A Bias-Correction Approach for the Identification of Piecewise Affine Output-Error Models

Manas Mejari*  Valentina Breschi**  Vihangkumar V. Naik***  Dario Piga*

* IDSIA Dalle Molle Institute for Artificial Intelligence, SUPSI-USI, Via Cantonale 2C, 6928 Manno, Switzerland. (e-mail: {manas.mejari, dario.piga}@supsi.ch).
** Dipartimento di Elettronica e Informazione, Politecnico di Milano, 20133 Milan, Italy (e-mail: valentina.breschi@polimi.it).
*** ODYS S.r.l., Via A. Passaglia 185, Lucca, 55100, Italy. (e-mail: vihang.naik@odys.it)

Abstract: The paper presents an algorithm for the identification of PieceWise Affine Output-Error (PWA-OE) models, which involves the estimation of the parameters defining affine submodels as well as a partition of the regressor space. For the estimation of affine submodel parameters, a bias-correction scheme is presented to correct the bias in the least squares estimates which is caused by the output-error noise structure. The obtained bias-corrected estimates are proven to be consistent under suitable assumptions. The bias-correction method is then combined with a recursive estimation algorithm for clustering the regressors. These clusters are used to compute a partition of the regressor space by employing linear multi-category discrimination. The effectiveness of the proposed methodology is demonstrated via a simulation case study.

Keywords: Hybrid systems, PWA regression, bias corrected least-squares, output-error models

1. INTRODUCTION

PieceWise Affine (PWA) models are defined by a set of affine functions, each associated to a polyhedral region of the input (or regressor) space. PWA models can be used to describe the behavior of hybrid dynamical systems, which are governed by both continuous and discrete states. One of the main advantages of the PWA modeling paradigm is that tools developed for analysis and control of hybrid systems can also be applied to systems represented in a PWA form (Bemporad and Morari, 1999; Bemporad et al., 2000; Heemels et al., 2001). Furthermore, due to the universal approximation property of PWA maps, they are able to approximate any nonlinear function with arbitrary accuracy (Breiman, 1993). This makes PWA models also suitable to describe nonlinear systems that do not necessarily exhibit a switching behavior.

The problem of identifying PWA models consists of estimating the parameters defining the affine sub-models as well as the discrete mode sequence representing the active submodel at each time step. This is an NP-hard problem (Lauer, 2015), and many heuristics have been developed for learning PWA models from data (see the survey papers (Garulli et al., 2012; Paoletti et al., 2007)). The main methods proposed in the literature include, among others, bounded-error or set-membership approaches (Bemporad et al., 2005; Ozay et al., 2015); sparse optimization based algorithms (Bako, 2011; Ohlsson and Ljung, 2013; Piga and Tóth, 2013); mixed-integer programming methods (Roll et al., 2004; Naik et al., 2017); clustering-based two-stage approaches (Nakada et al., 2005; Juloski et al., 2005; Ferrari-Trecate et al., 2003; Bako et al., 2011; Bemporad et al., 2018; Breschi et al., 2016).

The aforementioned approaches are developed for the identification of either Switched AutoRegressive with eXogenous inputs (SARX) or PieceWise Affine AutoRegressive with eXogenous inputs (PWARX) models. The underlying assumption in these methods is that the output is generated by ARX submodels. With this assumption, standard linear Least Squares (LS) can be used to obtain consistent estimates of the submodel parameters, given the discrete mode sequence. In this paper, we relax this assumption and develop an identification algorithm for PWA output-error (PWA-OE) model structure. The PWA-OE models makes the identification problem more challenging, as the affine functions of the PWA map depend on the (unmeasured) “noise-free” regressors.

Very few works in the literature have addressed the identification of PWA-OE models. In (Canty et al., 2012), PWA-OE algorithm is proposed with an Instrumental Variable (IV) scheme. The correct choice of IVs is critical as they need to be uncorrelated with the system noise and correlated with the inputs. A Prediction Error Method (PEM) is proposed in (Rosenqvist and Karlström, 2005) for identifying piecewise linear models under the assumption that the

* This work was partially supported by the European H2020-CS2 project ADMITTED, Grant agreement no. GA832003.
discrete mode sequence is known. In (Breschi et al., 2019), PEM is used to estimate more general Box-Jenkins models, along with the estimation of the hidden mode sequence. The PEM approach leads to an non-convex optimization problem (even in the linear case) which thus requires an accurate initial guess. Although the global minimum of the optimization problem is a consistent estimate of the system parameters, iterative gradient-based numerical algorithms may get trap in a local minima, giving inaccurate estimates. In (Juloski and Weiland, 2006), PWA-OE identification problem is treated in a Bayesian framework, and a suboptimal algorithm is derived for estimating the model parameters along with the unknown mode sequence.

In this work, we propose a bias-correction approach, combined with a two-stage clustering based algorithm for the identification of PWA-OE models. Specifically, the bias-corrected least squares estimates of the affine submodel parameters are derived and are proven to be consistent under the assumption that the active mode sequence is known. This assumption is then relaxed and mode sequence is estimated using an iterative clustering-based algorithm.

The main idea of bias-correction methods is to remove the bias from the LS estimates in order to obtain a consistent estimate of the model parameters. Bias-correction methods have been proposed in the literature for the identification of LTI systems (Hong et al., 2007; Zheng, 2002; Zheng and Feng, 1997; Gilson and Van den Hof, 2001), for the identification LPV models with noisy scheduling variable observations (Piga et al., 2015; Mejari et al., 2018) as well as for the identification of nonlinear system (Piga and Tóth, 2014). To the best of our knowledge, this paper presents the first contribution adapting the bias-correction method in a hybrid modelling framework.

Furthermore, for estimating the unknown active mode sequence, we adapt the two-stage clustering based algorithm recently proposed by some of the authors in (Breschi et al., 2016b). Specifically, in the first stage, training samples are clustered and bias-corrected least squares estimates of the affine submodel parameters are computed. The clusters are formed by assigning each regressor to a specific submodel according to a chosen clustering criterion. At the second stage, linear multi-category classification techniques are used to separate the clusters, thus partitioning the regressor space into polyhedral regions.

The paper is organized as follows: The identification problem of PWA-OE models is formalized in Section 2. Bias-corrected least-squares estimates for the affine submodels parameters are derived in Section 3. Section 4 presents a two-stage clustering-based iterative algorithm for simultaneous estimation of the submodel parameters and the unknown mode sequence in the first stage and for computing the partition of the regressor space in the second stage. A numerical case study is reported in Section 5 to show the effectiveness of the presented algorithm.

2. PROBLEM FORMULATION

2.1 Data generating system

We consider the following discrete-time, single-input single-output nonlinear data-generating system \( \mathcal{S}_o \), with Output-Error (OE) structure,
\[
\begin{align}
y_o(k) &= f(x_o(k)), \\
y(k) &= y_o(k) + e_o(k),
\end{align}
\]
where \( y_o(k) \in \mathbb{R} \) and \( y(k) \in \mathbb{R} \) are the noise-free and noise-corrupted output of the system at time \( k \), respectively. The noise-free regressor, which depends on the past \( n_a \) values of the output \( y_o \) and the past \( n_b \) values of the input \( u \), namely,
\[
x_o(k) = [y_o(k-1) \cdots y_o(k-n_a) u(k-1) \cdots u(k-n_b)]^\top,
\]
and \( e_o(k) \in \mathbb{R} \) is a zero-mean additive white Gaussian noise with variance \( \sigma_e^2 \), which is assumed to be statistically independent of the regressor \( x_o(k) \).

The map \( f : X \to \mathbb{R} \) is PieceWise Affine (PWA) and it is defined as
\[
f(x_o) = \begin{cases} 
(\theta_1^\top \begin{bmatrix} x_o \\ 1 \end{bmatrix}) & \text{if } x_o \in X_1, \\
\vdots & \\
(\theta_s^\top \begin{bmatrix} x_o \\ 1 \end{bmatrix}) & \text{if } x_o \in X_s,
\end{cases}
\]
where \( s \in \mathbb{N} \) is the number of modes (i.e., the number of affine functions defining \( f \)), \( \theta_i \in \mathbb{R}^{(n_a+1) \times 1} \) is the parameter vector associated to the \( i \)-th affine submodel, and the set \( X_i \subseteq X \) is a polyhedron defined as:
\[
X_i = \{x_o \in \mathbb{R}^{n_a} : H_i x_o \leq D_i\},
\]
with \( H_i \) and \( D_i \) being real matrices, for \( i = 1, \ldots, s \). The polyhedra \( \{X_i\}_{i=1}^s \) form a complete polyhedral partition of the regressor space \( X \). Note that, unlike PWARX models, the polyhedral partition \( \{X_i\}_{i=1}^s \) is defined over the space of noise-free regressors.

Since the noise \( e_o \) does not affect the evolution of the output \( y_o \) and the function \( f \) is piecewise affine, we will refer to the hybrid system \( \mathcal{S}_o \) described in (1) as a PieceWise Affine-Output Error (PWA-OE) representation. Note that, the noise-free output \( y_o \) and consequently the noise-free regressor \( x_o \) are not available for measurements.

2.2 Model structure

In order to describe the true PWA-OE system \( \mathcal{S}_o \) in (1), the following parameterized model structure \( \mathcal{M}_\theta \) is introduced,
\[
y(k) = \begin{cases} 
\theta_1^\top \begin{bmatrix} x(k) \\ 1 \end{bmatrix} + \epsilon(k), & \text{if } x(k) \in X_1, \\
\vdots & \\
\theta_s^\top \begin{bmatrix} x(k) \\ 1 \end{bmatrix} + \epsilon(k), & \text{if } x(k) \in X_s,
\end{cases}
\]
where \( \epsilon(k) \) is the residual term modeling the mismatch between the true system and model output, \( x(k) \) is the regressor vector constructed from the past \( n_a \) samples of the measured output \( y \) and past \( n_b \) samples of the input \( u \), i.e.,
\[
x(k) = [y(k-1) \cdots y(k-n_a) u(k-1) \cdots u(k-n_b)]^\top.
\]

\footnote{The collection \( \{X_i\}_{i=1}^s \) is a complete partition of \( X \) if \( \bigcup_{i=1}^s X_i = X \)
and \( \bar{X}_i \cap \bar{X}_j = \emptyset, \forall i \neq j \), with \( \bar{X}_i \) denoting the interior of \( X_i \).}
For an output-error model structure \( \mathcal{M} \), the residual term \( \epsilon \) is colored. We remark that, the problem of model structure selection (i.e., estimation of the model orders \( n_a, n_b \) and the number of modes \( s \)) is not addressed in this paper. Thus, we assume that both the true system \( S_o \) and the model \( M_o \) share the same parameters \( n_a, n_b \) and \( s \), which are assumed to be known.

The identification problem addressed in this paper is formalized as follows.

**Problem 1.** For a fixed model structure \( M_o \) and given a set of \( N \) input-output observations \( \{(u(k), y(k))\}_{k=1}^N \), generated by the system \( S_o \) in (1), compute consistent estimates of the true parameters \( \{\theta_i^o\}_{i=1}^s \) characterizing the affine submodels of the PWA map \( f \), and find a polyhedral partition \( \{X_i\}_{i=1}^s \) of the regressor space \( \mathcal{X} \), over which the local affine submodels are defined.

To this aim, a novel identification algorithm based on bias-corrected least squares is presented in the next sections. The algorithm iteratively computes the estimates of the submodel parameters and the sequence of active modes. Then, multi-category linear classification techniques are employed to compute a polyhedral partition of the regressor space, based on the estimated mode sequence.

### 3. Bias-Corrected Least Squares

It is known that for an output-error model structure, ordinary least squares give an asymptotically biased estimate of the model parameters (Ljung, 1999). To overcome this issue, in this section we quantify the bias in the LS estimates and show how to eliminate it, in order to compute a bias-corrected estimate of the model parameters \( \{\theta_i^o\}_{i=1}^s \). Since the main focus of this section is the estimation of the submodel parameters, the expression of the bias-corrected least squares estimates is derived under the assumption that the true polyhedral partition of the regressor space \( \mathcal{X} \) is known. Under this assumption, we prove that the bias-corrected least squares estimates are consistent.

Let us define the mode function \( \sigma: \mathbb{N} \to \mathbb{N} \), such that,
\[
\sigma(k) = i \iff x_o(k) \in X_i,
\]
i.e., the active mode \( \sigma(k) \in \{1, \cdots, s\} \), at time \( k \), represents the partition which the regressor \( x_o(k) \) belongs to. In this section, we assume that the sequence of active modes \( \sigma(k) \) is known. This assumption is relaxed later.

Let \( N_i \) be the number of regressor/output data points associated to the \( i \)-th affine submodel, with \( \sum_i N_i = N \). Given the mode sequence \( \{\sigma(k)\}_{k=1}^N \), we define the column vector \( \gamma_i \in \mathbb{R}^{N_i} \) as the output vector associated to the \( i \)-th affine submodel, which is constructed from the output sequence \( \{y(k)\}_{k=1}^N \) such that,
\[
y(k) \text{ is a row of } \gamma_i \iff \sigma(k) = i,
\]
and, analogously, let \( X_i \in \mathbb{R}^{N_i \times (n_a+n_b+1)} \) be the regressor matrix constructed from the sequence \( \{x(k)\}_{k=1}^N \) by stacking the extended regressors associated to the \( i \)-th affine submodel, i.e.,
\[
\begin{bmatrix}
x(k) \\
1
\end{bmatrix}^\top \text{ is a row of } X_i \iff \sigma(k) = i.
\]

Using the notation introduced above, we define the noise-free output vector \( \gamma^o_i \in \mathbb{R}^{N_i} \), the noise-free regressor matrix \( X^o_i \in \mathbb{R}^{N_i \times (n_a+n_b+1)} \) and the measurement noise vector \( e^o_i \in \mathbb{R}^{N_i} \), by stacking the noise-free outputs, noise-free extended regressors and the samples of the measurement noise \( e_o \) associated to the \( i \)-th mode, respectively. Note that, \( \gamma^o_i, X^o_i \) and \( e^o_i \) cannot be constructed in practice, since the sequence of noise-free output/regressor pairs \( \{y^o_i(k), x^o_i(k)\}_{k=1}^N \) and the sequence of measurement noise \( \{e_o(k)\}_{k=1}^N \) are not available.

#### 3.1 Computation of the bias in the least squares estimate

Consider the least squares estimate of the \( i \)-th affine submodel parameter, which is given by,
\[
\theta_i^L S = \left( \frac{X_i^\top \gamma_i}{N_i} \right)^{-1} \frac{X_i^\top \gamma_i}{N_i},
\]
where \( \Gamma_{N_i} \) is assumed to be invertible. In order to quantify the bias in the least squares estimate, we compute the difference between the LS estimate \( \theta_i^L S \) and the true model parameter \( \theta_i^o \). To this aim, let us rewrite the output vector \( \gamma_i \) using (1a), (1b) and (1d) as follows:
\[
\gamma_i = \gamma^o_i + e^o_i = x_o \theta_i^o + e^o_i = x_o \theta_i^o + \Delta X_i \theta_i^o + e^o_i.
\]

where \( \Delta X_i \) is defined as the difference between noise-free and noisy regressor matrices, namely,
\[
\Delta X_i = X^o_i - X_i.
\]

By substituting the output equation (6) in (5), the difference between the LS estimate \( \theta_i^L S \) and the true parameter \( \theta_i^o \) is expressed as follows:
\[
\theta_i^L S - \theta_i^o = \left( \frac{X_i^\top \gamma_i}{N_i} \right)^{-1} \frac{X_i^\top \gamma_i}{N_i} \Delta X_i \theta_i^o + \left( \frac{X_i^\top \gamma_i}{N_i} \right)^{-1} \frac{X_i^\top e^o_i}{N_i}.
\]

Note that, as the measurement noise \( e_o \) is assumed to be zero-mean white noise, the noise vector \( e^o_i \) is uncorrelated with the regressors \( X_i \). Thus, the term \( B \) in (8) asymptotically (as \( N_i \to \infty \)) converges to 0 with probability 1 (w.p. 1), i.e.,
\[
\lim_{N_i \to \infty} \left( \frac{X_i^\top \gamma_i}{N_i} \right)^{-1} \frac{X_i^\top e^o_i}{N_i} = 0 \quad \text{w.p. 1}.
\]

However, the term \( B \) is not guaranteed to converge to 0 in general, inducing a non-zero bias in the LS estimate. This proves that the LS estimate \( \theta_i^L S \) in (5) is not consistent, i.e.,
\[
\lim_{N_i \to \infty} \theta_i^L S \neq \theta_i^o,
\]
even for the known true mode sequence.

#### 3.2 Elimination of the bias in the least squares estimate

Since the bias term \( B \) depends on the true model parameter \( \theta_i^o \) as well as on the noise-free regressor matrix \( X^o_i \) (see Definition of \( \Delta X_i \) in (7)), it cannot be computed based on the observed input/output data. Therefore, the bias cannot be simply eliminated from the LS estimate \( \theta_i^L S \).
The first difficulty due to the dependence of the bias $B_\Delta(\theta_i^*, X_i, \Delta X_i)$ on $\theta_i^*$ can be overcome, by introducing a corrected LS estimate $\theta_i^{\text{CLS}}$, following the same rationale used in (Piga and Tóth, 2014, Section 4.2). In particular, we define the corrected LS estimate $\theta_i^{\text{CLS}}$ as follows:

$$\theta_i^{\text{CLS}} = \theta_i^{\text{LS}} - B_\Delta(\theta_i^{\text{CLS}}, X_i, \Delta X_i), \quad (9)$$

where,

$$B_\Delta(\theta_i^{\text{CLS}}, X_i, \Delta X_i) = \left( X_i^\top X_i / N_i \right)^{-1} X_i^\top \Delta X_i / N_i \psi_i^{\text{CLS}} \quad (10)$$

The main idea behind the definition of $\theta_i^{\text{CLS}}$ in (9) is that, the LS estimate is corrected by eliminating the bias term $B_\Delta$, which is evaluated at the estimate $\theta_i^{\text{CLS}}$ rather than at the unknown true model parameter $\theta_i^*$.

By substituting equations (5) and (10) in (9) and with simple algebraic manipulations, we obtain

$$\theta_i^{\text{CLS}} = \left( X_i^\top X_i + X_i^\top \Delta X_i / N_i \right)^{-1} X_i^\top \psi_i / N_i \quad (11)$$

**Proposition 1.** Under the assumption that the following limit exist:

$$\lim_{N_i \to \infty} \left( X_i^\top X_i + X_i^\top \Delta X_i / N_i \right)^{-1},$$

then, the corrected LS estimate $\theta_i^{\text{CLS}}$ in (11) is a consistent estimate of the true model parameter $\theta_i^*$, i.e.,

$$\lim_{N_i \to \infty} \theta_i^{\text{CLS}} = \theta_i^*, \text{ w.p. 1}$$

**Proof.** See Appendix A.1.

Note that, since $\Delta X_i$ depends on the noise-free regressors $X_{i,o}$ (see definition in (7)), the estimates $\theta_i^{\text{CLS}}$ in (11) cannot be computed based on the available input/output measurements. To overcome this problem, the term $X_i^\top \Delta X_i$ in (11) is replaced by a bias-eliminating matrix $\Psi_i$, which is constructed in such a way that it can be computed from the available information and it satisfies the following property:

$$C1: \lim_{N_i \to \infty} \frac{1}{N_i} X_i^\top \Delta X_i = \lim_{N_i \to \infty} \frac{1}{N_i} \Psi_i, \text{ w.p. 1}$$

A bias-eliminating matrix $\Psi_i$ which satisfies condition C1, can be constructed (following similar ideas used in (Mejari et al., 2018, Section 4.2)) by evaluating the expected value of the matrix $E \{ X_i^\top \Delta X_i \}$, such that $E \{ \Psi_i \} = E \{ X_i^\top \Delta X_i \}$ is satisfied by construction.

**Proposition 2.** The bias-eliminating matrix $\Psi_i$ is given by,

$$\Psi_i = -\sigma_o^2 N_i \begin{bmatrix} I_{n_u \times n_u} & 0_{n_u \times (n_u+1)} \\ 0_{(n_u+1) \times n_u} & 0_{(n_u+1) \times (n_u+1)} \end{bmatrix} \quad (12)$$

where $\sigma_o^2$ is the variance of the measurement noise $e_o$ and $N_i$ is the number of training data points associated with the $i$-th mode. The matrix $\Psi_i$ in (12) satisfies condition C1.

**Proof.** See Appendix A.2.

**Remark 1.** The bias-eliminating matrix in (12) depends on the noise variance $\sigma_o^2$ which is assumed to be known. In case the variance is unknown, it can be estimated either through dedicated experiments, or can be tuned via cross-validation treating it as a hyper-parameter to be optimized.

### 3.3 Bias-corrected estimate

Using the results presented in the previous paragraphs, the bias-corrected least squares estimate is obtained by replacing $X_i^\top \Delta X_i$ in (11) with the bias-eliminating matrix $\Psi_i$ (eq. (12)), i.e.,

$$\theta_i^{\text{BC}} = \left( X_i^\top X_i + \Psi_i / N_i \right)^{-1} X_i^\top \psi_i / N_i \quad (13)$$

**Proposition 3.** Assume that the following limit exists:

$$\lim_{N_i \to \infty} \frac{X_i^\top X_i + \Psi_i}{N_i} = \frac{X_i^\top \psi_i}{N_i},$$

then, the bias-corrected estimate $\theta_i^{\text{BC}}$ in (13) is a consistent estimate of the true model parameter $\theta_i^*$, i.e.,

$$\lim_{N_i \to \infty} \theta_i^{\text{BC}} = \theta_i^*, \text{ w.p. 1}$$

**Proof.** By substituting the output $\psi_i$ (see eq. (6)) into the right-hand side of (13) we get,

$$\lim_{N_i \to \infty} \theta_i^{\text{BC}} = \lim_{N_i \to \infty} \left( \frac{X_i^\top X_i + \Psi_i}{N_i} \right)^{-1} \left( X_i^\top X_i + X_i^\top \Delta X_i / N_i \right) \theta_i^*$$

$$+ \left( X_i^\top X_i + \Psi_i / N_i \right)^{-1} X_i^\top \psi_o / N_i \quad (14)$$

Since the regressor $[x^\top (k)]^\top$ is uncorrelated with the zero-mean white noise $e_o(k)$, the second term in (14) asymptotically (as $N_i \to \infty$) converges to 0 with probability 1, i.e.,

$$\lim_{N_i \to \infty} \left( \frac{X_i^\top X_i + \Psi_i}{N_i} \right)^{-1} X_i^\top \psi_o / N_i = 0, \text{ w.p. 1.}$$

Moreover, as $\Psi_i$ satisfies condition C1 (see Proposition 2), the first term in (14) asymptotically converges to the identity matrix $I$, with probability 1, i.e.,

$$\lim_{N_i \to \infty} \left( \frac{X_i^\top X_i + \Psi_i}{N_i} \right)^{-1} \left( X_i^\top X_i + X_i^\top \Delta X_i / N_i \right) = I, \text{ w.p. 1.}$$

With the above considerations and using (14), Proposition 3 follows.

**Remark 2.** For consistency proof, we assume that the input signal $u$ has been chosen in such a way that all the affine dynamics (i.e. all $s$ modes) of the PWA system are sufficiently excited, so that as the number of input/output observations $N \to \infty$ also implies $N_i \to \infty$, for all $i = 1, \ldots, s$.

### 4. PWA-OE IDENTIFICATION ALGORITHM

In the previous section, consistent bias-corrected estimates of the affine submodel parameters have been derived under the assumption that the underlying discrete mode sequence is known. In this section, we relax this assumption and we estimate the mode sequence along with the bias-corrected parameter estimates in an iterative manner. Using the estimated mode sequence, the partition of the regressor space $\mathcal{X}$ is computed using multi-category discrimination techniques.
4.1 Iterative clustering and parameter estimation

Algorithm 1 summarizes the main ideas of the proposed approach which involves the computation of the bias-corrected estimates of the affine submodel parameters \( \{ \theta_i \} \) as well as the unknown discrete state \( \{ \sigma(k) \} \) and the clusters \( \{ C_i \} \) characterizing the regressor space partition. The cluster \( C_i \) is constructed by stacking all (estimated) regressors \( \hat{x}(k) \) associated to mode \( i \) and its centroid \( c_i \) is defined as:

\[
c_i = \frac{1}{N} \sum_{k} x(k)_{i} \in C_i, \hat{x}(k).
\]

Given an initial guess \( \{ \sigma^0(k) \} \) of the mode sequence, at each iteration \( m \geq 1 \), Algorithm 1 alternates between the computation of the bias-corrected model parameters \( \{ \theta^m_i \} \) for a fixed mode sequence \( \{ \sigma^{m-1}(k) \} \) obtained at iteration \( (m-1) \) (see Step 1.1), and the estimation of the mode sequence \( \{ \sigma^m(k) \} \), for fixed bias-corrected model parameters \( \{ \theta^m_i \} \) obtained at the \( m \)-th iteration (Step 1.3). The details of Steps 1.1 and 1.3 are described as follows:

At Step 1.1, for each mode \( i = 1, \ldots, s \), based on the estimated mode sequence \( \{ \sigma^{m-1}(k) \} \) obtained at the previous iteration, the matrices \( X_i \) and \( Y_i \) are constructed from the data (eqs. (3) and (4)) and the bias-correcting matrix \( \Psi_i \) is computed as (12). The bias-corrected estimates \( \{ \theta^m_i \} \) are then computed and used at Step 1.3 to estimate the mode sequence.

For each time index \( k = \max(n_a, n_b) + 1, \ldots, N \), the simulated regressor \( \hat{x}(k) \) is computed based on the simulated outputs \( \hat{y}(k) = y(k) \) for \( k = 1, \ldots, \max(n_a, n_b) \). At Step 1.3.2, the prediction-error \( e_i(k) \) is computed for all modes \( i = 1, \ldots, s \). Note that, \( e_i(k) \) is the output-error computed based on the bias-corrected estimate \( \theta^m_i \) and simulated regressor \( \hat{x}(k) \). Step 1.3.3 selects the best mode \( \sigma^m(k) \) to which the regressor \( \hat{x}(k) \) is associated with. In particular, the clustering criterion for choosing the best mode at 1.3.3, minimizes the prediction-error \( e_i(k) \) and the distance between \( \hat{x}(k) \) and the centroid \( c_i \) of the cluster \( C_i \) which contains all the regressors already assigned to mode \( i \). A positive hyper-parameter \( \lambda \) is used to weigh these two terms. The centroid penalty \( |\hat{x}(k) - c_i|^2 \) takes into account the assumption that regressors “close” to each other are likely to belong to the same cluster.

Algorithm 1 Iterative bias-corrected parameter estimation and clustering.

**Input:** Observation sequence \( \{ x(k), y(k) \} \); number of modes \( s \); tuning parameter \( \lambda \); initial guess of the mode sequence \( \{ \sigma^0(k) \} \); number of iterations \( M \).

1. iterate for \( m = 1, \ldots, M \) do
   1.1. compute \( \{ \theta^m_i \} \) for fixed \( \{ \sigma^{m-1}(k) \} \):
   \[
   \theta^m_i = \left( \frac{X_i Y_i + \Psi_i}{N_i} \right)^{-1} \frac{X_i Y_i}{N_i}; \quad \forall i = 1, \ldots, s.
   \]
   1.2. set \( C_i = \emptyset, \forall i = 1, \ldots, s; \)
   1.3. for \( k = \max(n_a, n_b) + 1, \ldots, N \) do
      1.3.1. let \( \hat{x}(k) = [\hat{y}(k-1) \ldots \hat{y}(k-n_a) u(k-1) \ldots u(k-n_b)]^T \)
      1.3.2. let for \( i = 1, \ldots, s \)
         \( e_i(k) = y_{i}(k) - (\theta^m_i)^T \hat{x}(k) \);
      1.3.3. let \( \sigma^m(k) = \arg \min_{i=1,\ldots,s} \lambda e_i^2(k) + ||\hat{x}(k) - c_i||^2_2 \);
      1.3.4. let \( \hat{y}(k) = (\theta_{\sigma^m(k)})^T \hat{x}(k) \);
      1.3.5. let \( C_{\sigma^m(k)} \) = \( C_{\sigma^m(k)} \cup \{ \hat{x}(k) \} \);
      1.3.6. let \( N_{\sigma^m(k)} \leftarrow N_{\sigma^m(k)} + 1 \);
      1.3.7. update centroid
         \[c_{\sigma^m(k)} \leftarrow \frac{1}{N_{\sigma^m(k)}} \sum_{\hat{x}(k) \in C_{\sigma^m(k)}} \hat{x}(k) \]
   1.4. end for;
   2. end for;

**Output:** Estimated parameters \( \theta^M, \ldots, \theta^M \); clusters \( C^M, \ldots, C^M \); sequence of active modes \( \{ \sigma^M(k) \} \).

4.2 Partitioning the regressor space

Given the clusters \( \{ C_i \} \) obtained from Algorithm 1, the partition \( \{ X_i \} \) of the regressor space \( X \) can be computed using the computationally efficient linear multiclassity discrimination algorithm proposed in (Breschi et al., 2016) and briefly discussed in this subsection to make the paper self-contained.

In order to separate the clusters \( C_1, \ldots, C_s \), we search for a piecewise-affine separator function \( \phi : \mathbb{R}^{n_x} \to \mathbb{R} \) defined as

\[
\phi(\hat{x}) = \max_{i=1,\ldots,s} \left( [\hat{x}^T - 1] [\gamma^i_{1 \gamma^i}] \right),
\]  

where \( \omega^i \in \mathbb{R}^{n_x} \) and \( \gamma^i \in \mathbb{R} \) are unknown parameters to be computed.

Let \( N_i \) denote the cardinality of the \( i \)-th cluster \( C_i \) and \( M_i \in \mathbb{R}^{N_i \times n_x} \) be the matrix obtained by stacking the regressors \( \hat{x}^T(k) \) belonging to \( C_i \) in its rows.

According to the formulation introduced in (Bennett and Mangasarian, 1994), if the clusters \( \{ C_i \} \) are linearly separable, then the following conditions are satisfied,

\[
|M_i \ - 1_N| [\gamma^i_{1 \gamma^i}] \geq |M_i \ - 1_N| [\gamma^j_{1 \gamma^j}] + 1_N, \quad i, j = 1, \ldots, s, \quad i \neq j,
\]  

where \( 1_N \) is an \( N_i \) dimensional vector of ones. Consequently, the piecewise-affine dimensional vector of ones. Consequently, the piecewise-affine-dimensional vector of ones. Consequently, the piecewise-affine-dimensional vector of ones. Consequently, the piecewise-affine-dimensional vector of ones.
\[
\begin{align*}
\phi(\bar{x}) &= \begin{bmatrix} \bar{x}^\top - 1 \end{bmatrix} \begin{bmatrix} \omega_i \gamma_i \end{bmatrix}, \quad \forall x \in C_i, \quad i = 1, \ldots, s, \\
\phi(\bar{x}) &\geq \begin{bmatrix} \bar{x}^\top - 1 \end{bmatrix} \begin{bmatrix} \omega_j \gamma_j \end{bmatrix} + 1, \quad \forall x \in C_i, \quad i \neq j.
\end{align*}
\]
(17)

Thus, based on (17), each polyhedron \( \{X_i\}_{i=1}^{s} \) is defined as

\[
X_i = \left\{ \bar{x} \in \mathbb{R}^{n_x} : [\bar{x}^\top - 1] \begin{bmatrix} \omega^+_j \gamma^+_j \end{bmatrix} \geq 1, \quad j = 1, \ldots, s, \quad j \neq i \right\}.
\]

The parameters \( \{\omega^+_i, \gamma^+_i\}_{i=1}^{s} \), are computed by minimizing the averaged squared 2-norm of the violation of the inequalities in (16), via the solution of the following convex optimization problem:

\[
\min_{\xi} \frac{1}{2} \sum_{i=1}^{s} \left( \left\| [\omega^+_i, \gamma^+_i]^\top \right\|^2 + (\gamma_i)^2 \right) + \sum_{i=1}^{s} \sum_{j=1}^{s} \frac{1}{N_i} \left\| \left( [M_i, -N_i, [\omega^+_j, \gamma^+_j] + 1]_{N_i} \right) \right\|_2^2,
\]
(18)

where \( \xi \) is the set of optimization variables, i.e., \( \xi = \{[\omega^+_1, \gamma^+_1]^\top \ldots [\omega^+_s, \gamma^+_s]^\top\} \), and for a given \( x \in \mathbb{R}^n \), \( x_+ \) denotes a vector whose \( i \)-th element is max\( \{x_i, 0\} \). The regularization parameter \( \kappa > 0 \) makes sure that the optimization problem (18) is strongly convex and has a unique global minimizer.

### 5. NUMERICAL EXAMPLE

The effectiveness of the proposed identification algorithm is shown via a numerical example. All computations are carried out on an i7 1.9-GHz Intel core processor with 32 GB of RAM running MATLAB R2019a.

We consider the data-generating system \( S_{\text{true}} \) introduced in (Bemporad et al., 2005), modified to the output-error structure in (1) as follows:

\[
y_0(k) = \begin{cases} 
-0.4y_0(k-1) + u(k-1) + 1.5, & \text{if } 4y_0(k-1) - u(k-1) + 10 < 0, \\
0.5y_0(k-1) - u(k-1) - 0.5, & \text{if } 4y_0(k-1) - u(k-1) + 10 \geq 0, \\
-0.3y_0(k-1) + 0.5u(k-1) - 1.7, & \text{if } 5y_0(k-1) + u(k-1) - 6 < 0, \\
0.3u(k-1) + 0.5u(k-1) - 1.7, & \text{if } 5y_0(k-1) + u(k-1) - 6 \geq 0,
\end{cases}
\]
(19)

The system is characterized by \( s = 3 \) discrete modes. The input signal \( u \) is generated from a uniform random distribution taking values in the interval \([-4, 4]\). The noise \( \epsilon_o \), corrupting the output signal is generated by a zero-mean white Gaussian process with variance \( \sigma^2_o = 0.64 \), which corresponds to the Signal-to-Noise Ratio (SNR),

\[
\text{SNR} = 10 \log \frac{\sum_{k=1}^{N} \epsilon_o^2(k)}{\sum_{k=1}^{N} y_0^2(k)} = 11.7 \text{ dB}.
\]

The training dataset consist of \( N = 5000 \) input/output samples gathered from the system (19), of which 1313, 2015 and 1672 samples are generated from each of the \( s \) affine subsystems.

For the identification, we consider the PWA-OE model structure defined in (2) with \( s = 3 \) modes, and model orders \( n_a = 1, n_b = 1 \). The tuning hyper-parameter \( \lambda \) is set to the inverse of the noise variance, i.e., \( \lambda = \sigma^2_o = 1.56 \). The model parameters and the unknown mode sequence are estimated by running Algorithm 1 for \( M = 20 \) iterations with a randomly generated initial guess \( \{\sigma^0(k)\}_{k=1}^{s} \) of the mode sequence. The computation time to run Algorithm 1 is 2.6 sec.

The estimated affine submodel parameters obtained with the proposed bias-correction approach are reported in Table 1, along with the ones obtained via standard least squares (i.e., using the LS estimate (5) in Algorithm 1). From the obtained results, it can be seen that the LS estimates are biased while the bias-corrected estimates match closely with the true system parameters. This is further highlighted by the obtained norms of the parameter estimation errors \( \|\theta^0 - \hat{\theta}\| \) reported in Table 2, for bias-corrected and LS estimates.

The accuracy of the estimated active mode sequence is expressed by the Mode-Fit (MF) index

\[
\text{MF} = \left( \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(\sigma^M(k) = \sigma^*(k)) \right) \times 100\%,
\]
(20)

where \( \mathbb{I}(\cdot) \) is the indicator function, \( \sigma^M(k) \) and \( \sigma^*(k) \) are the estimated and the true \( s \) modes at time \( k \), respectively. The MF index achieved over the training data is 98.76% with the bias-corrected estimates while it is 98.50% with the LS estimates.

Based on the clusters estimated through Algorithm 1, the polyhedral partition of the regressor space is computed by solving the linear multicategory discrimination problem (18), with regularization parameter \( \kappa = 10^{-5} \). The true and the estimated polyhedral partitions of the regressor space are shown in Fig. 1.

The estimated PWA-OE model is validated on a new dataset of length \( N_{\text{val}} = 500 \). Based on the estimated submodel parameters and the polyhedral partitions, the output is simulated in open-loop as shown in Fig. 2 with bias-corrected and LS estimates. Specifically, in validation, using the estimated parameters \( \{\omega^+_i, \gamma^+_i\}_{i=1}^{s} \) defining the partition of the regressor space, the active mode is selected based on the value of the separator function (15) and corresponding model parameter \( \theta^B_{IC} \) is used to simulate the output. For the sake of better visualization, only a subset of validation data is plotted. The MF index achieved over

<table>
<thead>
<tr>
<th>Mode</th>
<th>( \theta^0 )</th>
<th>( \theta^{LS} )</th>
<th>( \theta^{BC} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s = 1 )</td>
<td>(-0.4000)</td>
<td>(-0.3933)</td>
<td>(-0.1889)</td>
</tr>
<tr>
<td>( s = 1 )</td>
<td>1.0000</td>
<td>1.0007</td>
<td>0.9691</td>
</tr>
<tr>
<td>( s = 2 )</td>
<td>1.5000</td>
<td>1.4816</td>
<td>2.2521</td>
</tr>
<tr>
<td>( s = 3 )</td>
<td>-0.5000</td>
<td>-0.4943</td>
<td>-0.6273</td>
</tr>
<tr>
<td>( s = 3 )</td>
<td>-0.3000</td>
<td>-0.3132</td>
<td>-0.2489</td>
</tr>
<tr>
<td>( s = 3 )</td>
<td>0.5000</td>
<td>0.5056</td>
<td>0.5122</td>
</tr>
<tr>
<td>( s = 3 )</td>
<td>-1.7000</td>
<td>-1.6765</td>
<td>-1.9091</td>
</tr>
</tbody>
</table>

\[ \text{Bias-corrected and LS estimates. Specifically, in validation, using the estimated parameters } \{\omega^+_i, \gamma^+_i\}_{i=1}^{s} \text{ defining the partition of the regressor space, the active mode is selected based on the value of the separator function (15) and corresponding model parameter } \theta^B_{IC} \text{ is used to simulate the output. For the sake of better visualization, only a subset of validation data is plotted. The MF index achieved over} \]

\[ \text{Table 1. True } (\theta^*) \text{ and estimated model parameters: Least-squares } (\theta^{LS}) \text{ vs Bias-corrected } (\theta^{BC}) \text{ estimates} \]

---

\[ \text{The true mode sequence } \sigma^*(k) \text{ is considered only for evaluating the MF index and it is not used in the estimation algorithm.} \]
Table 2. Norm of the parameter estimation error. Least squares vs bias-corrected estimates.

<table>
<thead>
<tr>
<th>Mode</th>
<th>$|\theta^n - \theta^{BCE}|$</th>
<th>$|\theta^n - \theta^{LS}|$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s=1$</td>
<td>0.0196</td>
<td>0.7818</td>
</tr>
<tr>
<td>$s=2$</td>
<td>0.0670</td>
<td>0.2091</td>
</tr>
<tr>
<td>$s=3$</td>
<td>0.0275</td>
<td>0.2156</td>
</tr>
</tbody>
</table>

Future research activities include the problem of model structure selection (i.e., estimation of model orders $n_a$, $n_b$ and number of modes $s$) directly from available measurements. This requires to define a proper criterion to be optimized, taking into account the output-error model structure in the fitting cost.

REFERENCES


Appendix A. PROOFS

A.1 Proof of Proposition 1

By substituting the output equation (6) in (11), we get

\[
\theta_{i}^{\text{CLS}} = \left( \frac{X_i^T X_i + \Delta X_i^T \Delta X_i}{N_i} \right)^{-1} \left( \frac{X_i^T X_i + \Delta X_i^T \Delta X_i}{N_i} \right) \theta_i^0 \\
+ \left( \frac{X_i^T X_i + \Delta X_i^T \Delta X_i}{N_i} \right)^{-1} \frac{X_i^T E_i}{N_i}
\]

(A.1)

According to the stated assumptions, the regressors \( X_i \) is uncorrelated with noise \( E_i \), and thus, the second term in (A.1) asymptotically (as \( N_i \to \infty \)) converges to 0 w.p.1. Proposition 1 easily follows.

A.2 Proof of Proposition 2

We remark that, the regressors \( x_o(k), x(k) \), output \( y(k) \) and noise samples \( e_o(k) \) considered hereafter in this proof are all associated to the \( i \)-th mode, and they are thus indicated by dropping the index \( i \), in order not to burden the notation.

The bias-eliminating matrix \( \Psi_i \) is constructed by evaluating the expected value of the matrix \( E \{ X_i^T \Delta X_i \} \), such that \( E \{ \Psi_i \} = E \{ X_i^T \Delta X_i \} \) holds.

By definition of \( X_i \) and \( \Delta X_i \) (see (7)), it follows that,

\[
E \{ \Psi_i \} = E \left[ \begin{bmatrix} x_o(k) & \cdots & x_o(k-n_o) \end{bmatrix} \begin{bmatrix} x_o(k-n_o) \\ \vdots \\ x_o(k-n_o-n_b) \end{bmatrix} \right] = 0
\]

\[
E \left[ \begin{bmatrix} n_o & \cdots & n_{o+n_b} \end{bmatrix} \begin{bmatrix} o_{n_o+n_b} & \cdots & o_{n_o+n_b+n_b} \end{bmatrix} \right] = 0
\]

Note that, since the value of the output at time \( k \) does not depend on the past or on the future values of the zero-mean white noise \( e_o \), we obtain,

\[
E \{ y(k - \tau) e_o(k-j) \} = 0, \quad \forall \tau \neq j.
\]

\[
E \{ e_o(k - \tau) \} = 0, \quad \forall \tau
\]

For \( \tau = j \), it follows that,

\[
- E \{ y(k-j) e_o(k-j) \} = - E \{ y(k-j) + e_o(k-j) \} e_o(k-j)
\]

\[
= - E \{ y_o(k-j) + e_o(k-j) \} e_o(k-j) - E \{ e_o(k-j) e_o(k-j) \}
\]

\[
= 0 - \sigma_e^2 = - \sigma_e^2
\]

as the noise-free output \( y_o \) is statistically independent of \( e_o \) and since \( e_o \) is zero-mean white noise with variance \( \sigma_e^2 \).

Furthermore, as the input \( u \) is statistically independent of \( e_o \), we have,

\[
E \{ u(k - \tau) e_o(k-j) \} = 0 \quad \forall \tau, j.
\]

Thus, using the above considerations it follows that,
\[ E \{ \Psi_i \} = \]

\[
E \left\{ \sum_{k=1}^{N_i} \begin{bmatrix}
-\sigma_c^2 & 0 & \cdots & 0 \\
0 & -\sigma_c^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & -\sigma_c^2
\end{bmatrix} 0_{n_a \times (n_b+1)} \right\}
\]

\[
= -\sigma_c^2 N_i \begin{bmatrix}
I_{n_a \times n_a} & 0_{n_a \times (n_b+1)} \\
0_{(n_b+1) \times n_a} & 0_{(n_b+1) \times (n_b+1)}
\end{bmatrix}
\]

\[ = \Psi_i, \]
as defined in Proposition 2.

By the above construction we have \( E \{ \Psi_i \} = E \{ X_i^T \Delta X_i \} \), then from law of large numbers the matrix \( \Psi_i \) satisfies property C1. See (Piga and Tóth, 2014, Appendix A3) for details.