A product-of-errors framework for linear hybrid system identification

Fabien Lauer, René Vidal, Gérard Bloch

A product-of-errors framework for linear hybrid system identification

Fabien Lauer * **René Vidal** ** **Gérard Bloch** *

* *Centre de Recherche en Automatique de Nancy, Nancy–University, CNRS, France, {fabien.lauer,gerard.bloch}@esstin.uhp-nancy.fr*
** *Center for Imaging Science, Johns Hopkins University, USA*

Abstract: We propose a general framework for identification of linear discrete-time hybrid systems in which arbitrary loss functions can be easily included. Our framework includes the algebraic (Vidal et al., 2003) and support vector regression (Lauer and Bloch, 2008a,b) methods as particular cases. Inspired by these approaches, we then propose an optimization framework that relies on the minimization of a *product of loss functions*. Here, the identification problem is recast as a nonlinear and non-convex, though continuous, optimization program that involves only the model parameters as variables. As a result, its complexity scales linearly with the number of data and it can easily be solved using standard global optimization methods. Moreover, we show that by choosing a saturated loss function, such as Hampel's loss function, the algorithm can efficiently deal with noise and outliers in the data. The final result is a general framework for linear hybrid system identification that can deal efficiently with noise, outliers, and large data sets. Numerical experiments demonstrate the efficiency and robustness of the proposed approach.

Keywords: switched systems; piecewise affine systems; linear hybrid systems; system identification; global optimization; large-scale problems; robustness to outliers.

1. INTRODUCTION

In this paper, we are concerned with the identification of a class of discrete-time hybrid systems of the ARX form

$$y_i = f_{\lambda_i}(\boldsymbol{x}_i) + e_i, \tag{1}$$

where $\boldsymbol{x}_i = [y_{i-1} \ \dots \ y_{i-n_a}, \ u_{i-n_k} \ \dots \ u_{i-n_k-n_c+1}]^T$ is the *continuous state* (or regression vector) of dimension p containing the lagged n_c inputs u_{i-k} and n_a outputs y_{i-k}, $\lambda_i \in \{1, \dots, n\}$ is the *discrete state* (or mode) determining which one of the n subsystems $\{f_j\}_{j=1}^n$ is active at time step i, and e_i is an additive noise term. In particular, we concentrate on the problem of finding a hybrid model $f = \{f_j\}_{j=1}^n$ of the form (1) from input–output data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. In the following, we assume the number of modes n to be known.

Related work. Recently, six main approaches to hybrid system identification have been developed: the clustering approach (Ferrari-Trecate et al., 2003; Nakada et al., 2005), the Bayesian approach (Juloski et al., 2005b), the mixed integer programming (MIP) approach (Roll et al., 2004), the bounded-error approach (Bemporad et al., 2005), the algebraic approach (Vidal et al., 2003; Ma and Vidal, 2005), and the Support Vector Regression (SVR) approach (Lauer and Bloch, 2008a; Lauer, 2008). The first four focus on the problem of PieceWise Affine (PWA) system identification, where the discrete state λ_i depends on the continuous state $\boldsymbol{x}_i$. The first two approaches rely on alternating minimization and are sensitive to initialization, while the next two involve solving a combinatorial optimization problem, which can be computationally intensive. Both the bounded-error and Bayesian approaches can also be used to identify a broader class of systems, known as switched linear systems, where the discrete state evolves independently of the continuous state. The algebraic approach (Vidal et al., 2003) gives a closed form solution to this latter problem. However, it is sensitive to noise compared to the clustering-based or bounded-error methods, as shown in Juloski et al. (2005a). The SVR approach provides a convenient way of dealing with noisy data by incorporating regularization into the optimization framework. However, it optimizes over a number of variables that grows with the number of data points, hence it is only applicable to small data sets. The bounded error approach suffers from the same drawback. To the best of our knowledge, other than the bounded error approach, none of the existing approaches deals explicitly with outliers in the data. We refer the reader to Paoletti et al. (2007) for a review and comparison of some of these methods.

Paper contributions. Our main contribution is a hybrid system identification algorithm that is computationally efficient with respect to the number of data points, and robust with respect to outliers in the data. The proposed method is based on the minimization of a *product of loss functions* plus a regularization term. This provides a general framework for hybrid system identification, in which any suitable loss function can be used. In particular, data corrupted by non-Gaussian noise can be treated by a suitable choice of the loss function. The interpretation of the method in a maximum likelihood framework also allows one to choose the loss functions with respect to the noise probability density function, as in standard estimation theory.

We also show that unbounded loss functions that are robust to outliers in classical linear estimation problems cannot guarantee robustness for the product-of-errors approach. Thus, we propose to use saturated loss functions, such as the Hampel's loss function (Cichocki and Unbehauen, 1993), which are shown to be robust in this case. Experiments show that this loss function outperforms the classical loss functions in the presence of outliers and additive noise.

The paper also proposes a reformulation of the hybrid system identification problem as an unconstrained optimization program. Though non-convex, this nonlinear program involves a low number of variables, equal to the number of parameters of the hybrid model. As we will see, its complexity scales linearly with the number of data and thus it is efficiently solvable by standard global optimization algorithms. Experiments show that the global optimum is always found in few seconds for data sets of up to hundreds of thousands of data points.

Paper organization. The paper starts by presenting a general optimization framework in §2, including its min-min and maximum likelihood interpretations (§2.1-2.2). We also show how two existing hybrid system identification algorithms can be interpreted as particular cases of this framework (§2.3-2.4). §3 presents the proposed product-of-errors approach to hybrid system identification, including an analysis of its robustness to outliers (§3.1), and the global optimization method (§3.2). The paper ends with numerical experiments in §4 and conclusions in §5.

2. GENERAL FRAMEWORK

The general principle behind all hybrid system identification methods is to find a collection of models $\{f_i\}_{i=1}^n$ that best fit the given collection of data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. As a consequence, one can pose the hybrid system identification problem as an optimization problem of the form

$$\min_{f_1,\dots,f_n} R(f_1,\dots,f_n) + J(f_1,\dots,f_n,\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N). \quad (2)$$

The first term is called the *regularizer*, and measures the model smoothness or the model complexity. The second term is called the *fitting error* and measures the fidelity of the model with respect to the data. Data fidelity is often measured using a *loss function* of the error $e_{ij} = y_i - f_j(\boldsymbol{x}_i)$ incurred by assigning the ith data point to the jth model.

The methods included in the framework differ in the choice of the regularizer, in the choice of the loss function, or in how these loss functions are combined across different models to form the fitting error. Before delving into the details, we first review some of the well known loss functions used in estimation theory and shown in Fig. 1.

Squared loss function. The squared loss function,

$$l(e) = e^2, \quad (3)$$

is perhaps the most well known. As will be seen in §2.3, the algebraic method (Vidal et al., 2003) can be interpreted in the proposed framework on the basis of this loss function.

Absolute loss function. Another well known loss function is the absolute loss, which given by

$$l(e) = |e|. \quad (4)$$

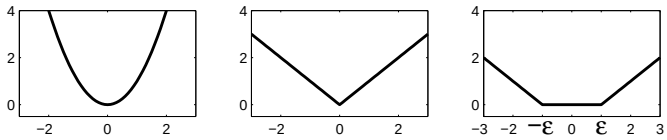

Fig. 1. Squared, absolute, and ε-*insensitive* loss functions $l(e)$ versus the error e.

In classical linear estimation theory, this loss function offers a certain robustness to outliers compared to the squared loss function. As will be seen in §3.1, this property does not hold for the proposed product-of-errors approach to hybrid system identification.

ε-insensitive loss function. The ε-insensitive loss function, defined by Vapnik (1995) for Support Vector Regression (SVR),

$$l(e) = \max(0,\ |e| - \varepsilon), \quad (5)$$

builds a tube of insensitivity of radius ε, inside which the errors are meaningless. Errors larger than ε are penalized linearly, which ensures a certain robustness to outliers in linear estimation theory. The ε-insensitive loss function has been used for hybrid system identification in Lauer and Bloch (2008a,b) as a relaxation to the bounded-error approach (Bemporad et al., 2005).

2.1 Min-min error estimator

The min-min estimator (MME) assigns sample $(\boldsymbol{x}_i, y_i)$ to the submodel f_j that gives the best estimate $f_j(\boldsymbol{x}_i)$ of y_i:

$$\hat{\lambda}_i = \arg \min_{j=1,\dots,n} l(y_i - f_j(\boldsymbol{x}_i)),\ i = 1,\dots,N, \quad (6)$$

where l is, for example, one of the loss functions above. Thus, the fitting error to be minimized is the sum of the errors made after assigning point i to submodel $\hat{\lambda}_i$, i.e.,

$$\min_{f_1,\dots,f_n} J^{MME} = \sum_{i=1}^{N} \left(\min_{j=1,\dots,n} l(y_i - f_j(\boldsymbol{x}_i)) \right). \quad (7)$$

One of the main difficulties with this MME framework is that it involves optimization over both discrete and continuous variables. To see this, notice that one can rewrite the optimization problem in (7) as

$$\min_{\{f_j\},\{\beta_{ij}\}} \sum_{i=1}^{N} \sum_{j=1}^{n} \beta_{ij} l(y_i - f_j(\boldsymbol{x}_i)), \quad (8)$$

where $\beta_{ij} \in \{0,1\}$ and $\sum_j \beta_{ij} = 1$. The discrete variables β_{ij} encode the assignment of points to submodels, while the continuous variables encode the parameters of each submodel. One way to solve this mixed (discrete and continuous) program is to use alternating minimization: given the submodels, compute the assignment of points to submodels using (6); and given the assignments, compute one submodel for each group of points. This approach is indeed effective when the submodels are linear and the squared loss (3) is used, because the estimation of each submodel is a linear system identification problem. However, this approach is sensitive to initialization. Note that problem (8) can also be solved by using mixed integer programming techniques, as proposed in Roll et al. (2004) for hinging hyperplane models. These latter optimization techniques can guarantee to find the global minimum, but, due to their high complexity, they can only be used in practice for small data sets.

2.2 Maximum likelihood of the most likely estimator

The Maximum Likelihood (ML) approach consists in finding the model f that most likely generated the data. The Maximum Likelihood of the Most Likely (MLML) estimator assigns each training sample $(\boldsymbol{x}_i, y_i)$ to the most likely submodel f_j, i.e. the one with maximal likelihood of the sample w.r.t. f_j. This leads to

$$f(\boldsymbol{x}_i) = f_{\hat{\lambda}_i}(\boldsymbol{x}_i), \text{ for } \hat{\lambda}_i = \arg\max_{j=1,\dots,n} p(y_i|\boldsymbol{x}_i, f_j). \quad (9)$$

As $p(y_i|\boldsymbol{x}_i, f) = p(y_i|\boldsymbol{x}_i, f_{\hat{\lambda}_i})$, the likelihood of a sample w.r.t. the hybrid model f can be written as

$$p(y_i|\boldsymbol{x}_i, f) \propto \max_{j=1,\dots,n} p(y_i|\boldsymbol{x}_i, f_j). \quad (10)$$

After replacing the maximization in (10) by the minimization of the negative log-likelihood, we obtain the following MLML estimator over all the samples

$$\min_{f_1,\dots,f_n} J^{MLML} = \sum_{i=1}^{N} \left(\min_{j=1,\dots,n} -\ln p(y_i|\boldsymbol{x}_i, f_j) \right). \quad (11)$$

Notice that the MLML problem amounts to a min-min optimization program, hence there is an equivalence between MLML and MME estimators. Indeed, choosing a particular loss function l for an MME estimator (7) corresponds to a particular choice of the noise probability density function, $p(y_i|\boldsymbol{x}_i, f_j)$, for an MLML estimator (11). Therefore, MLML suffers from the same drawbacks as MME. That is, the solution to problem (11) involves alternating minimization, which is sensitive to initialization, and possibly computationally costly for non-Gaussian errors.

2.3 Algebraic approach

One way of reducing the computational burden of mixed optimization is to use soft assignments $\beta_{ij} \in \mathbb{R}^+$ rather than hard assignments $\beta_{ij} \in \{0, 1\}$. The algebraic approach (Vidal et al., 2003) uses a clever choice of the assignments that leads to a closed form solution for linear models. Specifically, the algebraic approach is based on the observation that, in the absence of noise, $y_i = \boldsymbol{w}_{\lambda_i}^T \boldsymbol{x}_i$. This leads to the following hybrid decoupling constraints

$$\prod_{j=1}^{n} (y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i) = 0, \ i = 1, \dots, N, \quad (12)$$

from which one can algebraically solve for the model parameters as shown in Vidal et al. (2003); Ma and Vidal (2005). In case of noisy data, the authors solve the equations in (12) in a least squares sense by minimizing

$$\min_{\boldsymbol{w}_1,\dots,\boldsymbol{w}_n} J^A = \sum_{i=1}^{N} \prod_{j=1}^{n} (y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i)^2. \quad (13)$$

This cost function is the same as that in (8) with $l(e) = e^2$, but with soft assignments $\beta_{ij} = \prod_{k \neq j} l(y_i - \boldsymbol{w}_k^T \boldsymbol{x}_i)$.

Though in principle (13) is a nonlinear optimization problem, the authors find a linear solution by minimizing over the coefficients of the product polynomial, rather than over the model parameters $\boldsymbol{w}_j$. This approximation results in a convex (quadratic) optimization problem, which can be solved very efficiently using linear techniques. However, this approximation comes at the cost of sensitivity to noise and outliers. Whenever pertinent, we will refer to these two approaches as linear or nonlinear algebraic approach.

2.4 Support Vector Regression approach

None of the approaches described so far includes regularization over the submodel parameters. A natural way of including regularization is to use SVR with ℓ_2 regularization, which in the case of linear models, $f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T \boldsymbol{x}$, yields $\sum_j \boldsymbol{w}_j^T \boldsymbol{w}_j$. Inspired by the hybrid decoupling constraint (12), the work of Lauer and Bloch (2008a) combines this regularizer with a fitting error which is essentially the product of upper bounds on the errors $\xi_{ij} \geq |y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i| - \delta_j$, for thresholds δ_j. This leads to the following optimization program

$$\min_{\boldsymbol{w}_j, \xi_{ij} \geq 0} \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{w}_j^T \boldsymbol{w}_j + \frac{C}{N} \sum_{i=1}^{N} \prod_{j=1}^{n} \xi_{ij} \quad (14)$$

$$-\xi_{ij} - \delta_j \leq y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i \leq \delta_j + \xi_{ij}, \ i = 1, \dots, N, \ j = 1, \dots, n.$$

where C is the parameter that tunes the trade-off between the regularizer and the fitting error. A similar formulation of this method with ℓ_1-norm regularization can be found in (Lauer, 2008).

However, the problem with this formulation is that it involves a number of variables equal to $n \times (p + N)$, which grows with the number of samples N for a fixed number of modes n and a fixed number of parameters per mode p. To resolve this issue, notice that the SVR method can be reformulated in the proposed general framework by using the ε-insensitive loss function in (5). This leads to the optimization problem

$$\min_{\boldsymbol{w}_1,\dots,\boldsymbol{w}_n} \underbrace{\frac{1}{n} \sum_{j=1}^{n} \boldsymbol{w}_j^T \boldsymbol{w}_j}_{R^{SVR}} + \underbrace{\frac{C}{N} \sum_{i=1}^{N} \prod_{j=1}^{n} \max(0, |y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i| - \delta_j)}_{J^{SVR}},$$

which involves only $n \times p$ variables. This reformulation of the SVR approach is the basis for the product-of-errors approach, which we propose next.

3. PRODUCT-OF-ERRORS ESTIMATOR

In the previous section, we presented a general optimization framework for solving the hybrid system identification problem, and showed how different methods can be obtained via different choices of the regularizer and fitting error, as summarized in Table 1. Some methods, e.g., the algebraic approach, lead to simple optimization problems, but suffer from robustness issues. Other methods, e.g., the SVR approach, can be made robust to noise, but not to outliers and involve optimization over a large number of variables.

To circumvent these issues, in this section we propose a product-of-errors (PE) estimator of the form

$$\min_{f_1,\dots,f_n} \underbrace{\frac{1}{n} \sum_{j=1}^{n} R(f_j)}_{R^{PE}} + \underbrace{\frac{C}{N} \sum_{i=1}^{N} \prod_{j=1}^{n} l(y_i - f_j(\boldsymbol{x}_i))}_{J^{PE}}. \quad (15)$$

where $R(f_j)$ is a regularizer for submodel f_j and l is a robust loss function, to be defined below. The PE estimator has several important properties. First of all, notice that for linear models, $f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T \boldsymbol{x}$, the PE estimator is as efficient as the nonlinear algebraic approach, because it

Table 1. Particular choice of regularizer and loss function leading to the existing methods.

Method	Regularizer $R(f_1,\dots,f_n)$	Fitting error $J(f_1,\dots,f_n,\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N)$
SVR approach(Lauer and Bloch, 2008a)	$\frac{1}{n}\sum_{j=1}^n \|\boldsymbol{w}_j\|_2^2$	$\sum_{i=1}^N \prod_{j=1}^n \max(0, \|y_i - \boldsymbol{w}_j^T\boldsymbol{x}_i\| - \delta_j)$
Algebraic approach (Vidal et al., 2003)		$\sum_{i=1}^N \prod_{j=1}^n (y_i - \boldsymbol{w}_j^T\boldsymbol{x}_i)^2$
Product-of-errors estimator (this paper)	$\frac{1}{n}\sum_{j=1}^n R(\boldsymbol{w}_j)$, for any $R(\boldsymbol{w}_j)$	$\sum_{i=1}^N \prod_{j=1}^n l(y_i - f_j(\boldsymbol{x}_i))$, for any $l(y_i - \boldsymbol{w}_j^T\boldsymbol{x}_i)$

optimizes over a number of variables that does not depend on the number of data points. In fact, J^{PE} coincides with the algebraic error in (13) when $l(e) = e^2$. However, the PE estimator can be made robust to outliers by a suitable choice of the loss function, as we will show in §3.1. Notice also that the PE estimator coincides with the SVR approach when $R(\boldsymbol{w}_j) = \boldsymbol{w}_j^T\boldsymbol{w}_j$ and l is chosen as the ε-insensitive loss function. However, we will see in §3.1 that the product of ε-insensitive loss functions is not robust to outliers, and propose an alternative choice. Finally, although the PE estimator involves solving a nonlinear and non-convex program, we will show in §3.2 that global optimization techniques can be used to find the global minimum, under the mild requirement of continuity of the loss function.

3.1 Robustness to outliers

In this paper, we consider only outliers with respect to the output value, i.e., points with arbitrary value y_i, but exact regression vector $\boldsymbol{x}_i$. For an estimator to be robust to such outliers, the effect of a single point on the estimation must be bounded. For instance, in classical linear estimation problems, the influence function of the squared loss (3), i.e. its derivative with respect to y_i, is unbounded. On the other hand, the absolute loss (4) has an influence function bounded by 1 and thus leads to robust estimators. However, a simple calculation shows that this property does not hold when using a PE estimator for hybrid systems. In this case, the influence of a single point on the objective function (15) becomes

$$\frac{\partial J^{PE}}{\partial y_i} = \sum_{j=1}^n \frac{\partial l(y_i - f_j(\boldsymbol{x}_i))}{\partial y_i} \prod_{k\in\{1,\dots,n\}\setminus j} l(y_i - f_k(\boldsymbol{x}_i)), \quad (16)$$

where $\partial l(y_i - f_j(\boldsymbol{x}_i))/\partial y_i = \partial l(y_i - f_j(\boldsymbol{x}_i))/\partial(y_i - f_j(\boldsymbol{x}_i)) \times \partial(y_i - f_j(\boldsymbol{x}_i))/\partial y_i = l'(e_{ij})$. Though the derivative $l'(e)$ of the absolute loss is bounded by 1, the values $l(y_i - f_k(\boldsymbol{x}_i))$ cannot be bounded for an unbounded y_i. Thus the influence of a single point cannot be bounded.

In order to obtain a robust PE estimator, one needs a loss function l leading to $l'(e) = 0$, for large values of the error $|e|$. For this purpose, we propose to use a *saturated absolute loss* function defined as

$$l(e) = \min(\delta,\ |e|), \quad (17)$$

with derivative $l'(e) = -1$, if $e \in [-\delta, 0[$, 1, if $e \in [0, \delta[$, and 0 otherwise.

A similar methodology can be applied to define other robust loss functions, such as the saturated squared loss $l(e) = \min(\delta^2, e^2)$, also known as Talvar's loss function, or the smoother Hampel's loss function

$$l(e) = \begin{cases} \delta^2/\pi\,(1 - \cos(\pi e/\delta)), & \text{if } |e| \le \delta, \\ 2\delta^2/\pi, & \text{otherwise.} \end{cases} \quad (18)$$

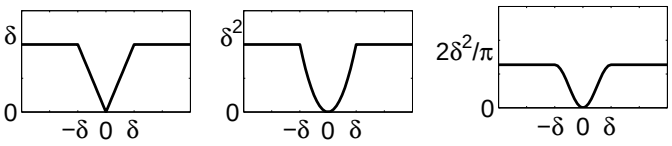

Fig. 2. Saturated absolute, saturated squared (or Talvar's), and Hampel's loss functions $l(e)$ versus the error e.

These loss functions are plotted in Fig. 2.

3.2 Optimization of the product-of-errors function

For linear submodels of the form $f_j(\boldsymbol{x}) = \boldsymbol{w}_j^T\boldsymbol{x}$, $j = 1,\dots,n,$, the product-of-error estimator is given by

$$\min_{\boldsymbol{w}_1,\dots,\boldsymbol{w}_n} \frac{1}{n}\sum_{j=1}^n R(\boldsymbol{w}_j) + \frac{C}{N}\sum_{i=1}^N \prod_{j=1}^n l(y_i - \boldsymbol{w}_j^T\boldsymbol{x}_i), \quad (19)$$

where the parameter vectors $\boldsymbol{w}_j$ to estimate are of dimension $p = n_a + n_c$ (or $p = n_a + n_c + 1$ for affine submodels). Thus the number of variables involved in (19), $n \times p$, is small and fixed for any number of data N. However, the objective function in (19) requires to compute a sum over N terms, hence the linear complexity of the algorithm w.r.t. N. This allows the optimizer to find the global minimum in reasonable time, despite the NP-hard nature of the problem. Here we propose to solve (19) with the Multilevel Coordinate Search (MCS) algorithm [1] (Huyer and Neumaier, 1999), that is guaranteed to converge if the objective is continuous in the neighborhood of the global minimizer. This optimizer uses only function values (when required, derivatives are estimated from these) and alternates between global and local search. The local search, done via sequential quadratic programming (SQP), speeds up the convergence once the global part has found a point in the bassin of attraction of the global minimizer.

4. NUMERICAL EXPERIMENTS

In this section, we present some numerical results supporting the approach. In particular, we consider the identification of a switched linear system from data with noise and outliers in §4.1, and large-scale experiments in §4.2.

All the experiments are performed on a 2.4GHz Core 2 duo laptop with 2GB of memory using Matlab. Though the MCS algorithm can deal with unbounded variables, box constraints are used to limit the search space and restrain the variables to the interval $[-10, 10]$ (in practice, these bounds can be larger with little influence). Beside this, the default parameters of MCS are used. For all the problems, N samples are generated by

$$y_i = \boldsymbol{\theta}_{\lambda_i}^T\boldsymbol{x}_i + v_i, \quad i = 1,\dots,N, \quad (20)$$

[1] Available at http://www.mat.univie.ac.at/~neum/software/mcs/.

where the $\boldsymbol{\theta}_j \in \mathbb{R}^p$ are the true parameters to recover and $v_i \sim \mathcal{N}(0, \sigma_v^2)$ is a Gaussian noise. The methods are compared on the basis of the normalized Mean Squared Error (MSE) on the parameters, NMSE $= \sum_{j=1}^{n} \|\boldsymbol{\theta}_j - \boldsymbol{w}_j\|_2^2 / \|\boldsymbol{\theta}_j\|_2^2$, where the $\boldsymbol{w}_j$ are the estimated parameters.

4.1 Switched linear system identification

Consider the example taken from Vidal (2008). The aim is to recover, from $N = 1000$ samples, the parameters $\boldsymbol{\theta}_1 = [0.9, 1]^T$ and $\boldsymbol{\theta}_2 = [1, -1]^T$ of a dynamical system, arbitrarily switching between $n = 2$ modes, with continuous state $\boldsymbol{x}_i = [y_{i-1}, u_{i-1}]^T$.

Linear vs. nonlinear algebraic approach. As seen in §2.3, the algebraic method can be implemented either as the direct minimization over the model parameters, as in (13), or as a linear problem solved w.r.t. the product of parameters. These two algorithms are compared in Table 2, which shows the average and standard deviation of the NMSE over 100 trials with different noise sequences. These results highlight the gain in solving the problem directly for the model parameters $\boldsymbol{w}_j$ versus the optimization over the product of parameters. The larger error obtained with the linear method for large noise levels is due to the fact that this method neglects one constraint: the polynomial must be factorable into the original form (12). However, the gain in NMSE obtained by solving directly for the $\boldsymbol{w}_j$ comes at the cost of solving a nonlinear optimization program instead of a linear problem. This leads to computing times about 30 times larger.

Table 2. Average NMSE and computing time over 100 trials.

	σ_v	Nonlinear algebraic	Linear algebraic
NMSE	0.00	0.00000	0.00000
	0.02	0.00000 ± 0.00000	0.00003 ± 0.00012
	0.10	0.00007 ± 0.00007	0.00447 ± 0.02962
	0.20	0.00036 ± 0.00028	0.03968 ± 0.21762
	0.30	0.00069 ± 0.00055	0.04449 ± 0.19270
Time (sec.)		0.32 ± 0.02	0.01 ± 0.00

Robustness to outliers. We now study the robustness to outliers of the PE estimator (19) without regularization. The data are corrupted with 20% of outliers by forcing the target outputs y_i, at random time steps i, to take uniformly distributed random values in the interval $[-10, 10]$. Table 3 shows the resulting NMSE for the squared, absolute, saturated absolute and Hampel's loss functions with parameter δ set to $\delta = 2$.

These results emphasize that the squared loss function cannot be used to accurately estimate the parameters in the presence of outliers. On the other hand, the absolute loss function provides a certain level of robustness even for a PE estimator and still leads to accurate estimates. However, the loss functions satisfying the robustness condition for PE estimators, i.e. $l'(e) = 0$ for $|e| > \delta$, lead to a lower NMSE. In these experiments, the minimum NMSE is always obtained with Hampel's loss function. The benefits of using saturated loss functions compared to the absolute loss are also emphasized by the plots in Fig. 3, which show the NMSE for an increasing percentage of outliers and additive Gaussian noise ($\sigma_v = 0.3$).

Table 3. Average and standard deviation of the NMSE for a data set with 20% of outliers.

Loss function	σ_v	NMSE	Time (sec.)
Squared	0.0	0.22898 ± 0.10684	0.36 ± 0.06
Absolute	0.0	0.00708 ± 0.01618	0.54 ± 0.02
Saturated absolute	0.0	0.00165 ± 0.00442	0.65 ± 0.04
Hampel's	0.0	$\mathbf{0.00009 \pm 0.00007}$	0.76 ± 0.21
Squared	0.1	0.21267 ± 0.10700	0.37 ± 0.05
Absolute	0.1	0.00276 ± 0.00801	0.61 ± 0.03
Saturated absolute	0.1	0.00152 ± 0.00418	0.65 ± 0.03
Hampel's	0.1	$\mathbf{0.00013 \pm 0.00010}$	0.68 ± 0.10
Squared	0.2	0.18548 ± 0.10161	0.41 ± 0.07
Absolute	0.2	0.00099 ± 0.00128	0.63 ± 0.03
Saturated absolute	0.2	0.00033 ± 0.00035	0.69 ± 0.05
Hampel's	0.2	$\mathbf{0.00024 \pm 0.00021}$	0.71 ± 0.13
Squared	0.3	0.22177 ± 0.12003	0.40 ± 0.06
Absolute	0.3	0.00170 ± 0.00141	0.64 ± 0.04
Saturated absolute	0.3	0.00075 ± 0.00059	0.69 ± 0.04
Hampel's	0.3	$\mathbf{0.00050 \pm 0.00042}$	0.74 ± 0.12

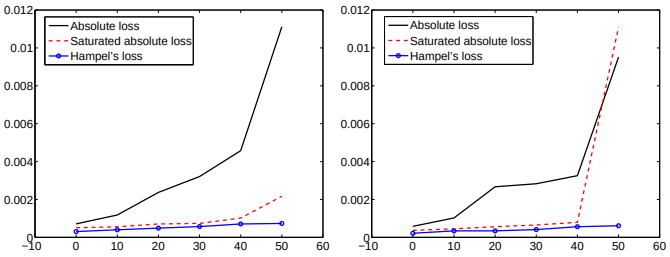

Fig. 3. Average (left) and standard deviation (right) of the NMSE over 100 trials vs. the percentage of outliers.

4.2 Large-scale experiments

Large data sets. The method is now evaluated over 100 large-scale problems, generated by (20) with uniformly distributed random parameters $\boldsymbol{\theta}_j \in [-2, 2]^p$, $j = 1, 2$, and regression vectors $\boldsymbol{x}_i \in [-5, 5]^p$, a random switching sequence $\{\lambda_i\}$ and $\sigma_v = 0.2$. The PE estimator (19) with squared loss is considered without regularization. For $n = 2$ and $p = 3$, Table 4 shows the resulting NMSE and the computing times, averaged over the 100 problems. Note that the algorithm, including the MCS optimization, is entirely implemented in non-compiled Matlab code.

Table 4. Average error and computing time over 100 randomly generated problems.

Number of data N	NMSE ($\times 10^{-3}$)	Time (seconds)
10 000	0.0046 ± 0.0213	1.27 ± 0.17
50 000	0.0010 ± 0.0012	4.77 ± 0.56
100 000	0.0008 ± 0.0010	11.62 ± 1.92
500 000	0.0007 ± 0.0010	63.70 ± 8.23

The times in Table 4 show that the proposed algorithm has a complexity scaling linearly with the number of data N. As a result, the method can be applied to large data sets with hundreds of thousands of data.

Large model structures. The computing time of the method heavily relies on the number of model parameters $n \times p$. Thus the method may not be suitable for hybrid models with numerous modes. However, as shown by Fig. 4, the average computing time remains below 4 minutes for models with up to 40 parameters. The average is evaluated over 10 runs of the method using Hampel's loss

function for $N = 10\,000$ data points, generated by random sets of parameters.

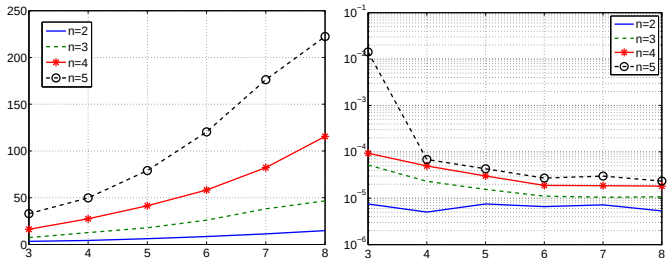

Fig. 4. Average computing time in seconds (left) and NMSE (right) vs. the number of parameters per submodel for n modes.

4.3 Application to real data

The method is now applied to real data from a pick-and-place machine, described in the comparison paper by Juloski et al. (2005a). In this paper, most methods (except the algebraic approach) had to subsample the data set of 60 000 samples to 750 samples in order to be applied in reasonable time. This subsampled data set is then divided in two overlapping subsets of 500 samples. The PE estimator (19) with squared loss and without regularization is applied to the first subset (with normalized output and a 1 appended to the regression vector) to build a hybrid model with 2 modes and orders $n_a = n_c = 2$. This model, obtained in 2 seconds, is then tested on the second subset, leading to a one-step-ahead prediction MSE $= \sum_{i=1}^{N}(y_i - f_{\lambda_i}(\boldsymbol{x}_i))^2 = 0.0590$. When considering the entire data set, divided in two subsets of 30 000 samples, the model, obtained in 18 seconds, leads to MSE $= 5.6248 \times 10^{-6}$. The results reported in Ma and Vidal (2005) on this application are 0.1195 and 5.3426×10^{-6} for the two settings, respectively. In comparison, the PE estimator leads to a lower MSE for the subsampled data set. Figure 5 shows the simulation of the model on the entire data set, where the data y_i and the model output $f(\boldsymbol{x}_i)$ cannot be distinguished.

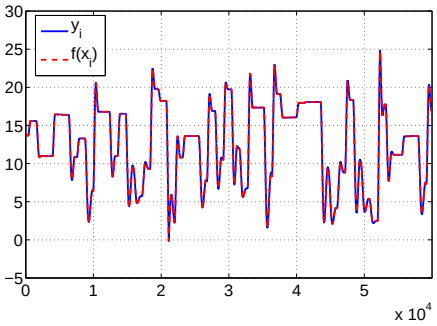

Fig. 5. Model simulation on the real data.

5. CONCLUSION

We have proposed a new framework for hybrid system identification, which includes some of the most recent approaches as special cases. We also proposed a specific optimization framework, which relies on the minimization of a *product of loss functions* to circumvent some computational issues. The loss function can be chosen on the basis of the noise model thanks to a maximum likelihood interpretation. In addition, we gave a condition for a loss function to be robust to outliers in the product-of-errors context. Numerical experiments showed that the resulting algorithm can identify linear hybrid systems from both large data sets and data sets corrupted with outliers.

Future work will focus on the estimation of the number of modes and of the subsystem orders in the presence of outliers, as well as on the extension of the proposed framework to nonlinear hybrid system identification, e.g. with nonlinear submodels in kernel form as studied by Lauer and Bloch (2008b). In addition, robustness to outliers in the regressors also requires further investigation.

REFERENCES

A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Trans. on Automatic Control*, 50(10):1567–1580, 2005.

A. Cichocki and R. Unbehauen. *Neural Networks for Optimization and Signal Processing*. Wiley, 1993.

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.

W. Huyer and A. Neumaier. Global optimization by multilevel coordinate search. *Journal of Global Optimization*, 14(4):331–355, 1999.

A. L. Juloski, W. Heemels, G. Ferrari-Trecate, R. Vidal, S. Paoletti, and J. H. G. Niessen. Comparison of four procedures for the identification of hybrid systems. In *Proc. of the 8th Int. Conf. on Hybrid Systems: Computation and Control (HSCC), Zurich, Switzerland*, volume 3414 of *LNCS*, pages 354–369, 2005a.

A. L. Juloski, S. Weiland, and W. Heemels. A Bayesian approach to identification of hybrid systems. *IEEE Trans. on Automatic Control*, 50(10):1520–1533, 2005b.

F. Lauer. *From Support Vector Machines to Hybrid System Identification*. PhD thesis, Université Henri Poincaré Nancy 1, 2008.

F. Lauer and G. Bloch. A new hybrid system identification algorithm with automatic tuning. In *Proc. of the 17th IFAC World Congress, Seoul, Korea*, pages 10207–10212, 2008a.

F. Lauer and G. Bloch. Switched and piecewise nonlinear hybrid system identification. In *Proc. of the 11th Int. Conf. on Hybrid Systems: Computation and Control (HSCC), St. Louis, MO, USA*, volume 4981 of *LNCS*, pages 330–343, 2008b.

Y. Ma and R. Vidal. Identification of deterministic switched ARX systems via identification of algebraic varieties. In *Proc. of the 8th Int. Conf. on Hybrid Systems: Computation and Control (HSCC), Zurich, Switzerland*, volume 3414 of *LNCS*, pages 449–465, 2005.

H. Nakada, K. Takaba, and T. Katayama. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*, 41(5):905–913, 2005.

S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: a tutorial. *European Journal of Control*, 13(2-3):242–262, 2007.

J. Roll, A. Bemporad, and L. Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, 2004.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA, 1995.

R. Vidal. Recursive identification of switched ARX systems. *Automatica*, 44(9):2274–2287, 2008.

R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proc. of the 42nd IEEE Conf. on Decision and Control (CDC), Maui, Hawaï, USA*, pages 167–172, 2003.