Identification of switched linear regression models using sum-of-norms regularization

Henrik Ohlsson and Lennart Ljung

Linköping University Post Print

N.B.: When citing this work, cite the original article.

Original Publication:

Henrik Ohlsson and Lennart Ljung, Identification of switched linear regression models using sum-of-norms regularization, 2013, Automatica, (49), 4, 1045-1050.

<http://dx.doi.org/10.1016/j.automatica.2013.01.031>

Copyright: Elsevier

<http://www.elsevier.com/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-92612>

Identification of Switched Linear Regression Models Using Sum-of-Norms Regularization [★]

Henrik Ohlsson ^{*}, Lennart Ljung

Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden

Abstract

This paper proposes a general convex framework for the identification of switched linear systems. The proposed framework uses over-parameterization to avoid solving the otherwise combinatorially forbidding identification problem and takes the form of a least-squares problem with a sum-of-norms regularization, a generalization of the ℓ_1 -regularization. The regularization constant regulates complexity and is used to trade off fit and the number of submodels.

Key words: Regularization; system identification; sum-of-norms; switched linear systems, piecewise affine systems.

1 Introduction

We shall in this contribution consider the following problem: Given measured values of an “output” $y(t)$, a “regression vector” $\varphi(t)$ and a “regime variable” $p(t)$, $t = 1, \dots, N$, find a switched model $\theta_i, \mathcal{H}_i, i = 1, \dots, d$,

$$y(t) = \theta_i^T \varphi(t) + e(t), \quad \text{if } p(t) \in \mathcal{H}_i, \quad (1)$$

where $e(t)$ is some additive noise. The problem is to find both the number of linear submodels d , the partitioning of the regime variable space $\Omega = \cup_{i=1}^d \mathcal{H}_i$, and the associated parameter values θ_i .

1.1 LPV and PWA Models

The problem formulation contains several common estimation tasks such as model and signal segmentation, piece-wise affine systems, hybrid system modeling and linear parameter-varying (LPV) system identification. Estimation of such models is of considerable current interest, and we may refer to several recent publications for interesting related results, [1,20,31,30,23].

Perhaps the most obvious case is when the regime variable $p(t)$ is the regression vector $\varphi(t)$ itself. Then the

model (1) becomes a nonlinear model that is piecewise affine in the regressor space:

$$y(t) = \theta_i^T \varphi(t) + e(t), \quad \text{if } \varphi(t) \in \mathcal{H}_i \quad (2)$$

This is the *Piecewise Affine* (PWA) model class. It contains as a special case the *Piecewise ARX* (PWARX) models where the regressors are made up of delayed inputs and outputs.

PWA systems serve as popular models of nonlinear systems due to their universal approximation properties [21,6]. In addition, it can also be shown that PWA systems are equivalent to certain types of hybrid systems, see *e.g.*, [2,14]. This makes PWA systems a very important class of systems with an increasing interest. Five methods that have attained special attention in the literature are the clustering-based approach [10], the bounded error approach [3], the mixed integer quadratic programming approach [4,35], the Bayesian approach [19] and the algebraic approach [39]. For an overview of contributions see [32,11]. The identification of PWA models is a complex task in which, simultaneously, both the partitions and the linear models have to be found. The underlying problem is often non-convex and most methods can be seen as local searches. These are then highly dependent on a good initialization for delivering a satisfying model. See *e.g.*, [33] for an overview. Very few methods can guarantee the finding of the global optimum and hence being independent of initialization. One approach that achieves this is the mixed integer quadratic programming approach [4,35]. However, such programs are known to be hard to solve (NP-hard in the

[★] A preliminary version of this article [26] was presented at the 18th IFAC World Congress, Milan, Italy, 2011.

^{*} Corresponding author. Tel.: +46 13 281000; fax: +46 13 282622.

Email addresses: ohlsson@isy.liu.se (Henrik Ohlsson), ljung@isy.liu.se (Lennart Ljung).

worst case [35]) and the approach is therefore practically applicable only to very small problems. Our approach, which we first discussed in [25], is rather different from the above mentioned in that it approximates the underlying optimization problem with a convex relaxed problem. It is therefore insensitive to initialization, since it is convex, while being solvable for problems of practical sizes. A number of papers have followed our methodology and developed new methods for PWA identification using relaxations, convex optimization and sparsity, see *e.g.*, [23,1].

1.2 Segmentation of Models and Signals

Another application of (1) of considerable interest is when $p(t) = t$ *i.e.*,

$$y(t) = \theta_i^T \varphi(t) + e(t), \quad \text{if } t \in \mathcal{H}_i, \quad (3)$$

which is a piecewise constant model over time; a *segmented model*. There are two important special cases of (3):

- the time segments are intervals $\mathcal{H}_i = [t_i, t_{i+1}]$. This corresponds to a time-varying system with piecewise constant model, and was considered *e.g.*, in [28,31].
- the time segments contain several intervals, like $\mathcal{H}_i = [[t_1, t_2], [t_3, t_4], \dots]$. This describes the situation where the model parameters may return to previous values after a while, like in a Markov chain. Note that also a PWA model is a special case.

2 Proposed Method

2.1 Basic Idea

Let h be a function that takes a set of vectors as argument and returns the number of different elements in the set:

$$h(\theta(1), \theta(2), \dots, \theta(N)) = \# \text{ of different } \theta \text{ vectors} \quad (4)$$

Assume that $\{(y(t), \varphi(t), p(t))\}_{t=1}^N$, $y(t) \in R$, $\varphi(t) \in R^{n_\varphi}$, $p(t) \in R^{n_p}$, is available and consider a model of the form

$$y(t) = \theta^T(t) \varphi(t). \quad (5)$$

Let the fit to data be measured by the least-squares (LS) criterion. We hence find the model parameters using

$$\min_{\theta(t), t=1, \dots, N} \sum_{t=1}^N (y(t) - \theta^T(t) \varphi(t))^2. \quad (6a)$$

This will of course always give a perfect fit and a model of the form (1) with $d = N$. Say that we seek $d < N$ submodels, we then solve (6a) subject to the constraint

$$h(\theta(1), \theta(2), \dots, \theta(N)) = d. \quad (6b)$$

This constraint can be added to the criterion with a Lagrange multiplier $\lambda > 0$ or as a regularization term:

$$\min_{\theta(t), t=1, \dots, N} \sum_{t=1}^N (y(t) - \theta^T(t) \varphi(t))^2 + \lambda h(\theta(1), \theta(2), \dots, \theta(N)) \quad (7)$$

2.2 Counting the Number of Different Vectors

The function h is well defined, but a bit awkward to handle algorithmically, due to its combinatorial nature. A possibility is to approximate it by

$$h(\theta(1), \theta(2), \dots, \theta(N)) \approx \sum_{t=1}^N \sum_{s=1}^N \left\| \|\theta(s) - \theta(t)\| \right\|_0. \quad (8)$$

Here, and in the rest of the paper, $\|x\| \triangleq \sqrt{x^T x}$ is the regular 2-norm, and $\|\cdot\|_0$ is the “zero norm”, counting the number of non-zero elements of the vector. Note that since the argument of the zero norm is just a scalar in (8), the zero norm simply returns a 1 if its argument is nonzero and 0 otherwise.

It is important to realize that (8) is a rather crude approximation: Many different vectors will be counted more than once. In fact, if there are d different models, such that model j is the same for k_j values of t ($N = \sum_{j=1}^d k_j$), (8) takes the value $\sum_{i,j=1}^d k_i k_j$ instead of d . This means that partitions with many θ in some and few in others carry a less penalty than more equally weighted partitions. To alleviate this, a weighing or a “kernel” $K(p(t), p(s)) : R^{n_p} \times R^{n_p} \rightarrow R^+$ can be used to decrease the penalty on two parameter vectors being different, when this is “natural” as seen from the regime variables. We hence have

$$h(\cdot) \approx \sum_{t=1}^N \sum_{s=1}^N K(p(t), p(s)) \left\| \|\theta(s) - \theta(t)\| \right\|_0. \quad (9)$$

We shall give examples of useful kernels shortly.

2.3 A Convex Relaxation

The criterion (7) with (9) is still hard to minimize due to the combinatorial aspects of the zero norm. We therefore suggest to replace the zero norm by the ℓ_1 -norm. This is entirely in line with the much used Lasso method [37] and compressive sensing [7,8]. Note that the term inside the zero norm in (9) is always positive. Replacing the zero norm with the ℓ_1 -norm therefore leads to the sum-

of-norms regularized least squares criterion

$$\min_{\theta(t), t=1, \dots, N} \sum_{t=1}^N (y(t) - \theta^T(t) \varphi(t))^2 + \lambda \sum_{t=1}^N \sum_{s=1}^N K(p(t), p(s)) \|\theta(s) - \theta(t)\|. \quad (10)$$

Note that the ℓ_1 -norm vanishes since all its arguments are positive.

There are two key reasons why the criterion (10) is attractive for identifying models of the form (1):

- It is a convex optimization problem, so the global solution can be computed efficiently.
- The sum-of-norms form of the regularization favors sparse solutions where “many” (depending on λ) of the regularized variables come out as exactly zero in the solution. In this case, this implies that many of the estimates of $\theta(t)$ become identical, so the criterion (10) returns d distinct model vectors $\theta_i, i = 1, \dots, d$, where d is controlled by λ . It then remains to associate these vectors with corresponding partitions, see Section 2.5.

The downside of using an ℓ_1 -norm instead of the zero norm is that the ℓ_1 -norm penalizes the size of the regularized variable and not only if it is non zero or not, like the zero norm. The regularized variable ($K(p(t), p(s)) \|\theta(s) - \theta(t)\|$ in (10)) will therefore be biased toward zero.

2.4 Kernel

If the kernel is chosen so that

$$K(p(t), p(s)) \triangleq \begin{cases} 1 & \text{if } p(t) \in \mathcal{H}_i \text{ and } p(s) \in \mathcal{H}_i, \\ & \text{for some } i = 1, \dots, d, \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

then there is a $\lambda^* > 0$ such that the true partition of the observed data is recovered whenever $\lambda > \lambda^*$. The kernel (11) is of course not available in practice since it relies on that the partition is already known. However, this insight tells us that the kernel should be chosen as a prior for the clustering. We will use the following kernel, which is suitable for finding PWA models, in many of our examples:

$$K_n(p(t), p(s)) \triangleq \begin{cases} 1 & \text{if } p(t) \text{ is one of the } n \text{ closest} \\ & \text{neighbors of } p(s) \text{ among all} \\ & \text{the observations,} \\ 0 & \text{otherwise,} \end{cases} \quad (12)$$

where “closest” is measured using the 2-norm and n is a positive integer. Note that if this kernel is used, we implicitly assume that the function is locally linear. This assumption was also discussed in [10] and there are therefore some interesting connections between proposed method and the method proposed in [10].

2.5 Estimating the Shapes of the Partitions

The output of (10) is an estimate of θ for each output-regressor-regime variable triple in the training data set $\{(y(t), \varphi(t), p(t))\}_{t=1}^N$. To find the shapes of the partitions, we apply a classification algorithm to $\{(p(t), \theta(t))\}$.

Let $p(t)$ be the regressor and $\theta(t)$ the classification label. A classification algorithm can then be used to partition the regime variable space into regions, each region having a different θ -value (or class label) associated with it. If the number of distinct θ -estimates of (10) is d , we denote these $\theta_1, \dots, \theta_d$ and the associated partitions by $\mathcal{H}_1, \dots, \mathcal{H}_d$, we hence have a switched model of the form (1).

In the coming numerical illustrations we have chosen to use a *Support Vector Machines* (SVM, [38]) classifier. The SVM classifier was applied using a *one-versus-the-rest* approach. That is, a set of training data points having the same estimated θ -value was seen as belonging to one class and the rest of the training data to a second class. This simple version of SVM was sufficient in our example but would have performed poorly on more complicated regions. We refer the interested reader to [32, Sect. 4.2] for a discussion on alternative approaches.

2.6 Estimating the Number of Submodels and Critical Parameter Value

The only design parameter of the proposed method is the regularization parameter λ . If d is known, λ can easily be tuned to give d different θ vectors. If this is not the case but a validation data set is available, this should preferably be used to find a suitable choice for λ . If neither d or a validation data set is available, a λ -value can be found by studying how the fit to the training data varies with λ . In doing so, it can be very useful to know in what interval it is sensible to look for a suitable λ -value.

A basic result from convex analysis (see Remark 2 in [29], also cf. pp. 277–278 in [5]) tells us that there is a value $\lambda^{\max}$ for which the solution of the problem is constant, i.e., $\theta(t), t = 1, \dots, N$, does not vary with t , if and only if $\lambda \geq \lambda^{\max}$. In other words, $\lambda^{\max}$ gives the threshold above which there is only one submodel. Reasonable values for λ are typically on the order of $0.01\lambda^{\max}$ to $\lambda^{\max}$.

Proposition 1 (Critical Parameter Value $\lambda^{\max}$)

Let θ^{ls} be the optimal constant parameter vector, i.e.,

the solution of the least squares problem

$$\min_{\theta} \sum_{t=1}^N (y(t) - \varphi^T(t)\theta)^2. \quad (13)$$

We then find $\lambda^{\max}$ as the solution to the optimization problem

$$\min_{\lambda, z(s,t), s,t=1,\dots,N} \lambda \quad (14a)$$

$$s.t. \|z(s,t)\| \leq \lambda, \quad (14b)$$

$$z(s,s) = 0, \quad (14c)$$

$$z(s,t) = z(t,s), \quad (14d)$$

$$2(y(t) - \varphi^T(t)\theta^{\text{ls}})\varphi^T(t) = \sum_{r=1}^{t-1} (K(p(t), p(r)) + K(p(r), p(t)))z(t,r) - \sum_{r=t+1}^N (K(p(t), p(r)) + K(p(r), p(t)))z(r,t), \quad (14e)$$

$$s, t = 1, \dots, N. \quad (14f)$$

This is a convex optimization problem and $\lambda^{\max}$ is hence readily computed. For the special case when $p(t) = t$ and the kernel given by

$$K(t,s) = \begin{cases} 1 & \text{if } t = s - 1, \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

which was the setup studied in [28], the expression for $\lambda^{\max}$ can be simplified into

$$\lambda^{\max} = \max_{t=2,\dots,N} \left\| \sum_{r=1}^t 2(y(r) - \varphi^T(r)\theta^{\text{ls}})\varphi^T(r) \right\|. \quad (16)$$

Due to page limitation, the interested reader is referred to the technical report [27] for the proof.

3 Solution Algorithms and Software

Many standard methods of convex optimization can be used to solve the problem (10). Systems such as CVX [13,12] or YALMIP [22] can readily handle the sum-of-norms regularization, by converting the problem to a cone problem and calling a standard interior-point method. Recently many authors have developed fast first order methods for solving ℓ_1 regularized problems, see for example, [34, §2.2].

The simulations shown in this paper were carried out in MATLAB using CVX. A code-package using CVX is available for download on <http://www.control.isy.liu.se/~ohlsson/code.html>.

4 Numerical Illustrations

Example 1 A One Dimensional PWARX System

Consider the one-dimensional PWARX system (introduced in [10])

$$y(t) = \begin{cases} u(t-1) + 2 + e(t), & -4 \leq u(t-1) \leq -1, \\ -u(t-1) + e(t), & -1 < u(t-1) < 2, \\ u(t-1) + 2 + e(t), & 2 \leq u(t-1) \leq 4. \end{cases} \quad (17)$$

Generate $\{u(t)\}_{t=1}^{50}$ by sample a uniform distribution $U(-4, 4)$ and let $e(t) \sim N(0, 0.05)$. Figure 1 shows the dataset $\{(y(t), u(t))\}_{t=1}^{50}$. To identify a PWARX model,

Fig. 1. Data used in Example 1 showed with $\triangle$, $\circ$ and $\diamond$ symbols. The data marked with the same symbol got the same θ -estimate using the proposed criteria (10),(12). Dashed lines show the estimated partitions obtained by applying SVM. The solid lines show the affine submodels.

we should let the kernel reflected the underlying assumption of a PWARX. That is, the kernel should reflect the assumptions that there are only a limited number of values that the model parameter can take and that its likely to return to the same value over time. We hence choose to use the kernel defined by (12) and set $\varphi(t) = [1 \ u(t-1)]^T$ and $p(t) = \varphi(t)$. Now, with $\lambda = 0.0015$ and $n = 8$ in the proposed criterion (10) with (12), the result shown in Figure 1 is produced. The results compare well with the result reported in [10]. Note however that even though the setup is the same as in [10] the data is not. Also note that since the kernel only expresses our desire to have identical model parameter values in a n -neighborhood of each regressor, we obtain 3 sub-models even though the true model parameter is the same for two of these.

Example 2 Segmented ARX

The system (17) is essentially a PWARX system. But it can also be modeled as a time-varying system where coefficients are piecewise constant in time, as in the first bullet of Section 1.2. In Figure 2 the result of the proposed criterion (10) with $p(t) = t$ and the kernel (15) is

shown with a dashed line. The same figure also shows how the true coefficient in front of $u(t-1)$ changes over time and the result obtained by using the kernel given in (12) with $p(t) = \varphi(t)$ and $n = 8$. Note the difference in using the kernel (12) and (15). The first mentioned grants that, over time, the model parameter value return to the same value as it previously has taken. The second kernel only grants that the model parameter value changes infrequently. Since the data was generated by a PWARX system, modeling using a PWARX model and the kernel (12) will, as seen, give a better fit to the measurement.

Fig. 2. Solid line: The system parameter coefficient for $u(t-1)$ in (17). Dashed line: The estimated parameter using the scheme proposed in Section 2 with kernel (15). The kernel (12) with $p(t) = t$ gives estimates that are indistinguishable from the true values in the resolution of the figure.

Example 3 A Hammerstein System

Consider the system

$$y(t) = -a_1 y(t-1) - a_2 y(t-2) + b_1 v(t-1) + e(t) \quad (18)$$

where v is a saturated version of u ,

$$v(t) = \begin{cases} u_{max} & \text{if } u(t) > u_{max}, \\ u(t) & \text{if } u_{min} \leq u(t) \leq u_{max}, \\ u_{min} & \text{if } u(t) < u_{min}. \end{cases} \quad (19)$$

The system, from u to y , is commonly referred to as a Hammerstein system. With $a_1 = 0.5$, $a_2 = 0.1$, $b_1 = 1$, $u_{max} = 1$ and $u_{min} = -1$ the system was simulated with u being white Gaussian noise with a variance of 4. The measurement noise was also set to white Gaussian noise but with a variance of 0.04. 250 samples were generated. This setup was also used in [24].

Since the saturation (19) is a piecewise affine function, the whole system between u and y will be piecewise affine.

It can be shown that it is given by

$$y(t) = \begin{cases} [-a_1 & -a_2 & 0 & b_1 u_{max}] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -u_{max}] \varphi(t) > 0, \\ [-a_1 & -a_2 & b_1 & 0] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -u_{max}] \varphi(t) < 0 \text{ and} \\ & [0 & 0 & 1 & -u_{min}] \varphi(t) < 0, \\ [-a_1 & -a_2 & 0 & b_1 u_{min}] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -u_{min}] \varphi(t) < 0, \end{cases}$$

with $\varphi(t) = [y(t-1) \ y(t-2) \ u(t-1) \ 1]^T$. If we let $p(t) = \varphi(t)$, use the kernel K_n as in (12), $n = 8$ and set $\lambda = 0.008$, we get an estimate

$$\hat{y}(t) = \begin{cases} [-0.45 & -0.05 & -0.02 & 2.08] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -2.24] \varphi(t) > 0, \\ [-0.50 & -0.10 & 0.92 & 0.06] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -2.24] \varphi(t) < 0 \text{ and} \\ & [0 & 0 & 1 & 1.56] \varphi(t) < 0, \\ [-0.49 & -0.10 & 0.00 & -1.04] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -1.56] \varphi(t) < 0, \end{cases} \quad (20)$$

using the scheme proposed in Section 2. The shapes of the partitions were estimated using a SVM classifier. These results compare well with the results

$$\hat{y}(t) = \begin{cases} [-0.49 & -0.12 & -0.05 & 1.81] \varphi(t) + e(t) & \text{if} \\ & [0.01 & 0.01 & -0.5 & -u_{max}] \varphi(t) > 0, \\ [-0.46 & -0.05 & 0.95 & 0.00] \varphi(t) + e(t) & \text{if} \\ & [0.01 & 0.01 & -0.5 & -u_{max}] \varphi(t) < 0 \text{ and} \\ & [0.01 & -0.04 & 1.01 & -u_{min}] \varphi(t) < 0, \\ [-0.53 & -0.13 & 0.05 & -0.86] \varphi(t) + e(t) & \text{if} \\ & [0.01 & -0.04 & 1.01 & -u_{min}] \varphi(t) < 0, \end{cases}$$

reported in [24]. Note however that even though the setup is the same as in [24] the data is not.

If we use an oracle to give what submodel respective sample comes from and then estimate the parameters using a least-squares fit, we obtain

$$\hat{y}(t) = \begin{cases} [-0.47 & -0.05 & -0.03 & 2.09] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -u_{max}] \varphi(t) > 0, \\ [-0.51 & -0.10 & 0.99 & 0.00] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -u_{max}] \varphi(t) < 0 \text{ and} \\ & [0 & 0 & 1 & -u_{min}] \varphi(t) < 0, \\ [-0.49 & -0.10 & 0.02 & -0.99] \varphi(t) + e(t) & \text{if} \\ & [0 & 0 & 1 & -u_{min}] \varphi(t) < 0, \end{cases}$$

which is very similar to what we obtained in (20) without the help of an oracle.

Example 4 A Pick-and-Place Machine

In this example we study a pick-and-place machine. The pick-and-place machine considered is used to place electronic components on a circuit board. Two modes can be distinguished, the free mode and the impact mode. When operating in the free mode, the pick-and-place machine is carrying an electronic component but is not in contact with the circuit board while in the impact move, the electronic component is in contact with the circuit board. For details on the setup see [16].

The data used is of a real physical process and also used in [15, 17, 3, 18]. It consists of a 15 s recording of the voltage input to the motor of the mounting head of the pick-and-place machine (will be referred to as input) and the vertical position of the mounting head (will be referred to as output). The input and output was sampled at 50 Hz and are shown in Figure 3. The first 8 s of data was used

Fig. 3. Top plot, measurements of the vertical position of the pick and place machine. Bottom plot, input voltage to the pick and place machine.

for estimation and the last 7 s for validation. A PWARX

$$\begin{aligned} y(t) &= \theta_i^T \varphi(t) + e(t) \quad \text{if } \varphi(t) \in \mathcal{H}_i \\ \varphi(t) &= [y(t-1) \ y(t-2) \ u(t-1) \ u(t-2)] \end{aligned} \quad (21)$$

with two submodels was identified using the proposed scheme using $\lambda = 1$, the kernel given in (12) and $n = 14$. The one-step ahead predictor obtain an almost perfect fit to the observed outputs.

A one-step ahead predictor is not of much interest if the model is going to be used for model predictive control. It is therefore of interest to see how well the model can handle simulation. The fit on the validation data using simulation was 78.6 %. Figure 4 shows the simulated output using the identified PWARX model and the measured output. This is slightly better than the result reported in [18] for a PWARX of the form (21) with two submodels. Note that this is a considerably more difficult task than the one-step ahead prediction. Here the measured inputs

and estimated outputs are used to form the regressors (which are used to choose what submodel that is active). The deviation between measured and simulated output can to some extent be explained by dry friction. Notice for example that the input is changing but not the output between time 10 and 11 in Figure 3.

Fig. 4. Validation output (head position) for Example 4. Simulated output (dashed) along with the measured system output (solid).

5 Conclusion

A novel method for switched regression has been presented. The formulation takes the shape of a least-squares problem with sum-of-norms regularization over regressor parameter differences, a generalization of ℓ_1 -regularization. The regularization constant is used to trade off fit and the number of submodels. Numerical illustrations on previously known examples from the literature shows that the proposed method performs well in comparison to known identification methods.

There are several interesting extensions of proposed scheme. For example, a piecewise nonlinear function could be estimated by applying a regularization as in (10) to *Support Vector Regression* (SVR, see e.g., [36]). See also [9].

6 Acknowledgment

Partially supported by the Swedish foundation for strategic research in the center MOVIII and by the Swedish Research Council in the Linnaeus center CADICS. Also partial support from the European Research Council under the advanced grant LEARN, contract 267381, which is gratefully acknowledged. Ohlsson is also supported by a postdoctoral grant from the Sweden-America Foundation, donated by ASEA's Fellowship Fund, and by postdoctoral grant from the Swedish Science Foundation.

References

- [1] Laurent Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668–677, April 2011.
- [2] A. Bemporad, G. Ferrari-Trecate, and M. Morari. Observability and controllability of piecewise affine and hybrid systems. *IEEE Transactions on Automatic Control*, 45(10):1864–1876, October 2000.
- [3] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50(10):1567–1580, October 2005.
- [4] A. Bemporad, J. Roll, and L. Ljung. Identification of hybrid systems via mixed-integer programming. In *Proceedings of the 40th IEEE Conference on Decision and Control*, volume 1, pages 786–792, December 2001.
- [5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [6] L. Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013, May 1993.
- [7] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, February 2006.
- [8] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.
- [9] T. Falck, H. Ohlsson, L. Ljung, J. A.K. Suykens, and B. De Moor. Segmentation of times series from nonlinear dynamical systems. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, 2011.
- [10] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.
- [11] A. Garulli, S. Paoletti, and A. Vicino. A survey on switched and piecewise affine system identification. In *Proceedings of the 16th IFAC Symposium on System Identification, SYSID 2012*, pages 344–355, July 2012.
- [12] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. D. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag, 2008. http://stanford.edu/~boyd/graph_dcp.html.
- [13] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. <http://cvxr.com/cvx>, August 2010.
- [14] W. P. M. H. Heemels, B. De Schutter, and A. Bemporad. Equivalence of hybrid dynamical models. *Automatica*, 37(7):1085–1091, July 2001.
- [15] A. L. Juloski, W. P. M. H. Heemels, and G. Ferrari-Trecate. Identification of an experimental hybrid system. In *IFAC Conference on the Analysis and Design of Hybrid Systems (ADHS 03)*, pages 39–44, St. Malo, France, June 2003.
- [16] A. L. Juloski, W. P. M. H. Heemels, and G. Ferrari-Trecate. Data-based hybrid modelling of the component placement process in pick-and-place machines1. *Control Engineering Practice*, 12(10):1241–1252, 2004.
- [17] A. L. Juloski, W. P. M. H. Heemels, G. Ferrari-Trecate, R. Vidal, S. Paoletti, and J. H. G. Niessen. Comparison of four procedures for the identification of hybrid systems. In M. Morari and L. Thiele, editors, *Hybrid Systems: Computation and Control*, volume 3414 of *Lecture Notes in Computer Science*, chapter 23, pages 354–369. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2005.
- [18] A. L. Juloski, S. Paoletti, and J. Roll. Recent techniques for the identification of piecewise affine and hybrid systems. In L. Menini, L. Zaccarian, and C. T. Abdallah, editors, *Current Trends in Nonlinear Systems and Control: In Honor of Petar Kokotovic and Turi Nicosia*. Birkhuser, 2006.
- [19] A. L. Juloski, S. Weiland, and W. P. M. H. Heemels. A Bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control*, 50(10):1520–1533, October 2005.
- [20] F. Lauer, G. Bloch, and R. Vidal. A continuous optimization framework for hybrid system identification. *Automatica*, 47(3):608–613, 2011.
- [21] J.-N. Lin and R. Unbehauen. Canonical piecewise-linear approximations. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 39(8):697–699, August 1992.
- [22] J. Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.
- [23] I. Maruta and T. Sugie. Identification of PWA models via data compression based on l1 optimization. In *Proceedings of the 50th IEEE Conference on Decision and Control*, pages 2800–2805, December 2011.
- [24] H. Nakada, K. Takaba, and T. Katayama. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*, 41(5):905–913, 2005.
- [25] H. Ohlsson. *Regularization for Sparseness and Smoothness - Applications in System Identification and Signal Processing*. Linköping Studies in Science and Technology. Dissertations. No. 1351, Linköping University, November 2010.
- [26] H. Ohlsson and L. Ljung. Piecewise affine system identification using sum-of-norms regularization. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, 2011.
- [27] H. Ohlsson and L. Ljung. Identification of switched linear regression models using sum-of-norms regularization. Technical report, Linköping University, 2012.
- [28] H. Ohlsson, L. Ljung, and S. Boyd. Segmentation of ARX-models using sum-of-norms regularization. *Automatica*, 46(6):1107–1111, 2010.
- [29] M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000.
- [30] N. Ozay, C. Lagoa, and M. Sznaiar. Robust identification of switched affine systems via moments-based convex optimization. In *Proceedings of the 48th IEEE Conference on Decision and Control*, pages 4686–4691, December 2009.
- [31] N. Ozay, M. Sznaiar, C. Lagoa, and O. Camps. A sparsification approach to set membership identification of a class of affine hybrid systems. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 123–130, December 2008.
- [32] S. Paoletti, A.L. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: a tutorial. *European Journal of Control*, 13(2-3), 2007.
- [33] J. Roll. *Local and Piecewise Affine Approaches to System Identification*. Linköping studies in science and technology. thesis no 802, Linköping University, April 2003.

- [34] J. Roll. Piecewise linear solution paths with application to direct weight optimization. *Automatica*, 44:2745–2753, 2008.
- [35] J. Roll, A. Bemporad, and L. Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, 2004.
- [36] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
- [37] R. Tibsharani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B (Methodological)*, 58(1):267–288, 1996.
- [38] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
- [39] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proceedings of the 42nd IEEE Conference on Decision and Control (CDC)*, pages 167–172, Hawaii, USA, December 2003.