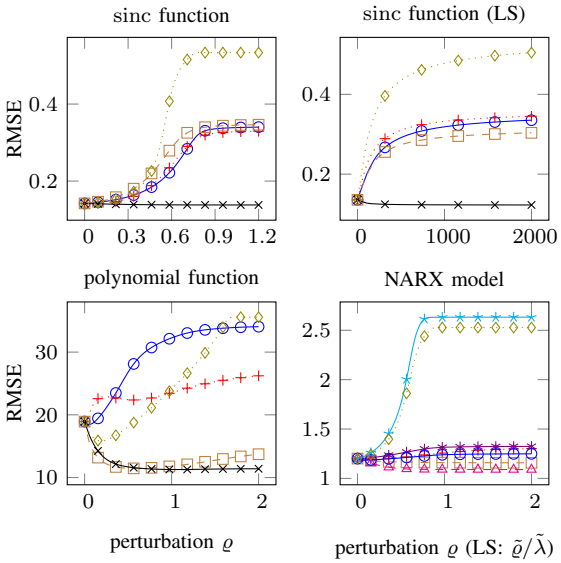

Fig. 1: Sensitivity of different inputs in a kernel based model

A. Sensitivity of different inputs

By assuming that all components of the regression vector except for one are not perturbed, the influence on the model of a perturbation of the remaining one can be analysed. Therefore we added an additional independent variable to the models and then analysed the influence of each component. The behavior of the overall system is given as a reference. From Figure 1 the independent variable can be clearly identified. As it only adds noise in the training the RMSE is improved if regularization is added to that component. The sinc example shows that variables with the same influence show similar behavior, whereas for the polynomial function different sensitivities among the inputs are detected.

B. Robustness of different kernels

Figure 2 shows the influence of the choice of the kernel function. It can be seen that the Gaussian RBF is the more robust kernel for the toy examples. For the sinc function it has better prediction performance to start with and is able to retain the performance level for a larger amount

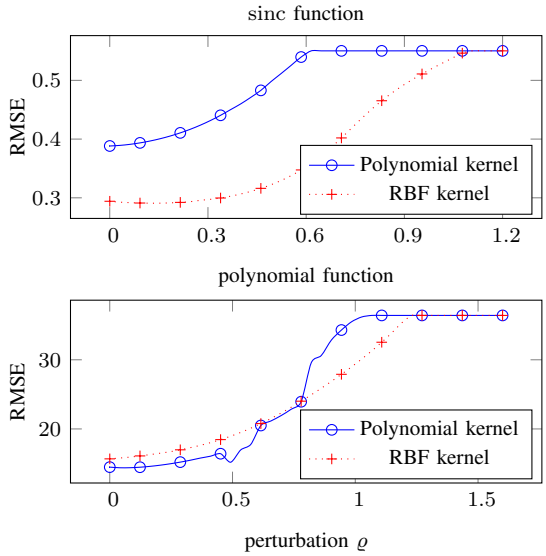

Fig. 2: Sensitivity of kernel based model depending on the choice of the used kernel function

of perturbation. In case of the polynomial example the polynomial kernel initially outperforms the Gaussian RBF kernel in terms of prediction performance. The degradation of the Gaussian kernel though is not as fast as that of the polynomial kernel. Although the polynomial is the better choice in terms of prediction performance for this problem if robustness is needed the RBF has better properties.

VI. Conclusions

We have shown how a robust formulation of Least Squares Support Vector Machines can be obtained by adding an additional regularization term, where the regularization term has a direct interpretation as a bound on the perturbations. The dual formulation can still be expressed in terms of the kernel function. In case the interpretation of the additionally introduced parameter is not crucial, we have shown that a computationally more efficient solution in terms a simple system of linear equations can be found. This solution corresponds to standard LS-SVM with a modified kernel matrix. The obtained one-step ahead prediction rules explicitly incorporate the bound on the perturbations. These results were then applied to analyse the influence of different inputs and different choices for the used kernel function. In both cases the results closely match the expectations. Some of the remaining challenges in this context are robust model selection, improved computational tractability (the size of the matrix $\mathbf{\Omega}''$ is $(N \cdot n)^2$) and structured perturbations.

Acknowledgments

Research Council KUL: GOA AMBioRICS, CoE EF/05/006 OPTEC, IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0302.07, G.0320.08, G.0558.08, G.0557.08, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite+ Belgian Federal Science Policy Office: IUAP P6/04 DYSCO; EU: ERNSI; Contract Research: AMINAL. J. Suykens is a professor and B. De Moor is a full professor at the K.U. Leuven.

References

[1] L. Ljung, *System identification: Theory for the User*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1999.
[2] 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*, 31(12), 1691–1724, 1995.
[3] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjoberg, and Q. Zhang, "Nonlinear black-box models in system identification: Mathematical foundations," *Automatica*, 31(12), 1725–1750, 1995.
[4] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.
[5] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, 2002.
[6] M. Espinoza, J. A. K. Suykens, R. Belmans, and B. De Moor, "Electric load forecasting - using kernel based modeling for nonlinear system identification," *IEEE Control Systems Magazine*, 27(5), 43–57, 2007.
[7] G. Wahba, *Spline Models for Observational Data*. SIAM, 1990.
[8] J. A. K. Suykens and J. Vandewalle, "Recurrent least squares support vector machines," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 47(7), 1109–1114, 2000.
[9] M. Espinoza, J. A. K. Suykens, and B. De Moor, "Kernel based partially linear models and nonlinear identification," *IEEE Transactions on Automatic Control*, 50(10), 1602–1606, 2005.
[10] L. El Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM J. Matrix Anal. Appl.*, 18(4), 1035–1064, 1997.
[11] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed, "An efficient algorithm for a bounded errors-in-variables model," *SIAM J. Matrix Anal. Appl.*, 20(4), 839–859, 1999.
[12] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem : Computational Aspects and Analysis, Frontiers in Applied Mathematics Series, Vol. 9*. SIAM, Philadelphia, 1991.
[13] J. B. Rosen, H. Park, and J. Glick, "Structured total least norm for nonlinear problems," *SIAM J. Matrix Anal. Appl.*, 20(1), 14–30, 1998.
[14] G. A. Watson, "Robust solutions to a general class of approximation problems," *SIAM J. Sci. Comp.*, 25(4), 1448–1460, 2003.
[15] ——, "Robust counterparts of errors-in-variables problems," *Computational Statistics and Data Analysis*, 1080–1089, 2007.
[16] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola, "Second order cone programming approaches for handling missing and uncertain data," *Journal Machine Learning Research*, 7, 1283–1314, 2006.
[17] T. B. Trafalis and R. C. Gilbert, "Robust classification and regression using support vector machines," *Eur. J. Oper. Res.*, 173(3), 893–909, 2006.
[18] R. A. Renault, H. Guo, and W. J. Chen, "Regularised total least squares support vector machines," Presentation, May 2005. [Online]. Available: http://math.asu.edu/~rosie/mypresentations/Rosie_talk_svmc.pdf
[19] D. J. C. MacKay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, 11(5), 1035–1068, 1999.
[20] M. Grant and S. P. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer, 2008, to appear. [Online]. Available: http://stanford.edu/~boyd/graph_dcp.html
[21] ——, "CVX: Matlab software for disciplined convex programming (web page and software)," Mar. 2008. [Online]. Available: http://stanford.edu/~boyd/cvx
[22] K. Toh, M. Todd, and R. Tutuncu, "SDPT3 — a matlab software package for semidefinite programming," *Optimization Methods and Software*, 11, 545–581, 1999.
[23] R. Tutuncu, K. Toh, and M. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming Series B*, 95, 189–217, 2003.