A bayesian framework for large-scale identification of nonlinear hybrid systems

Madary, Ahmad; Momeni, Hamid Reza; Abate, Alessandro; Larsen, Kim G.

Published in:
IFAC-PapersOnLine

DOI (link to publication from Publisher):
10.1016/j.ifacol.2021.08.508

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-5 (2021) 259–264

A Bayesian Framework for Large-Scale Identification of Nonlinear Hybrid Systems

Ahmad Madary [*,**] **Hamid Reza Momeni** [*]
Alessandro Abate [***] **Kim G. Larsen** [****]

[*] *School of Electrical and Computer Engineering, Tarbiat Modares University, Iran, (email: {ahmad.madary,momeni_h}@modares.ac.ir)*
[**] *Mechanical and Production Engineering department, Aarhus University, Denmark,(amadary@mpe.au.dk)*
[***] *Department of Computer Science, University of Oxford, UK (email: aabate@cs.ox.ac.uk)*
[****] *Department of Computer Science at Aalborg University, Denmark (email: kgl@cs.aau.dk)*

Abstract: In this paper, a two-level Bayesian framework is proposed for the identification of nonlinear hybrid systems from large data sets by embedding it in a four-stage procedure. At the first stage, feature vector selection techniques are used to generate a reduced-size set from the given training data set. The resulting data set then is used to identify the hybrid system using a Bayesian method, where the objective is to assign each data point to a corresponding sub-mode of the hybrid model. At the third stage, this data assignment is used to train a Bayesian classifier to separate the original data set and determine the corresponding sub-mode for all the original data points. Finally, once every data point is assigned to a sub-mode, a Bayesian estimator is used to estimate a regressor for each sub-system independently. The proposed method tested on three case studies.

Keywords: Nonlinear hybrid systems, Switched nonlinear ARX models, Bayesian inference, System identification, Occam's Razor principle, Large data sets

1. INTRODUCTION

Hybrid Systems (HSs) have gained considerable attention in the past few years, since many of the present-day embedded systems are in essence hybrid. Furthermore, HSs can be used to model complex nonlinear system by representing them a collection of simpler linear sub-systems. In this work, a HS is in the form of a Switched Auto-Regressive Exogenous (SARX) system, and is defined as

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

where $\boldsymbol{x}_i = [y_{i-1} \quad \dots \quad y_{i-n_a} \quad u_{i-1-n_k} \quad \dots \quad u_{i-n_b-n_k}]$ is the continuous state composed of n_b and n_a samples of lagged input u and output y respectively, n_k is the number of delayed samples, and e_i is the measurement noise. Also, $\lambda_i \in \{1, ..., n\}$ is an exogenous, time-dependent variable that denotes the discrete mode and determines which of the n sub-systems λ_i is active at a specific time: the corresponding dynamics are characterised by the terms f_{λ_i}. If the sub-systems f_{λ_i} are nonlinear, then the resulting system is a Switched Nonlinear ARX system (SNARX), which is assumed throughout this work.
The nonlinear sub-systems $\{f_j\}_{j=1}^n$ can be expressed as a weighted summation of kernel functions. Assuming a training data set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, this summation takes the following form, Lauer et al. (2010):

$$f_j(\boldsymbol{x}; \boldsymbol{\alpha}_j, b_j) = \sum_{i=1}^{N} \alpha_{ij} k_j(\boldsymbol{x}_i, \boldsymbol{x}) + b_j, \tag{1}$$

where the weights $\boldsymbol{\alpha}_j = [\alpha_{1j} \quad \dots \quad \alpha_{Nj}]^T$ and the bias term b_j are the parameters of j^{th} sub-system and $k_j(\cdot,\cdot)$ is a kernel function that satisfies Mercer's condition and represents the model structure $\mathcal{H}_j$. It should be noted that the parameters for each sub-system f_j are $\boldsymbol{\alpha}_j$ and b_j, while each model structure $\mathcal{H}_j$ has one or more hyper-parameters (e.g., the width of the Gaussian kernel).
The identification of nonlinear hybrid systems (NHS) includes a joint estimation of the best parameters for the nonlinear sub-systems $\{f_j\}_{j=1}^n$ and of the time-dependent switching signal $\lambda_i \in \{1, ..., n\}$, from the training data set $\mathcal{S}$. Therefore, this problem comprises of two sub-problems that should be solved jointly: *the identification of the switching signal* and *the estimation of each sub-system*. This leads to the following optimization problem:

$$\min_{\boldsymbol{\alpha}_j, \boldsymbol{b}_j} \left(\sum_{j=1}^{n} \frac{1}{n} \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j + \frac{C}{N} \sum_{i=1}^{N} \min_{j=1,\dots,n} \left(y_i - \hat{f}_j(\boldsymbol{x}_i) \right)^2 \right), \tag{2}$$

where C is a parameter (discussed below) and $\hat{f}_j$ takes the form of (1) and also encompass the parameters b_j. This optimization results in both the model parameters and the modes λ_i estimation based on the nested minimisation.

Several methods have been proposed for the identification of *linear* HSs (Paoletti et al. (2007)), while the identification of NHSs has been much less researched.
Research in Lauer and Bloch (2008) adopted the rep-

10.1016/j.ifacol.2021.08.508

resentation (1) and proposed a "product of error (PE)" estimation, which results in the following optimization problem:

$$\min_{\boldsymbol{\alpha}_j,\boldsymbol{b}_j} \left(\sum_{j=1}^{n} \frac{1}{n} \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j + \frac{C}{N} \sum_{i=1}^{N} \prod_{j=1}^{n} \left(y_i - \hat{f}_j(\boldsymbol{x}_i) \right)^2 \right).$$

While this method possess good generalization, it suffers from a clear setback: it has $n(N+1)$ variables that require a considerable amount of time and memory to solve the optimization for large data sets. Furthermore, its performance deteriorates significantly and the optimization is prone to assign nearly all the data points to only one of the sub-systems due to parameter over-fitting. Besides, the parameter C that controls the trade off between model complexity and data fitness should be determined by user. Finally, the PE estimator is not an accurate estimation for the minimum function, which reduces the performance of the method. A Gaussian approach with stochastic simulations is proposed in Scampicchio et al. (2018) to identify a switched system consisting of one linear and one nonlinear sub-system for a small data set. A sparse optimization technique, based on $\ell_0 - \ell_1$ norms, is proposed in Bako et al. (2010).

The identification of NHSs gets complicated when the number of the data points increases. Bianchi et al. (2018) models nonlinear systems as a linear combination of polynomial functional expansions, and uses a randomized approach to assign each data point to a sub-model and to select the model structure of the local models.

The Expectation Maximization (EM) framework is used in Brusaferri et al. (2020) and Xiujun et al. (2020) to identify a specific class of NHSs in the from of Switched Markov Nonlinear ARX (SMNARX) systems and switched nonlinear systems with multiple Hammerstein models, respectively.

Bianchi et al. (2020) works with a large data set, by modeling the nonlinear functions as finite-dimensional parameterized polynomial expansions. Le et al. (2013) applies sparse optimization techniques to NHSs by using the kernel expansion form (1). A reduced-size kernels technique is proposed in Lauer et al. (2010); Bloch et al. (2011) to make the method developed in Lauer and Bloch (2008) applicable to large data sets. The method uses a pre-processing step to form a smaller data set based on principal component analysis.

Considering the limitations of the polynomial expansions, and the presence of many hyper-parameters and prior knowledge required by the switching signal, we conclude that the kernel expansion method presented in Lauer et al. (2010); Bloch et al. (2011) possess better generalization, fewer hyper-parameters, and less assumptions. Nevertheless, while the Lauer et al. (2010) is applicable to large data sets, the other two issues are still in place: the accuracy of the PE estimator and the best choice of the compromise between data fitness and model complexity. The reason is that in Lauer et al. (2010) the identified model using the reduced data set makes the mode estimator to assign all the data to their respective sub-system. In this regard, having the optimal value for the trade-off coefficient, an accurate estimation of the "minimum function" and an accurate mode estimator is even more crucial for an accurate identification.

In this paper, a multi-stage Bayesian framework is introduced for the identification of NHSs. First, a subset of data is generated from the complete data set through feature vector selection techniques. Then, the reduced data set is fed into a two-level Bayesian framework for an initial identification. This framework employs a new and accurate estimation that greatly improves the accuracy of the procedure. Furthermore, the best values for the hyper-parameters that control the model complexity and data fitness are obtained automatically, which subsequently provides a model with the best trade-off between complexity and data-fitness. The output of this stage is the mode assignment for the data points in the reduced data set. Instead of using a mode estimator based on minimum error using the rough estimated model, a Bayesian classifier is used to assign all the data points in the complete data set to sub-systems, using the mode assignments from the pre-identification step. Once all data points are assigned to sub-systems, a Bayesian regressor is used to estimate each sub-system. It is worth mentioning that unlike the existing kernel expansion methods and randomized methods, the predictions provided by the proposed method subsumes the uncertainty in the model parameters and generates a probability distribution: this can be later used for instance in Markov chain Monte Carlo method to generate random samples for further applications.

2. GENERATING THE REDUCED-SIZE DATA SET

Identification of NHSs with large data sets using kernel expansion requires considerable resources (time and memory), since kernel expansion methods treat each points as a variable. However, a kernel space can be spanned with only a sub-set of the data points and this subset can be sufficient to represent the entire data set. In this paper, a reduced-size data set is constructed from the original data $\mathcal{S}$ and is used as the input to the proposed identification procedure to estimate a HS and to assign the data points to their respective sub-systems. To construct the reduced data set $\mathcal{D} = \{\mathcal{D}_j\}_{j=1}^{n}$ from the complete data set $\mathcal{S} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where each $\mathcal{D}_j = \{\boldsymbol{x}_i, y_i\}_{i=1}^{M_j}$ is a sub-set of M_j data points corresponding to the $\mathcal{H}_j$ kernel, the Feature Vector Selection (FVS) method presented in Baudat and Anouar (2003) is used. This method tries to find a relevant subset from the given data and to form a basis in a feature space. This way, the subset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$ contains M data points from the original data set, where $M = \sum_{j=1}^{n} M_j$.

3. BAYESIAN SETUP

Identification problem for SNARX systems comprises estimating several sets of parameters and hyper-parameters. In a Bayesian framework, this can be achieved by maximizing their respective posterior probabilities according to Bayes' rule in two levels of inference, where the evidence of the first level is the likelihood of the next level. The required parameters and hyper-parameters for this framework are:

- The vector of the model parameters, $\boldsymbol{\theta} = [\boldsymbol{\alpha}, \boldsymbol{b}]^T$, where $\boldsymbol{\alpha}$ is the vector of the model weights and $\boldsymbol{b}$ is the

vector of bias terms: $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1 ... \boldsymbol{\alpha}_n]^T$, $\boldsymbol{b} = [b_1 ... b_n]^T$ (n is the number of sub-systems);
- The vector of the model hyper-parameters, $\boldsymbol{\mathcal{X}} = [\boldsymbol{\mu}, \beta]$, which contains the variances for the prior distributions of the weights and the estimated noise;
- The family of kernels $\mathcal{H} = \{\mathcal{H}_j | j = 1, ..., n\}$: this is the family of the models with different structures and/or different values for the parameters (e.g. the width of the Gaussian kernel or the degree of the polynomial kernel).

3.1 First level of inference: Model parameters

The first level of inference is dedicated to calculating the vector of the model parameters through maximizing their posterior probabilities. Assuming a reduced-size training data set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^M$ consisting of M data points, the vector of hyper-parameters $\boldsymbol{\mathcal{X}}$ and the family of the kernels $\mathcal{H}$, the conditional posterior probability of the model parameters can be calculated according to the Bayes' rule where $P(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}}, \mathcal{H})$ is the prior probability distribution of the model parameters and $P(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\mathcal{X}}, \mathcal{H})$ is the likelihood of the data points.

$$P(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\mathcal{X}}, \mathcal{H}) = \frac{P(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\mathcal{X}}, \mathcal{H}) P(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}}, \mathcal{H})}{P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H})}, \quad (3)$$

The denominator of (3) is called the hyper-parameter evidence and as it will be shown, it is not a function of the model parameters. Therefore, it is usually ignored in the calculation process of the model parameters MacKay (1992). Assuming independence across sub-systems parameters, as well as independence of weights $\boldsymbol{\alpha}_j$ and bias terms b_j of sub-systems (cf. MacKay (1992, 1995)), one can write the conditional probability of the prior distribution over the model parameters as:

$$P(\boldsymbol{\alpha}, \boldsymbol{b}|\boldsymbol{\mathcal{X}}, \mathcal{H}) = \prod_{j=1}^{n} P(\boldsymbol{\alpha}_j|\boldsymbol{\mathcal{X}}, \mathcal{H}) P(b_j|\boldsymbol{\mathcal{X}}, \mathcal{H}). \quad (4)$$

Next, a normal distribution with zero mean and covariance matrix of $\mu_j^{-1} I_M$ is assumed for the prior distribution of the weights $\boldsymbol{\alpha}_j$ of the j^{th} sub-system:

$$P(\boldsymbol{\alpha}_j|\boldsymbol{\mathcal{X}}, \mathcal{H}) = \frac{1}{Z_{\boldsymbol{\alpha}_j}} e^{-\frac{\mu_j}{2}\boldsymbol{\alpha}^T\boldsymbol{\alpha}}; Z_{\boldsymbol{\alpha}_j} = \left(\frac{2\pi}{\mu_j}\right)^{\frac{M}{2}}. \quad (5)$$

In (5), μ_j shows how sure we are about the weights a priori and is discussed further in Section 4. The second term in (4) is the prior probability distribution on the bias terms which due to the lack of prior information, is usually considered to be uninformative, MacKay (1992). The conditional distribution of $P(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\mathcal{X}}, \mathcal{H})$ is the likelihood term. The complete likelihood can be written after the data points are assigned to their respective sub-systems. This can be achieved by using maximum likelihood principle, Lauer and Bloch (2008) that tries to assign each data point $(\boldsymbol{x}_i, y_i)$ to the sub-system that most likely generates the data point, i.e. the one that maximizes the likelihood of the data with respect to the estimated sub-system $\hat{f}_j$. This can be written as:

$$\begin{aligned} \hat{\lambda}_i &= \underset{j=1,...,n}{\arg\max}\ P(y_i|\boldsymbol{x}_i, \hat{f}_j), \\ P(y_i|\boldsymbol{x}_i, \hat{f}_j) &= \frac{e^{-\ell\left(y_i - \hat{f}_j(\boldsymbol{x}_i)\right)}}{Z_\delta}, \end{aligned} \quad (6)$$

where $\ell(.)$ is a proper loss function and Z_δ is a normalizing constant, while $\hat{f}_j$ is the estimated model of the j^{th} sub-system. Here a Gaussian distribution with the variance of $1/\beta$ is chosen as the likelihood function. The term $1/\beta$ represents our prior belief on the noise variance of the system. Besides, $y_i - \hat{f}_j(\boldsymbol{x}_i)$ is the prediction error and $P(y_i|\boldsymbol{x}_i, \hat{f}_j)$ is the probability density function of the prediction errors, Ljung (1999). A standard assumption in the system identification is that the prediction errors are independent Ljung (1999): under this assumption, one can write the the complete likelihood of the data as:

$$P(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\mathcal{X}}, \mathcal{H}) = \prod_{i=1}^{M} \frac{1}{Z_\delta} e^{\underset{j}{\arg\max} -\frac{\beta}{2}\left(y_i - \hat{f}_j(\boldsymbol{x}_i)\right)^2}, \quad (7)$$

where $Z_\delta = (\frac{2\pi}{\beta})$. At this stage, the the prior distribution of parameters and the complete likelihood of the data can be combined to obtain the posterior probability of the model parameters as:

$$\begin{aligned} P(\boldsymbol{\alpha}, \boldsymbol{b}|\mathcal{D}, \boldsymbol{\mathcal{X}}, \mathcal{H}) &= \frac{\prod_{i=1}^{M} Z_\delta^{-1} \prod_{j=1}^{n} Z_{\boldsymbol{\alpha}_j}^{-1} e^{-\mathcal{J}_1(\boldsymbol{\alpha}, \boldsymbol{b})}}{P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H})}, \\ \mathcal{J}_1(\boldsymbol{\alpha}, \boldsymbol{b}) &= \sum_{j=1}^{n} \frac{\mu_j}{2} \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j + \frac{\beta}{2} \sum_{i=1}^{M} \underset{j=1,...,n}{\arg\min} \left(y_i - \hat{f}_j(\boldsymbol{x}_i)\right)^2. \end{aligned} \quad (8)$$

The normalizing term $P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H})$ in (8) is the evidence of the hyper-parameters and will be used as the likelihood in the next level of inference. To obtain the model parameters, one should maximize the posterior probability distribution. This results in maximum a posteriori estimation of the parameters, designated as $\boldsymbol{\alpha}^{MAP}$ and $\boldsymbol{b}^{MAP}$. It is more convenient to minimize the negative logarithm of the posterior distribution, that is $\min_{\boldsymbol{\alpha}, \boldsymbol{b}} \mathcal{J}_1$. To circumvent the obtained mixed optimization on both continuous and discrete variables one can replace the min function on discrete variables with a smooth approximation of it: Lauer et al. (2009) proposes the PE estimator as such an estimation, however it is not the best smooth approximation. In this study, instead, the *minimum of logarithm of Summation of Exponential (MinLSE)* function is proposed to replace the min function. The MinLSE function is defined as follows.

Definition 1. *The MinLSE function for a set of* $\{x_j\}_{j=1}^n$ *is defined as*

$$MinLSE(x_1, ..., x_n) = -\kappa^{-1} \log\left(\sum_{j=1}^{n} \exp(-\kappa x_j)\right),$$

where $\kappa > 0$ *is a scale factor to further improve the accuracy of the approximation.*

Using the MinLSE function, the optimization problem (8) is re-written as follows.

$$\min_{\boldsymbol{\alpha}, \boldsymbol{b}} \sum_{j=1}^{n} \frac{\mu_j}{2} \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j + \frac{\beta}{2} \sum_{i=1}^{M} MinLSE_{j=1,...,n}\left(\left(y_i - \hat{f}_j(\boldsymbol{x}_i)\right)^2\right),$$

After calculating the optimal values for the sub-system parameters, the estimated sub-systems $\hat{f}_j$ is calculated using $\hat{f}_j(\boldsymbol{x}; \boldsymbol{\alpha}_j, b_j) = \sum_{i=1}^{M} \alpha_{ij} k_j(\boldsymbol{x}_i, \boldsymbol{x}) + b_j$. At this stage, since the estimated sub-systems are known, the discrete mode of each data point can be calculated by utilizing the maximum likelihood principle: the probability

of each data point belonging to all the sub-system is calculated. The data point belongs to the sub-system with the highest probability. Substituting the optimal values of the sub-system parameters obtained earlier in the maximum likelihood estimation in (6) results in:

$$\hat{\lambda_i} = \underset{j=1,...,n}{\arg\max}\ P(y_i|\boldsymbol{x}_i, \hat{f}_j(.;\boldsymbol{\alpha}^{MAP}, b^{MAP})), \qquad (9)$$

where $i = 1, ..., M$. The posterior distribution of the model parameters can be summarized using the calculated values for $\boldsymbol{\alpha}^{MAP}, \boldsymbol{b}^{MAP}$ and the confidence interval on these maximum a-posteriori parameters. The confidence intervals are calculated from the curvature of the posterior distribution MacKay (1992). The posterior can be approximated locally with a Gaussian distribution as:

$$P(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\mathcal{X}}, \mathcal{H}) \approxeq P(\boldsymbol{\theta}^{MAP}|\mathcal{D}, \boldsymbol{\mathcal{X}}, \mathcal{H}) \exp\left(-\frac{1}{2} \triangle \boldsymbol{\theta}^T \Sigma \triangle \boldsymbol{\theta}\right), \qquad (10)$$

where $\boldsymbol{\theta}^{MAP} = [\boldsymbol{\alpha}^{MAP}, \boldsymbol{b}^{MAP}]^T$ and $\triangle\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}^{MAP}$. In (10), Σ is the Hessian matrix, namely $\Sigma = -\bigtriangledown^2 \log P(\boldsymbol{\alpha}, \boldsymbol{b}|\mathcal{D}, \boldsymbol{\mathcal{X}}, \mathcal{H})$, and the covariance of $\mathcal{J}_1$ is equal to Σ^{-1}. The accuracy of this approximation depends on the problem. For the quadratic term that is used in this research, the approximation is exact MacKay (1992). After the most probable values of parameters have been obtained, the mode estimation will be done according to (9) and values of λ_i are calculated for each data point. The estimated modes can be encoded in a discrete variable B_{ij} that is defined as

$$B_{ij} \in \{0, 1\},\ \forall i = 1, ..., M\ \ j = 1, ..., n,$$
$$s.t\ \ B_{ij} = 1 \text{ iff } \lambda_i = j \text{ and } \sum_{j=1}^{n} B_{ij} = 1,$$

which encodes each data point to a sub-system. Introducing this discrete variable into $\mathcal{J}_1$ in (8), it can be re-written as

$$\mathcal{J}_1 = \sum_{j=1}^{n} \frac{\mu_j}{2} \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j + \frac{\beta}{2} \sum_{i=1}^{M} \sum_{j=1}^{n} B_{ij} \left(y_i - \hat{f}_j(\boldsymbol{x}_i)\right)^2.$$

The first term in this equation is called *regularization*, which expresses the kind of smoothness we expect from resulting model MacKay (1992). The second term is the data fitness.

3.2 Second level of inference: Hyper-parameters

The purpose of the second and third levels of inference is to obtain the optimal values for the model hyper-parameters. It is important to obtain the optimal values for these hyper parameters since the model parameters depend heavily on the values of prior variances of the weights and noise, as they can cause severe under-fitting or over-fitting MacKay (1992, 1995) (depending on the values of model parameters and the ratio β/μ_j). Furthermore, since the purpose is not only to fit models on the data, but also to estimate the switching sequence. Improper values for μ_j,β and model parameters may result in the wrong mode estimation.
The second level of inference is dedicated to maximizing the posterior distribution of the hyper-parameters given the data points and the model using Bayes formula. This posterior probability distribution is expressed as

$$P(\boldsymbol{\mathcal{X}}|\mathcal{D}, \mathcal{H}) = \frac{P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H}) P(\boldsymbol{\mathcal{X}}|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})},$$

where $P(\boldsymbol{\mathcal{X}}|\mathcal{H})$ is the prior distribution given the model set $\mathcal{H}$ and assumed to be flat over logarithmic scaleMacKay (1995), since before the training, little information is known about the optimum values of the hyper-parameters. Besides, the evidence of the model is $P(\mathcal{D}|\mathcal{H})$.

3.3 Second level of inference: Likelihood

The likelihood of the training data given the model hyper-parameters and model family $\mathcal{H}$ is $P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H})$. According to (3), this expression is the evidence of the first level of inference. Assuming a uniform prior for hyper-parameters one can maximize the likelihood of the second level to maximize the posterior distribution. By marginalizing over the model parameter using the following integral, the evidence of the first level can be calculated, MacKay (1992):

$$P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H}) = \int P(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\mathcal{X}}, \mathcal{H}) P(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}}, \mathcal{H}) d\boldsymbol{\theta}. \qquad (11)$$

It is expected that this posterior has a peak around the most probable values for the model parameters. By exploring this, one can approximate the evidence integral with the integrand's peak and its width $\triangle\boldsymbol{\theta}$, MacKay (1995). One can locally approximate the integral (11) as a Gaussian distribution with covariance matrix Σ, as follows:

$$P(\mathcal{D}|\boldsymbol{\mathcal{X}}, \mathcal{H}) = \prod_{j=1}^{n} Z_{\boldsymbol{\alpha}_j}^{-1} \prod_{i=1}^{M} Z_{\delta}^{-1} e^{-\mathcal{J}_1(\boldsymbol{\theta}^{MAP})} (2\pi)^{\frac{n(M+1)}{2}} |\Sigma|^{-\frac{1}{2}}, \qquad (12)$$

where $Z_{\boldsymbol{\alpha}_j} = \left(\frac{2\pi}{\mu_j}\right)^{\frac{M}{2}}$, $Z_{\delta} = \left(\frac{2\pi}{\beta}\right)^{\frac{1}{2}}$ and Σ is the Hessian matrix of the first-level cost function. The complexity of the model is controlled by the hyper-parameters μ_j. A large value for μ_j means low variance on prior distribution of weights. A model with such hyper-parameter will fits data within smooth functions, while a model with small μ_j (large freedom on the prior range of possible α) fits the data from possibly less smooth functions. The Occam's Razor principle states that this parameter should not be too high or too low, Tipping (2003). One of the most compelling aspects of the Bayesian approach is that it automatically applies the Occam's Razor principle by integrating out all the irrelevant variables. Namely, in the Bayesian framework simple models that sufficiently explain the data without unnecessary complexity are automatically the preferred choice Tipping (2003). This property holds even if the prior probability is completely uninformative, MacKay (1992). The most probable values of the hyper-parameters μ_j^{MAP} and β^{MAP}, can be obtained by minimizing the negative logarithm of the posterior probability. The output of this stage is an assignment of each data point in the reduced set $\mathcal{D}$ to a sub-system, forming a set of labels $\mathcal{L} = \{\lambda_i\}_{i=1}^{M}, \lambda_i \in \{1, ..., n\}$ that will be used to train a classifier to distinguish the data in the complete data set $\mathcal{S}$.

4. MODE ASSIGNMENT AND SUB-MODEL IDENTIFICATION

Having obtained the reduced-size sets of labels from the previous section, it is now possible to train a Bayesian

classifier to assign all the data points in $\mathcal{S}$ to their corresponding sub-system. To achieve this goal, an RVM classifier explained in Tipping et al. (2003) is used. This classifier has several advantages, including the possibility of choosing non-Mercer kernels, high sparsity, and probabilistic prediction, Tipping et al. (2003). The training data for the RVM classifier is called $\mathcal{T}$ and consists of triples of the input-output data in $\mathcal{D}$ and sub-system labels from $\mathcal{L}$: $\mathcal{T} = \{\boldsymbol{x}_i, y_i, \lambda_i\}_{i=1}^{M}$. To this stage, all the data points in $\mathcal{S}$ are assigned to a sub-system using the trained classifier. Now, it is possible to use an RVM estimator presented in Tipping et al. (2003) to estimate each single sub-system using the classified data. The following algorithm illustrates the proposed workflow of identification of a NHS with large data set.

Algorithm 1 Identification of NHSs from large data sets

1: *Collect N data from the HS under study*
2: *Set up the data set in appropriate format $\mathcal{S} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$*
3: *Apply the feature vector algorithm proposed in Baudat and Anouar (2003) to the raw data set $\mathcal{S}$ to find the reduced data set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{M}$*
4: *Apply the preliminary two-stage identification algorithm to the reduced data set $\mathcal{D}$ to identify system modes $\mathcal{L} = \{\lambda_i\}_{i=1}^{M}, \lambda_i \in \{1, ..., n\}$*
5: *Train a RVM classifier with the triples training set $\mathcal{T} = \{\boldsymbol{x}_i, y_i, \lambda_i\}_{i=1}^{M}$*
6: *Apply a RVM classifier from previous stage to the original data set $\mathcal{S}$ to assign all data to sub-systems*
7: *Apply a RVM regressor to the data assigned to each sub-system independently*

5. CASE STUDIES

In this section, three numerical examples are used to demonstrate the performance of the proposed method for the identification of NHSs. The simulations are implemented in Matlab R2019a and run on a laptop with Intel Core i5 1.6GHz CPU and 8GB of memory. The optimization problems are solved using the native Matlab *fminunc* function.

5.1 Full-data set vs the reduced-data set identification

In this part, a comparison is made between the identification based on the full and reduced data sets for a SNARX system. Here $N = 4000$ data points are generated from the following model:

$$y_i = \begin{cases} y_i = -0.9y_{i-1} + 0.5u_{i-1} + e_i & if \;\; \lambda_i = 1 \\ y_i = (0.8 - 0.5exp(-y_{i-1}^2))y_{i-1} & \\ \quad - 0.9y_{i-1}^2 + 0.9u_{i-1} + e_i & if \;\; \lambda_i = 2 \end{cases} \tag{13}$$

The system starts from a random initial condition y_0, with a random input uniformly distributed in the range $u_i \in [0\ \ 4]$ and a Gaussian noise with zero mean and standard deviation equal to 0.1. Two kernels are employed to identify this system: a linear kernel and a Gaussian kernel with parameter set to 0.22. The identification initiates from an initial hyper parameter $\boldsymbol{\mu} = [1\ \ 1]$ and $\beta = 100$. This system is identified with a data set of various sizes $N = \{100, 200, 400, 800, 1600, 3200\}$. The mode assignment results are presented in Table 1. In this table, %Total, %Sub-Sys1, and %Sub-Sys2 represent the percentage of correct data assignment over the N data points and for sub-system 1 and 2, respectively.
As it can be seen, the full data set identification has a very good performance when the number of the data points is small. However, as N increases its performance deteriorates considerably and assigns almost all data to only one sub-system. Although the reduced method has a lower performance compared with the full identification, it can handle larger data sets while maintaining a good performance. Notice the stark improvement in the runtime for the proposed method.

5.2 Comparison with existing methods

To perform a comparison against existing methods, a data set $\mathcal{S}$ with $N = 4000$ points is generated from the system from Bloch et al. (2011) with a uniformly distributed random sequence of $\lambda_i \in \{1, 2\}$, starting form the initial condition $y_0 = y_1 = 0.1$.

Table 1. Full/reduced data set identification

Type	N	%Total	%Sub-Sys1	%Sub-Sys2	Time(s)
Full	100	99	97.619	100	5.01
Reduced	86	70	54.76	81.03	3.56
Full	200	95.5	97.139	92.941	16.39
Reduced	97	74.5	51.764	91.30	8.11
Full	400	58.75	98.253	5.847	24.38
Reduced	96	68.5	50.877	81.66	9.24
Full	800	57.125	99.562	0.583	110.12
Reduced	103	64.124	77.551	54.048	8.68
Full	1600	43.06	99.27	0.766	776.1
Reduced	107	69.18	67.46	92.11	8.05
Full	3200	57.28	99.945	0.0654	3671.02
Reduced	111	68.15	41.527	88.22	7.71

This data is disturbed by a Gaussian noise with zero-mean and standard deviation $\sigma_e = 0.1$. This system is identified using the method presented in Bloch et al. (2011) that is based on the PE estimator, and the proposed MinLSE estimator explained earlier in Section 2. The nonlinear sub-system uses a Gaussian kernel with width $\sigma = 0.3$, while the regularization trade-off is set to $C = 100$. After performing data reduction using the method introduced in Section 3, the reduced data set $\mathcal{D}$ contains 102 data points. The average results for 100 trials are presented in Table 2. It must be noted that although the method in Bloch et al. (2011) does not have a classifier, but a minimum-error mode estimator, in the table the term "Classifier" is used for both methods and refers to the assignment of the data points to sub-systems.

Table 2. Benchmark against Bloch et al. (2011)

	Method	
	MinLSE	PE
Correct mode assignment on $\mathcal{S}$(%)	82.34	64.51
Classifier accuracy on $\mathcal{S}$(%)	77.45	60.64
Classifier accuracy on $\mathcal{D}$(%)	79.86	63.21
Time for data reduction	16.2	16.2
Time for identification	4.02	3.95
MSE	0.012	0.053

It can be seen that the proposed method (MinLSE) outperforms the existing method (PE) both in terms of data

assignment and estimation error. This is due to the higher accuracy of the MinLSE estimator compared with the PE one, and to the better performance of the Bayesian classifier compared with the minimum-error mode estimator used in Bloch et al. (2011). Our proposed method has performed identification on the reduced size data set $\mathcal{D}$, which has led to more correct data assignment to the corresponding sub-systems, and a more accurate classifier in terms of both model fitness (MSE).

5.3 Identification of a SNARX system

In this part, a SNARX system is identified using the proposed method: we focus on a comparison between the identification using optimized hyper-parameters introduced in Section 3.2, and using non-optimized values. This will illustrate the importance of obtaining optimized hyper-parameter and their influence on the pre-identification and the initial mode assignment, as well as on the overall performance of the identification procedure. We show that this can be done with no substantial increase in computational cost. System (13) from Section 5.1 is used, with the same parameters, hyper-parameters and initial conditions. The hyper-parameters have been optimized using the second level of inference from Section 3.2: their values have changed from $\boldsymbol{\mu} = [1\ 1]$ and $\beta = 100$ to $\boldsymbol{\mu} = [44.0102\ \ 53.7600]$ and $\beta = 1414$. The identification is then repeated with these new optimized hyper-parameters, and the results are reported in Table 3. It can be appreciated that optimizing the hyper-parameters considerably improves the performance of the initial mode assignment. Since this initial mode assignment is used to train a classifier to assign all the data points to sub-systems, this improvement enhances the overall performance of the procedure, both in terms of correct mode assignment and of identification error.

Table 3. Identification of a SNARX system

	Hyper-parameters	
	Not-optimized	Optimized
Correct mode assignment on $\mathcal{S}$(%)	81.34	98.34
Classifier accuracy on $\mathcal{S}$(%)	75.45	84.25
Classifier accuracy on $\mathcal{D}$(%)	79.32	91.56
Time for data reduction	16.6	16.6
Time for identification	3.72	4.35
MSE	0.044	0.023

6. CONCLUSIONS

In this paper, a multi-stage Bayesian framework has been developed for the identification of NHSs with large data sets. The method deals with the large data sets by using feature vector selection methods to obtain a reduced-size sub-set of the data and to assign these data to sub-systems. Numerical simulations show that under same conditions, the proposed method has better performance compared to existing methods. Furthermore, it is shown that the ability to produce optimized hyper-parameters greatly improve the performance of the identification.

REFERENCES

Bako, L., Boukharouba, K., and Lecoeuche, S. (2010). An ℓ_0-ℓ_1 norm based optimization procedure for the identification of switched nonlinear systems. In 49th IEEE Conference on Decision and Control (CDC), 4467–4472. IEEE.

Baudat, G. and Anouar, F. (2003). Feature vector selection and projection using kernels. Neurocomputing, 55(1-2), 21–38.

Bianchi, F., Prandini, M., and Piroddi, L. (2018). A randomized approach to switched nonlinear systems identification. IFAC-PapersOnLine, 51(15), 281–286.

Bianchi, F., Prandini, M., and Piroddi, L. (2020). A randomized two-stage iterative method for switched nonlinear systems identification. Nonlinear Analysis: Hybrid Systems, 35, 100818.

Bloch, G., Lauer, F., et al. (2011). Reduced-size kernel models for nonlinear hybrid system identification. IEEE Transactions on Neural Networks, 22(12), 2398–2405.

Brusaferri, A., Matteucci, M., and Spinelli, S. (2020). Estimation of switched markov polynomial narx models. arXiv preprint arXiv:2009.14073.

Lauer, F. and Bloch, G. (2008). Switched and piecewise nonlinear hybrid system identification. In International Workshop on Hybrid Systems: Computation and Control, 330–343. Springer.

Lauer, F., Bloch, G., and Vidal, R. (2010). Nonlinear hybrid system identification with kernel models. In 49th IEEE Conference on Decision and Control (CDC), 696–701. IEEE.

Lauer, F., Vidal, R., and Bloch, G. (2009). A product-of-errors framework for linear hybrid system identification. IFAC Proceedings Volumes, 42(10), 563–568.

Le, V.L., Lauer, F., Bako, L., and Bloch, G. (2013). Learning nonlinear hybrid systems: from sparse optimization to support vector regression. In Proceedings of the 16th international conference on Hybrid systems: computation and control, 33–42.

Ljung, L. (1999). System identification. Wiley encyclopedia of electrical and electronics engineering, 1–19.

MacKay, D.J. (1992). Bayesian interpolation. Neural computation, 4(3), 415–447.

MacKay, D.J. (1995). Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. Network: computation in neural systems, 6(3), 469–505.

Paoletti, S., Juloski, A.L., Ferrari-Trecate, G., and Vidal, R. (2007). Identification of hybrid systems a tutorial. European journal of control, 13(2-3), 242–260.

Scampicchio, A., Giaretta, A., and Pillonetto, G. (2018). Nonlinear hybrid systems identification using kernel-based techniques. IFAC-PapersOnLine, 51(15), 269–274.

Tipping, M.E. (2003). Bayesian inference: An introduction to principles and practice in machine learning. In Summer School on Machine Learning, 41–62. Springer.

Tipping, M.E., Faul, A.C., et al. (2003). Fast marginal likelihood maximisation for sparse bayesian models. In AISTATS.

Xiujun, C., Hongwei, W., Lin, W., and Zhengqing, X. (2020). Identification of switched nonlinear systems based on em algorithm. In 2020 39th Chinese Control Conference (CCC), 1337–1342. IEEE.