Segmentation of time series from nonlinear dynamical systems

Tillmann Falck * **Henrik Ohlsson** ** **Lennart Ljung** **
Johan A.K. Suykens * **Bart De Moor** *

* *Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, B-3001 Leuven, Belgium ({tillmann.falck, johan.suykens, bart.demoor}@esat.kuleuven.be).*
** *Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden ({ohlsson, ljung}@isy.liu.se).*

Abstract: Segmentation of time series data is of interest in many applications, as for example in change detection and fault detection. In the area of convex optimization, the sum-of-norms regularization has recently proven useful for segmentation. Proposed formulations handle linear models, like ARX models, but cannot handle nonlinear models. To handle nonlinear dynamics, we propose integrating the sum-of-norms regularization with a least squares support vector machine (LS-SVM) core model. The proposed formulation takes the form of a convex optimization problem with the regularization constant trading off the fit and the number of segments.

Keywords: Convex Optimization, System Identification, Support Vector Machines, Failure Detection, Nonlinear Systems.

1. INTRODUCTION

Segmentation of time-varying systems and signals into models whose parameters are piecewise constant in time is an important and well studied problem. The segmentation problem is often addressed using multiple detection techniques, multiple models and/or Markov models with switching regression, see, *e.g.,* Lindgren (1978); Tugnait (1982); Bodenstein and Praetorius (1977). The function `segment` for the segmentation problem in the System Identification Toolbox (Ljung, 2007), is based on a multiple model technique (Andersson, 1985).

A recently proposed method for segmentation is to use sum-of-norms regularization. This method has been used in trend estimation (Kim et al., 2009) and for the estimation of segmented ARX models (Ohlsson et al., 2010). The scheme proposed by Ohlsson et al. (2010) is limited to segmented ARX models,

$$f_c(\boldsymbol{x}_t) = \boldsymbol{w}_c^T \boldsymbol{x}_t, \quad t = t_{c-1}, \ldots, t_c - 1, \tag{1}$$

with $\boldsymbol{x}_t$ stacked past system outputs and inputs $\boldsymbol{x}_t = [y_{t-1}, \ldots, y_{t-p}, u_t, \ldots, u_{t-q}]^T$. For handling highly nonlinear systems, the linear structure of (1) is of little use.

The proposed method of this paper handles the segmentation of nonlinear systems by integrating the sum-of-norms regularization with a least squares support vector machine (LS-SVM) core model (Suykens et al., 2002, 2010). The support vector methodology is able to perform regression in high dimensional spaces through the use of positive semi-definite kernel functions and effective regularization and has successfully been applied to system identification problems, see *e.g.,* Espinoza et al. (2007) and reference therein. They typically start in terms of a high dimensional feature map and derive the Lagrange dual in terms of the kernel function. There are several related kernel based techniques like Support Vector Machines (SVMs) (Vapnik, 1998; Schölkopf and Smola, 2002), Splines (Wahba, 1990) or Gaussian Processes (*e.g.,* Rasmussen and Williams, 2006). Working with LS-SVMs has the advantage that the emerging optimization problems admit easy formulations. Simple regression and classification cases just require the solution of linear systems. The methodology is applicable to a wide range of problems in supervised and unsupervised learning (Suykens et al., 2010).

An important feature of the proposed scheme, the sum-of-norms regularization to detect changes in the parameters and the LS-SVM core model to handle nonlinearities, is that they result in a single convex optimization problem that can be solved efficiently.

The structure of the paper is as follows. The next section will state the general setting while Sec. 3 gives further information on the the sum-of-norms regularization. The kernel based model is derived in Sec. 4 and model selection is briefly discussed in Sec. 5. A short overview on algorithmic considerations is given in Sec. 6. The paper ends with an application to two motivational data sets in Sec. 7 and concludes in the last section.

2. PROBLEM FORMULATION

Assume a nonlinear system whose dynamics change at time instances $\{t_c\}_{c=1}^{C}$ can be modeled in discrete time by the parametric model

$$f_c(\boldsymbol{x}_t) = \boldsymbol{w}_c^T \boldsymbol{\varphi}_c(\boldsymbol{x}_t), \quad t = t_{c-1}, \ldots, t_c - 1. \tag{2}$$

Here we extended the linear model in (1) by using nonlinear basis functions $\boldsymbol{\varphi}_c$ to map the data into another space. The model parameters are $\boldsymbol{w}_c \in \mathbb{R}^{n_h}$ and the components

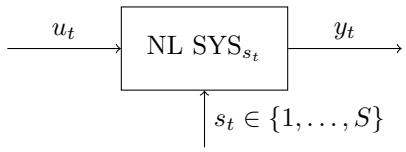

Fig. 1. Nonlinear dynamical system with inputs u_t, outputs y_t and (unknown) scheduling variable s_t.

of the nonlinear maps $\boldsymbol{\varphi}_c(\cdot) = [\varphi_c^1(\cdot), \ldots, \varphi_c^{n_h}(\cdot)]^T : \mathbb{R}^D \to \mathbb{R}^{n_h}$ are the corresponding basis functions. Both are defined for $c = 1, \ldots, C$ where $C < N$ and without loss of generality we assume that $t_0 = 1$.

Given measurement data $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$ of the system described by (2) we wish to estimate the number of changes C as well as their positions t_c and the model parameters $\boldsymbol{w}_c$ in each segment. We assume that in each segment c the model parameters can be estimated from a regularized least squares problem

$$\begin{aligned} \min_{\boldsymbol{w}_c, e_t} \quad & \frac{1}{2}\boldsymbol{w}_c^T\boldsymbol{w}_c + \frac{1}{2}\gamma \sum_{t=t_{c-1}}^{t_c-1} e_t^2 \\ \text{subject to} \quad & y_t = \boldsymbol{w}_c^T\boldsymbol{\varphi}_c(\boldsymbol{x}_t) + e_t, \quad t = t_{c-1}, \ldots, t_c - 1. \end{aligned} \tag{3}$$

The regularization parameter γ trades off the model fit measured by the squared residuals versus the model complexity quantified using the quadratic penalty term $\boldsymbol{w}_c^T\boldsymbol{w}_c$.

3. PIECEWISE NONLINEAR MODELING

If the change points t_c, $c = 1, \ldots, C$ were known, the task would simply be to estimate a model using (3) on each segment. Now, with a unknown number and position of change points, the problem becomes considerably more difficult. Here, we follow the approach in Ohlsson et al. (2010). However, we seek a model (2) instead of the linear model (1) used in Ohlsson et al. (2010). First, to make the problem tractable, the basis functions are fixed across segments, namely $\boldsymbol{\varphi}_c = \boldsymbol{\varphi}\ \forall\ c$. Therefore the basis functions have to be chosen rich enough to represent the dynamics of all segments. Then, to deal with the unknown change points t_c, $c = 1, \ldots, C$, we overparameterize and introduce a parameter $\boldsymbol{w}_t$ for each time instant t. We hence seek a model of the form

$$f_t(\boldsymbol{x}_t) = \boldsymbol{w}_t^T\boldsymbol{\varphi}_t(\boldsymbol{x}_t), \quad t = 1, \ldots, N. \tag{4}$$

In order to avoid a severe overfit, a sum-of-norms regularization $\sum_{t=2}^N \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2$ is added to the cost criteria to limit the flexibility of the model (4). The proposed formulation reads

$$\begin{aligned} \min_{\boldsymbol{w}_t, e_t} \quad & \|\boldsymbol{w}_1\|_2 + \sum_{t=2}^N \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2 + \frac{1}{2}\gamma \sum_{t=1}^N e_t^2 \\ \text{subject to} \quad & y_t = \boldsymbol{w}_t^T\boldsymbol{\varphi}(\boldsymbol{x}_t) + e_t, \quad t = 1, \ldots, N. \end{aligned} \tag{5}$$

The sum-of-norms regularization favors solutions where "many" of the regularized variables come out as exactly zero. Here, this means that there are only few changes of the parameter vector $\boldsymbol{w}_t$ over time t. The number of changes is roughly controlled by the regularization parameter γ.

The sum-of-norms regularization has strong similarities to the ℓ_1-regularization, which has been very popular recently, *e.g.,* in the much used Lasso method, Tibshirani (1996) or compressed sensing Donoho (2006); Candès et al. (2006). In fact, if we define

$$\Delta\boldsymbol{w} \triangleq \Big[\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2, \|\boldsymbol{w}_3 - \boldsymbol{w}_2\|_2, \ldots, \|\boldsymbol{w}_N - \boldsymbol{w}_{N-1}\|_2\Big]^T,$$

one could see (5) as a ℓ_1-regularized problem with the ℓ_1-regularization acting on the vector $\Delta\boldsymbol{w}$. The ℓ_1-regularization makes sure that the vector $\Delta\boldsymbol{w}$ becomes sparse. Another interpretation is to view the sum-of-norms penalty as a $L_{2,1}$ group norm regularization on the column of the matrix $[\boldsymbol{w}_2 - \boldsymbol{w}_1, \ldots, \boldsymbol{w}_N - \boldsymbol{w}_{N-1}]$ (Argyriou et al., 2008). Note that the new problem has only one regularization parameter that tunes the model complexity at the same time as the sparsity. This is at the same time an advantage and a limitation. With only one free parameter the selection is easier, but model complexity and the degree of sparsity cannot be tuned individually. We will come back to this issue in Section 5 on model selection.

4. NONPARAMETRIC KERNEL BASED FORMULATION

In the framework of least squares support vector machines (LS-SVMs) (Suykens et al., 2002, 2010) we can identify (5) as a combination of a LS-SVM core model with a special regularization scheme. This allows us to utilize the power of support vector machines for regression tasks. One key advantage is that the usually difficult choice of a good set of basis functions $\boldsymbol{\varphi}$ to model all different segments is simplified.

The key idea in SVMs (Vapnik, 1998) is not to define the basis functions $\boldsymbol{\varphi}$, called feature map, explicitly but to define them implicitly by means of their inner products in the dual optimization problem. Using Mercer's theorem the inner products of the feature map can be replaced by a positive semi-definite kernel function $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{y})$. Popular nonlinear kernel functions are the RBF kernel $K_{\text{RBF}}(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2/\sigma^2)$ with $\sigma \geq 0$ which corresponds to an infinite dimensional feature map and the polynomial kernel $K_{\text{poly}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + c)^d$ with $c \geq 0$ and $d \in \mathbb{N}$. The feature map of the polynomial kernel contains all monomials of order up to d.

4.1 Dual formulation

Due to the ℓ_2-norms (which are not squared) in (5), the overparametrized problem has to be solved as a second order cone programming problem (SOCP) instead of a simple linear system as for (3). To derive the dual problem we apply the procedure taken in Falck et al. (2009) which yields the following dual problem.

Lemma 1. Let $\boldsymbol{G}$ be a matrix square root of the kernel matrix $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ with $\Omega_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Furthermore let $\boldsymbol{l}_t = [\boldsymbol{0}_t, \boldsymbol{1}_{N-t}]^T$, where $\boldsymbol{0}_n \in \mathbb{R}^n$ and $\boldsymbol{1}_n \in \mathbb{R}^n$ are vectors of all zeros and ones respectively. Then the dual problem of (5) is

$$\begin{aligned} \max_{\alpha_t} \quad & \sum_{t=1}^N \alpha_t y_t - \frac{1}{2\gamma}\alpha_t^2 \\ \text{subject to} \quad & \|\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_t\|_2 \leq 1, \quad t = 1, \ldots, N, \end{aligned} \tag{6}$$

where α_t are the Lagrange multipliers corresponding to the equality constraints in (5) and $\boldsymbol{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_N)$.

Proof. To obtain a Lagrangian that gives rise to finite dimensional dual problem we express the $\|\cdot\|_2$ regularization terms in terms of their dual norm as in Shivaswamy et al. (2006)

$$\max_{\|\boldsymbol{y}\|_2 \leq 1} \boldsymbol{y}^T \boldsymbol{x}.$$

Using this definition the Lagrangian of (5) can be stated as

$$\mathscr{L}(\boldsymbol{w}_t, \boldsymbol{v}_t, e_t, \alpha_t) = \boldsymbol{v}_1^T \boldsymbol{w}_1 + \sum_{t=2}^{N} \boldsymbol{v}_t^T (\boldsymbol{w}_t - \boldsymbol{w}_{t-1}) + \frac{1}{2}\gamma \sum_{t=1}^{N} e_t^2 - \sum_{t=1}^{N} \alpha_t (\boldsymbol{w}_t^T \boldsymbol{\varphi}(\boldsymbol{x}_t) + e_t - y_t) \quad (7)$$

with $\|\boldsymbol{v}_t\|_2 \leq 1$ for $t = 1, \ldots, N$. The corresponding KKT conditions for optimality (see *e.g.,* Boyd and Vandenberghe, 2004) are

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}_t} &= \boldsymbol{0} : \ \boldsymbol{v}_t - \boldsymbol{v}_{t+1} = \alpha_k \boldsymbol{\varphi}(\boldsymbol{x}_t), \quad t = 1, \ldots, N-1, \\
\frac{\partial \mathscr{L}}{\partial \boldsymbol{w}_N} &= \boldsymbol{0} : \ \boldsymbol{v}_N = \alpha_N \boldsymbol{\varphi}(\boldsymbol{x}_N), \\
\frac{\partial \mathscr{L}}{\partial e_t} &= 0 : \ \gamma e_t = \alpha_t.
\end{aligned}$$

Substitution of the KKT conditions into the Lagrangian yields the dual optimization problem

$$\begin{aligned}
\max_{\boldsymbol{v}_t, \alpha_t} \quad & \sum_{t=1}^{N} \alpha_t y_t - \frac{1}{2\gamma} \alpha_t^2 \\
\text{subject to} \quad & \\
& \boldsymbol{v}_t - \boldsymbol{v}_{t+1} = \alpha_t \boldsymbol{\varphi}(\boldsymbol{x}_t), \quad t = 1, \ldots, N-1, \\
& \boldsymbol{v}_N = \alpha_N \boldsymbol{\varphi}(\boldsymbol{x}_N), \\
& \|\boldsymbol{v}_t\|_2 \leq 1, \quad t = 1, \ldots, N.
\end{aligned} \quad (8)$$

Depending on the feature map this problem may still be infinite dimensional. To obtain a finite dimensional problem first define $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N]$, $\boldsymbol{\Phi} = [\boldsymbol{\varphi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\varphi}(\boldsymbol{x}_N)]$ and

$$\boldsymbol{D} = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{N \times N}.$$

Then the equality constraints of (8) can be rewritten as $\boldsymbol{V}\boldsymbol{D} = \boldsymbol{\Phi}\boldsymbol{A}$. The inverse $\boldsymbol{D}$ is a lower triangular matrix of all ones and its t-th column is $\boldsymbol{l}_t$. Therefore $\boldsymbol{v}_l = \boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{l}_l$. Finally squaring the inequality constraints one obtains $\boldsymbol{l}_t^T \boldsymbol{A}^T \boldsymbol{\Omega} \boldsymbol{A} \boldsymbol{l}_l \leq 1$. This allows the possibly infinite dimensional problem (8) to be written just in terms of the finite number of Lagrange multipliers α_t as (6). □

4.2 Recovering the sparsity pattern and a predictive model

Instead of a problem in $N \cdot n_h$ in $\boldsymbol{w}_t$ as in (5) we reduced the problem to just N variables in (6). Yet to use the solution for prediction we also need to rewrite the model (2) in terms of the dual variables. As the primal problem is sparse we would also like to recover the sparsity pattern.

Lemma 2. Denote the value of the norm $\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2$ by ν_t and let $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_N]^T$. Then the sparsity pattern follows from

$$\min_{\boldsymbol{\nu}} \quad \|\boldsymbol{\nu}\|_1 \quad \text{subject to} \quad \boldsymbol{y} - \gamma^{-1}\boldsymbol{\alpha} = [\boldsymbol{\Omega}\boldsymbol{A}\boldsymbol{L}]_{\llcorner} \boldsymbol{\nu}, \quad (9)$$

where $L = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N]$ and $[\cdot]_{\llcorner}$ is an operator returning the lower triangular part of a matrix $([\boldsymbol{X}]_{\llcorner})_{ij} = X_{ij}$ if $j \leq i$ and zero otherwise.

Proof. On the one hand recall that $\boldsymbol{v}_t = \boldsymbol{\Phi}\boldsymbol{A}\boldsymbol{l}_t$ and on the other hand note that $\boldsymbol{w}_t - \boldsymbol{w}_{t-1} = \nu_t \boldsymbol{v}_t$. Solving the latter relation recursively for $\boldsymbol{w}_t$ one obtains $\boldsymbol{w}_t = \sum_{k=1}^{t} \nu_k \boldsymbol{v}_k$. Exploiting this and the former relation the primal problem (5) can be rewritten as (9). □

Remark 3. Note that the optimization problem in (9) is a special case of basis pursuit (Chen et al., 2001) for a specific matrix $[\boldsymbol{\Omega}\boldsymbol{A}\boldsymbol{L}]_{\llcorner}$. For its solution many efficient algorithms have been proposed like `SPGL1` (Van Den Berg and Friedlander, 2008) and `NESTA` (Becker et al., 2009).

Finally we can state a predictive equation in terms of the dual variables.

Corollary 4. A prediction at a new point $\boldsymbol{z}_t$ at time $t \in \{1, \ldots, N\}$ is obtained by

$$f_t(\boldsymbol{z}_t) = \sum_{k=1}^{N} \alpha_k^{(t)} K(\boldsymbol{x}_k, \boldsymbol{z}_t) \quad (10)$$

with $\boldsymbol{\alpha}^{(t)} = [\alpha_1^{(t)}, \ldots, \alpha_N^{(t)}]^T$ and $\boldsymbol{\alpha}^{(t)} = \sum_{k=1}^{t} \nu_k \boldsymbol{A}\boldsymbol{l}_k$.

Proof. First substitute the expressions for $\boldsymbol{w}_t$ obtained in the proof of Lemma 2 into the model equation (2). Then the dual model (10) follows from replacing the inner products of the feature map by the kernel function. □

Remark 5. Due to the sparsity induced by the sum-of-norms regularization the model will only change rarely and will otherwise stay constant over time. It fact many ν_k in the growing sum defining $\boldsymbol{\alpha}^{(t)}$ will be zero, which could be used to rewrite the equations depending on the identified segments as $f_c(\cdot)$ instead of dependent on time t.

Remark 6. Note that the predictions obtained from (10) depend on the time instant t at which a new point $\boldsymbol{z}_t$ is acquired. This requirement could be relaxed if the operation region c that generated the new point would be known. In general this will not be the case, therefore the primary use of this model is in validation schemes for model selection. This is in contrast to LS-SVM models (Suykens et al., 2002) whose predictions are independent of time.

In the following we will summarize the steps needed to obtain a predictive model in the dual.

Algorithm 1. (Model estimation).

(1) Choose a regularization constant γ.
(2) Compute the kernel matrix $\Omega_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and its matrix square root $\boldsymbol{G}$ such that $\boldsymbol{\Omega} = \boldsymbol{G}^T\boldsymbol{G}$.
(3) Solve the dual estimation problem (6) for the optimal Lagrange multipliers $\boldsymbol{\alpha}$.
(4) Solve (9) for $\boldsymbol{\nu}$ to recover the sparsity pattern of the primal problem.
(5) Evaluate (10) to obtain predictions.

5. MODEL SELECTION

As mentioned at the end of Section 2 the regularization constant γ needs to be selected. Additionally in the primal formulation (5) the basis functions need to be fixed or in the dual (6) a kernel function needs to be chosen. In the nonlinear setting described here the regularization parameter is crucial to control the complexity of the nonlinear model and at the same time to induce sparsity such that the change points $\{t_c\}_{c=1}^{C}$ can be discovered.

The main objective of this paper is to identify the correct change points. Estimating a single model on a known segment using (3) will always be able to outperform the joint model obtained from (5). A model specific for a single segment has better control over the complexity versus fit trade-off and is also able to select a better suited set of basis functions (or kernel in the dual). Therefore we suggest to re-estimate models on the individual segments using (3) once the change points are identified.

The modeling power of the global description (5) needs to be powerful enough to capture the nonlinear dynamics to a large extent to be able to detect changes. Therefore we propose to select γ according to generalization performance of the models using a validation scheme (see *e.g.,* Hastie et al., 2009). Once model parameters have been estimated they can be visualized as $\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2$ over t (see Fig. 2). In case sparsity is not perfect one can either perform thresholding, clustering or a combination thereof.

6. ALGORITHM

The primal problem (5), if finite dimensional and given an explicit expression for the feature map $\boldsymbol{\varphi}$, as well as the kernel based dual (8) can be solved using general purpose Second Order Cone Programming (SOCP) solvers like `Sedumi` (Sturm, 1999) which is especially easy with modeling tools like `YALMIP` (Löfberg, 2004). Yet the large number of constraints of any of the two problems makes their solution slow or even infeasible. Therefore we propose a multi stage procedure that makes it possible to tackle larger problems. Each of the techniques presented in the next sections can be used independently.

6.1 Active set strategy

For a clearer motivation consider an equivalent formulation of the dual optimization problem (6).

Lemma 7. Let $\boldsymbol{\alpha} = \gamma\boldsymbol{\alpha}'$ and $\boldsymbol{A}' = \operatorname{diag}(\alpha'_1, \ldots, \alpha'_N)$ then

$$\begin{aligned} \min_{\boldsymbol{\alpha}'} \quad & \|\boldsymbol{\alpha}' - \boldsymbol{y}\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{G}\boldsymbol{A}'\boldsymbol{l}_t\|_2 \leq \frac{1}{\gamma}, \quad t = 1, \ldots, N, \end{aligned} \tag{11}$$

is equivalent to (6).

Proof. For convenience we consider the equivalent minimization problem to (6) with the negated cost function $\frac{1}{\gamma}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \boldsymbol{y}^T\boldsymbol{\alpha}$. Note that adding constant terms to the objective function or rescaling it does not change the optimal solution. Therefore we can modify the cost to $\boldsymbol{\alpha}^T\boldsymbol{\alpha} - 2\gamma\boldsymbol{y}^T\boldsymbol{\alpha} + \gamma^2\boldsymbol{y}^T\boldsymbol{y} = \gamma^2\|\gamma^{-1}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2$. Performing a change of variables from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha}'$ the equivalent optimization problem (11) is obtained. □

The rewritten dual problem (11) suggests that depending on the value of the regularization constant γ many of the constraints in it will not be active, *i.e.,* $\|\boldsymbol{G}\boldsymbol{A}'\boldsymbol{l}_t\|_2 < \gamma^{-1}$. Therefore omitting these constraints does not change the solution. This motivates the use of an active set strategy. Starting with a single constraint, the most violating constraint is successively added to the set of active constraints as formalized in the following procedure.

Algorithm 2. (Active set strategy).

(1) Initialize $\mathcal{I} = \{1\}$.
(2) Solve
$$\min_{\boldsymbol{\alpha}} \quad \|\boldsymbol{\alpha} - \gamma\boldsymbol{y}\|_2^2 \text{ subject to } \|\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_i\| \leq 1, \quad i \in \mathcal{I}. \tag{12}$$
(3) Compute $i = \arg\max_{1 \leq t \leq N}\{\|\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_t\|\}$.
(4) If $\|\boldsymbol{G}\boldsymbol{A}'\boldsymbol{l}_i\| \leq \gamma^{-1}$ then terminate.
(5) Else $\mathcal{I} := \mathcal{I} \cup \{i\}$ and goto (2).

In our experiments (for an example see Fig. 7) we observed that only a small fraction of the whole number of constraints is needed to define the final solution. A similar approach has been presented in Jenatton et al. (2009).

6.2 Augmented Lagrangian

After an initial solution for (12) has been obtained the optimal solution will likely only change gradually over the iterations needed to satisfy all constraints. Therefore a good initial guess for the new solution is given by the solution of the last iteration. Interior-point solvers like `Sedumi` are in general hard to warm start such that there is no benefit of having a good initial guess. In contrast first order schemes are very easy to warm start and many efficient algorithms have been developed for related problems (*e.g.,* `SPGL1` and `NESTA` mentioned earlier). Most first order schemes are based on the idea of projected gradients. The main requirement of these algorithms is that the projection onto the constraint set is cheap. In its current form (12) that is not the case. By introducing new variables $\boldsymbol{\beta}_i = \boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_i$ the norm constraints simplify to $\|\boldsymbol{\beta}_i\|_2 \leq 1$. Now the projection on the norm constraint is just a matter of rescaling $\boldsymbol{\beta}_k$ such that it does not exceed unit norm. Unfortunately handling of the equality constraints for $\boldsymbol{\beta}_i$ is not directly possible in gradient projection algorithms. Therefore we propose to use an augmented Lagrangian algorithm (Nocedal and Wright, 2006) and define an augmented cost function that incorporates the equality constraints

$$\begin{aligned} g_\mu(\boldsymbol{\alpha}, \boldsymbol{\beta}_i, \boldsymbol{\lambda}_i) := \|\boldsymbol{\alpha} - \gamma\boldsymbol{y}\|_2^2 - \sum_{i \in \mathcal{I}} \boldsymbol{\lambda}_i^T(\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_i - \boldsymbol{\beta}_i) \\ + \frac{\mu}{2}\sum_{i \in \mathcal{I}} \|\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_i - \boldsymbol{\beta}_i\|_2^2. \end{aligned}$$

One can show that the following algorithm converges.

Algorithm 3. (Method of multipliers).

(1) Initialize μ.
(2) Set $\boldsymbol{\lambda}'_i = \boldsymbol{\lambda}_i - \mu(\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_i - \boldsymbol{\beta}_i)$ for $i \in \mathcal{I}$.
(3) Solve
$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}_i} \; g_\mu(\boldsymbol{\alpha}, \boldsymbol{\beta}_i, \boldsymbol{\lambda}_i) \text{ subject to } \|\boldsymbol{\beta}_i\| \leq 1 \; \forall \; i \in \mathcal{I}. \tag{13}$$
(4) Terminate if $\|\boldsymbol{G}\boldsymbol{A}\boldsymbol{l}_i - \boldsymbol{\beta}_i\|$ is small enough.
(5) Increase μ and goto (2).

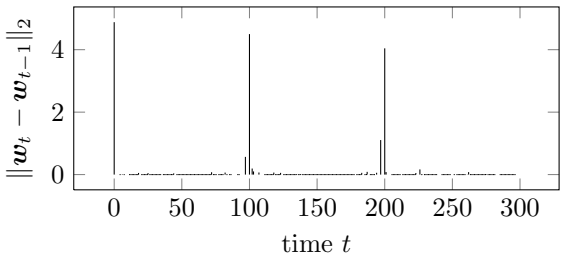

Fig. 2. Sparsity pattern obtained from (9) for the nonlinear system described in Sec. 7.1 switching to a different dynamic at $t_1 = 100$ and switching back to the initial dynamics at $t_2 = 200$.

For a detailed description see Alg. 17.4 in Nocedal and Wright (2006).

6.3 Accelerated gradient projection

To solve (13) in a timely fashion we applied Nesterov's optimal first order algorithm (Nesterov, 2005) for non smooth convex optimization. The convergence rate of the steepest descent algorithm can be increased from $\mathcal{O}(\frac{1}{\epsilon})$ to $\mathcal{O}(\frac{1}{\epsilon^2})$ by maintaining an additional sequence of linear combinations of past gradients. An accessible description of such an algorithm that can be easily adapted to (13) is presented in Liu et al. (2009).

7. EXPERIMENTS

7.1 NFIR Hammerstein system

We consider a simple Hammerstein type system with $y_t = [b_{1,t}\, b_{2,t}]\,\mathrm{sinc}(\boldsymbol{x}_t) + e_t$, $\boldsymbol{x}_t = [u_t, u_{t-1}]^T$ with $\mathrm{sinc}(\cdot)$ applied elementwise. The input signal u_t and the noise e_t are white and Gaussian. The noise is scaled such that the data has a signal to noise ratio of 10 dB, while the input signal has unit variance. The parameters are chosen as

$$(b_{1,t}, b_{2,t}) = \begin{cases} (5,-2), & 100 < t \leq 200, \\ (1,2), & \text{otherwise.} \end{cases}$$

We generate 900 equally spaced samples in the time interval $1 \leq t \leq 300$ which we split into three parts by taking every third sample, one for estimation, one for model selection and one for the final evaluation of the model performance. The model is used in its dual formulation (6) and a RBF kernel is applied with the bandwidth σ fixed to 1. The regularization parameter γ is selected according to prediction performance on the validation set. The obtained sparsity pattern in shown in Fig. 2. We observe that the procedure correctly isolated the two change points and the initial model. However, especially close to the change points, there are some small spurious components. In Fig. 3 the pairwise norms $\|\boldsymbol{w}_k - \boldsymbol{w}_l\|_2$ for all differences $\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2$ that are at least as big as 10^{-3} times the largest one are shown. One can clearly see only three segments are really significant and that the first and the third segment share the same dynamics.

The predictions on the independent test data as well as the residuals are shown in Fig. 4. The root mean squared error on the whole test data is 0.1905 for the piecewise nonlinear model, as a reference a LS-SVM trained on the whole data (without knowledge of the segments) achieves a RMSE of 0.6638 on the test set.

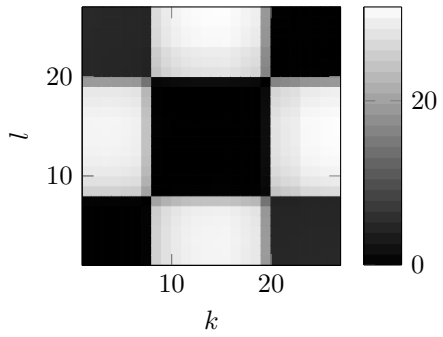

Fig. 3. Norm $\|\boldsymbol{w}_k - \boldsymbol{w}_l\|_2$ for $k, l \in \{t : \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2 \geq 10^{-3} \max_t \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2\}$. The corresponding nonlinear system is specified in Sec. 7.1.

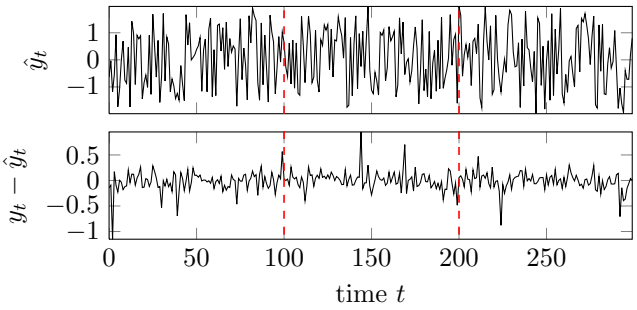

Fig. 4. Predictions (top panel) of the model described in Sec. 7.1 on independent test data and the corresponding modeling errors (bottom panel). The vertical dashed lines indicate the positions of the change points.

Table 1. Root mean squared error (RMSE) on independent test data for different sections (I: $1 \leq t < 100$, II: $100 \leq t < 200$, III: $200 \leq t < 300$) of Example 1. Piecewise nonlinear model Alg. 1 (PW NL), piecewise ARX model (Ohlsson et al., 2010) (PL ARX), LS-SVM given the true change points Eq. 3 (Suykens et al., 2002).

	PW NL (Alg. 1)		PW ARX	LS-SVM (Eq. 3)	
	RMSE	$\sigma = 1$, γ	RMSE	RMSE	(σ, γ)
I	0.206	0.468	1.045	0.163	(1.274, 33.6)
II	0.185	0.468	1.003	0.140	(1.274, 297.6)
III	0.180	0.468	1.118	0.126	(1.274, 33.6)

Let us now compare with a segmented ARX model. With $\boldsymbol{x}_t = [u_t\, u_{t-1}]^T$, a model (1) is sought using the same scheme as proposed in Ohlsson et al. (2010). If the prediction performance on the validation set is used to find the number of segments, no change points are found. The estimated ARX parameters are therefore equivalent to those of a least squares estimate on the whole estimation data set. The prediction performance on the test data yields a root mean squared error of 1.060.

Finally we compare with a LS-SVM (3) model given the true segmentation and apply full model selection *i.e.,* the bandwidth σ as well as the regularization parameter γ are

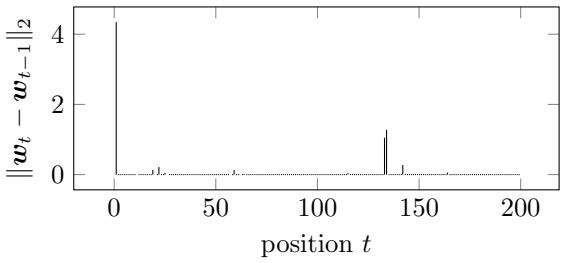

Fig. 5. Sparsity pattern obtained from (9) for the nonlinear system described in Sec. 7.2 switching to a different dynamic between positions 133 and 135.

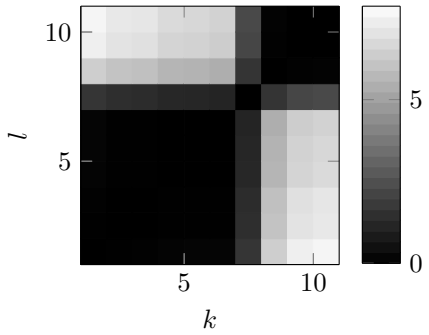

Fig. 6. Norm $\|\boldsymbol{w}_k - \boldsymbol{w}_l\|_2$ for $k, l \in \{t : \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2 \geq 10^{-3} \max_t \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2\}$. The corresponding nonlinear system is specified in Sec. 7.2.

selected based on prediction performance. We train one model for the second segment and one model with the combined data of the remaining two. The results of the piecewise nonlinear model, the piecewise ARX model and the segment-wise LS-SVM are reported in Table 1.

7.2 NARX Wiener system

As second example we consider a Wiener type system with ARX structure, defined by $y_t = \sin(\frac{\pi}{2}\,\boldsymbol{\theta}_t^T \boldsymbol{x}_t) + e_t$ and $\boldsymbol{x}_t = [y_{t-1}, y_{t-2}, u_t, u_{t-1}, u_{t-2}]^T$. The input signal u_t and the noise term e_t are zero mean white Gaussian. The noise has variance 0.1^2 and the input is scaled such that it is in the interval $[-1, 1]$. The parameter vector $\boldsymbol{\theta}_t$ is scaled to unit mean and chosen as $\boldsymbol{\theta}_t|_{t=1}^{400} = [-0.525,\ 0.096,\ 01.585,\ -0.562, 0.542,\ -0.135]^T$ and $\boldsymbol{\theta}_t|_{t=401}^{600} = [-1.168,\ -1.401,\ 2.178,\ 1.334,\ 0.247,\ -0.190]^T$. Again the data is split into three parts by taking every third sample. Therefore the estimation data at position 134 ($t = 400$) is governed by a different system than the corresponding sample ($t = 401$) in the validation data. The kernel function is again a RBF kernel with fixed bandwidth $\sigma = 1$ and the regularization parameter γ is selected based on validation performance. The resulting sparsity pattern is depicted in Fig. 5. The initial model at position 1 is clearly visible. Around position 134 we observe two significant peaks. The energy of this change is spread over two positions as there is a mismatch of dynamics in the estimation and the validation data sets at position 134. This can also be seen from the pairwise differences in Fig. 6. We clearly see two blocks that share the same dynamics, but the model at one position correlates well with the models before and after it. The root mean squared errors on an independent test set are for segment I: 0.223 (0.190), segment II: 0.592 (0.485) and total: 0.390 (0.319). The values given in parenthesis are the performances of a LS-SVM model given the true segmentation (3) and using a full model selection.

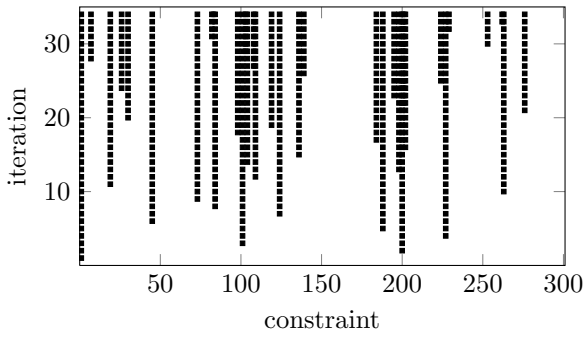

Fig. 7. Active constraints as a function of iterations in the active set scheme described in Sec. 6.1. Black pixels indicate that a constraint belongs to the active set.

Let us compare to a segmented ARX model. The scheme proposed by Ohlsson et al. (2010) was used to estimate a segmented ARX model (1) with $\boldsymbol{x}_t = [y_{t-1}, y_{t-2}, u_t, u_{t-1}, u_{t-2}]$. The change point at 134 was correctly detected and the root mean squared errors on an independent test set was for segment I: 0.310, segment II: 0.600 and total: 0.428, which is slightly worse than the proposed method.

7.3 Algorithm

For the example in Sec. 7.1 and the optimal value for γ the evolution of the active set along the iterations is shown in Fig. 7. We observe that only a fraction of all constraints determine the optimal solution, in this case 34 out of 300. Also observe that the first constraints that are included in the active set are the ones at $t = 100$ and $t = 200$ namely the positions of the two change points.

8. CONCLUSIONS

A novel method for segmenting time-series from nonlinear dynamical systems has been proposed. The proposed method uses sum-of-norms to trade-off the number of segments and fit. Two examples motivate the use of a nonlinear underlying model instead of a linear used in previous work. The fact that the method only has one design parameter, the regularization parameter, makes it extremely user friendly and attractive for *e.g.,* change detection, diagnosis and fault detection.

ACKNOWLEDGEMENTS

T. Falck, J.A.K. Suykens and B. De Moor are supported by Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and

optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940); Contract Research: AMINAL; Other: Helmholtz: viCERP, ACCM. J.A.K. Suykens is a professor and B. De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. H. Ohlsson and L. Ljung are supported by the Swedish foundation for strategic research in the center MOVIII and by the Swedish Research Council in the Linnaeus center CADICS.

REFERENCES

Andersson, P. (1985). Adaptive forgetting in recursive identification through multiple models. *International Journal of Control*, 42(5), 1175–1193.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3), 243–272.

Becker, S., Bobin, J., and Candes, E.J. (2009). Nesta: A fast and accurate first-order method for sparse recovery. Technical report, California Institute of Technology, Pasadena, CA, USA. URL `http://arxiv.org/pdf/0904.3367`.

Bodenstein, G. and Praetorius, H.M. (1977). Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*, 65, 642–652.

Boyd, S.P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Candès, E.J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52, 489–509.

Chen, S.S., Donoho, D.L., and Saunders, M.A. (2001). Atomic Decomposition by Basis Pursuit. *SIAM Review*, 43(1), 33–61.

Donoho, D.L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4), 1289–1306.

Espinoza, M., Suykens, J.A.K., Belmans, R., and De Moor, B. (2007). Electric Load Forecasting - Using kernel based modeling for nonlinear system identification. *IEEE Control Systems Magazine*, 27, 43–57.

Falck, T., Suykens, J.A.K., and De Moor, B. (2009). Robustness analysis for least squares kernel based regression: an optimization approach. In *Proc. 48th IEEE Conf. on Decision and Control CDC*, 6774–6779.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Jenatton, R., Audibert, J.Y., and Bach, F.R. (2009). Active Set Algorithm for Structured Sparsity-Inducing Norms. In *Proceedings of the 2nd NIPS Workshop on Optimization for Machine Learning*, 6. Whistler, Canada.

Kim, S.J., Koh, K., Boyd, S.P., and Gorinevsky, D. (2009). ℓ_1 Trend Filtering. *SIAM Review*, 51(2), 339–360.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, 5, 81–91.

Liu, J., Ji, S., and Ye, J. (2009). Multi-Task Feature Learning Via Efficient $L_{2,1}$-Norm Minimization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 10. Montreal, Canada.

Ljung, L. (2007). *The System Identification Toolbox: The Manual*. The MathWorks Inc. 1st edition 1986, 7th edition 2007, Natick, MA, USA.

Löfberg, J. (2004). YALMIP : A Toolbox for Modeling and Optimization in MATLAB. In *Proceedings of the 2004 IEEE International Symposium on Computer Aided Control Systems Design*, 284–289. Taipei, Taiwan.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming, Series A*, 103(1), 127–152.

Nocedal, J. and Wright, S.J. (2006). *Numerical Optimization*. Springer, New York, NY, USA, 2nd edition.

Ohlsson, H., Ljung, L., and Boyd, S.P. (2010). Segmentation of ARX-models using sum-of-norms regularization. *Automatica*, 46(6), 1107–1111.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian processes for machine learning*. Springer.

Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels*. MIT Press Cambridge, Mass.

Shivaswamy, P.K., Bhattacharyya, C., and Smola, A.J. (2006). Second Order Cone Programming Approaches for Handling Missing and Uncertain Data. *Journal Machine Learning Research*, 7, 1283–1314.

Sturm, J. (1999). Using SeDuMi 1.02, A Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1), 625–653.

Suykens, J.A.K., Alzate, C., and Pelckmans, K. (2010). Primal and dual model representations in kernel-based learning. *Statistics Surveys*, 4, 148–183. doi:10.1214/09-SS052.

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

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

Tugnait, J.K. (1982). Detection and estimation for abruptly changing systems. *Automatica*, 18, 607–615.

Van Den Berg, E. and Friedlander, M.P. (2008). Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2), 890–912.

Vapnik, V.N. (1998). *Statistical Learning Theory*. John Wiley & Sons.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.