Bayesian nonparametric identification of piecewise affine ARX systems

Johan Wågberg * Fredrik Lindsten ** Thomas B. Schön *

* Department of Information Technology, Uppsala University, Uppsala, Sweden, (e-mail: {johan.wagberg, thomas.schon}@it.uu.se).
** Department of Engineering, University of Cambridge, Cambridge, United Kingdom, (e-mail: fredrik.lindsten@eng.cam.ac.uk).

Abstract: We introduce a Bayesian nonparametric approach to identification of piecewise affine ARX systems. The clustering properties of the Dirichlet process are used to construct a prior over the partition of the regressor space as well as the parameters of each local model. This enables us to probabilistically reason about and to identify the number of modes, the partition of the regressor space, and the linear dynamics of each local model from data. By appropriate choices of base measure and likelihood function, we give explicit expressions for how to perform both inference and prediction. Simulations and experiments on real data from a pick and place machine are used to illustrate the capabilities of the new approach.

© 2015, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.

Keywords: Bayesian, Non-parametric identification, Dirichlet processes, Piecewise linear, Autoregressive models.

1. INTRODUCTION

Piecewise affine (PWA) systems are a class of models of nonlinear systems. These models have gained popularity by their universal approximation properties (Breiman, 1993; Lin and Unbehauen, 1992) and the fact that they rely on and generalize the well known linear systems theory. In this paper we study system identification of piecewise affine ARX (PWARX) systems.

Consider first a switched ARX system of order \((n_z, n_u)\). Let \(y_t\) and \(u_t\) denote the output and a known deterministic input at time \(t\) respectively. Furthermore, let

\[
x_t = [y_{t-1} \ldots y_{t-n_z} \ u_{t-1} \ldots u_{t-n_u}]^T
\]

(1)

denote the regressor vector and let \(\varphi_t = [x_t^T \ 1]^T\). The output is then given by

\[
y_t = \varphi_t^T \varphi_t + \epsilon_t,
\]

(2)

where \(\epsilon_t\) is white noise and \(z_t\) is a switching signal determining the active parameter vector \(\varphi_t\) at each time instance. A PWARX model is a special case of the switched ARX obtained by letting

\[
z_t = k \quad \text{if} \quad x_t \in \mathcal{R}_k,
\]

(3)

where \(\mathcal{R}_k\) is a convex polyhedral partition of the regressor space. That is, the dynamics of the system are local with respect to the regressor \(x_t\). A PWARX system can be written more explicitly as,

\[
y_t = \begin{cases} 
\varphi_1^T \varphi_t + \epsilon_t & \text{if } x_t \in \mathcal{R}_1, \\
\vdots & \\
\varphi_K^T \varphi_t + \epsilon_t & \text{if } x_t \in \mathcal{R}_K,
\end{cases}
\]

(4)

where the parameters of the subsystems are given by \(\varphi_k\). There exists many different approaches to identification of PWARX systems; see Garulli et al., 2002 for a recent survey. How difficult this identification problem is to solve heavily depends on what is assumed to be known. If the regions in the partition are known, the identification problem is reduced to \(K\) independent least squares problem where both frequentistic and Bayesian methods are easily applied. However, when the regions \(\mathcal{R}_k\) as well as the number of modes \(K\) are unknown, the problem becomes considerably more challenging. Existing methods includes for example clustering-based approach (Ferrari-Trecate et al., 2003), sum-of-norms regularization (Ohlsson and Ljung, 2013), mixed integer programming (Roll et al., 2004) and the bounded error approach (Bemporad et al., 2005b). We make use of a Bayesian nonparametric model based on the Dirichlet process introduced in Ferguson (1973). The Dirichlet process has successfully been used in Fox et al. (2011), but for temporally switching systems, where \(z_t\) is modeled by a discrete Markov chain, and not PWA systems.

The use of Bayesian methods have a long history in system identification, see Peterka (1981) for an early account of this development. For more recent accounts, see e.g. Ninness and Henriksen (2010) and Juloski et al. (2005a), where the latter considers a Bayesian solution for the problem under study in this paper. However, there, the number of modes is assumed to be known. Despite their capabilities, Bayesian methods are still not broadly employed. Historically, the reason for this has been due to the computational challenges inherent in the approach.

Our main contribution is that we construct a Bayesian nonparametric model allowing for simultaneous and automatic identification of: 1. the partition of the regressor space, 2. the number of modes, and 3. the local linear dynamics of each mode. This construction together with our specifically tailored Gibbs sampler allows us to reason about uncertainties over the involved partitionings in a principled way. In particular, we are able to make probabilistic predictions, i.e. predictions equipped with confidence intervals. This can be very useful, for instance in control applications where the control action may depend on the belief in the predictions made by the model. Our method is inspired by Hannah et al. (2011), where...
regression based on generalized linear models is studied in a similar way, but for a different class of problems.

2. THE DIRICHLET PROCESS

2.1 Definition

The Dirichlet process (DP) (Ferguson, 1973) is a popular construction which makes it possible to reason about uncertainty over partitions. In this respect, the DP is a Bayesian nonparametric model. Nonparametric does not mean that the model has no parameters – instead it refers to a model having an unbounded number of parameters or that the number of parameters grows with data. The DP can be seen as a probability distribution over discrete probability measures with an infinite number of point masses. That is, a draw \( G \) from the DP is a random probability distribution. For \( G \) to be DP distributed, all of its marginal distributions have to be Dirichlet (Dir) distributed. More precisely, let \( \alpha \) be a positive scalar and let \( G_0 \) be a (possibly continuous) probability measure over the measurable space \((\Omega, \mathcal{F})\). If \( G \) is a random measure on \((\Omega, \mathcal{F})\) such that

\[
(G(A_1), \ldots, G(A_r)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_r)),
\]

for every finite measurable partition \( A_1, \ldots, A_r \) of \( \Omega \), then \( G \) is DP distributed. We write \( G \sim \text{DP}(\alpha, G_0) \), where \( \alpha \) is called the concentration parameter and \( G_0 \) is called the base measure. With probability 1, \( G \) is a probability measure with an infinite number of point masses, i.e. we can write

\[
G = \sum_{k=1}^{\infty} a_k \delta_{\theta_k},
\]

(6)

Here, \( \delta_\theta \) is a point mass located at \( \theta \). Both the weights \( \{a_k\}_{k=1}^{\infty} \) and the point mass locations \( \{\theta_k\}_{k=1}^{\infty} \) are random variables. Furthermore, \( \sum_{k=1}^{\infty} a_k = 1 \) with probability 1, implying that \( G \) is indeed a random probability measure.

2.2 Predictive Distribution

While the definition (5) and the explicit representation (6) are useful for understanding the properties of the DP, they are not very practical. Instead, we will work directly with the predictive distribution, which is an equivalent representation of the DP. The predictive distribution also highlights an important feature of the DP that we will make use of, namely its clustering properties. Let

\[
G \sim \text{DP}(\alpha, G_0),
\]

\[
\theta_1 | G \sim \text{i.i.d.} G, \quad i = 1, 2, \ldots
\]

(7a)

(7b)

and assume that \( n \) values \( \theta_{1:n} \triangleq (\theta_1, \ldots, \theta_n) \) have been drawn. The predictive distribution for \( \theta_{n+1} \) can then be written as

\[
P(\theta_{n+1} | \theta_{1:n}) = \int G(\theta_{n+1}) P(G | \theta_{1:n}) \, dG.
\]

(8)

Since \( G \) is discrete (see (6)), there is a non-zero probability that \( \theta_{n+1} = \theta_i \) for some \( i \leq n \). By marginalizing over \( G \), it is shown by Blackwell and MacQueen (1973) that the predictive distribution is given by

\[
\theta_{n+1} | \theta_{1:n} \sim \alpha \delta_{\theta_0} + \frac{1}{\alpha + n} \sum_{k=1}^{K} n_k \delta_{\theta_k}.
\]

Here, \( \theta_{1:K} \) are the \( K \) unique values among \( \theta_{1:n} \) and \( n_k \) is the number of parameters where \( \theta_i = \theta_k \), i.e.

\[
n_k = \sum_{i=1}^{n} \mathbb{1}(\theta_i = \theta_k),
\]

(9)

where \( \mathbb{1}(\cdot) \) is the indicator function. Equation (9) clearly shows the clustering property of the model. The probability of a new draw coming from a certain “cluster”, i.e. \( \theta_{n+1} \) having the exact same value as a previous draw, is proportional to the number of past draws from the same “cluster” – we obtain a “rich gets richer” property. The effect of this is that the expected number of unique values among \( \theta_{1:n} \) grows slowly, \( E[K] = \frac{n}{\alpha + 1} = O(n \log n) \) (Teh, 2010). The role of the concentration parameter \( \alpha \) can also be interpreted as the probability of a new draw coming from a previously unseen cluster, as this is proportional to \( \alpha \). Hence, a large \( \alpha \) implies many new clusters and the measure \( G \) is more concentrated around the base measure \( G_0 \).

3. DIRICHLET PROCESSES FOR PWARX IDENTIFICATION

3.1 Model construction

We now turn to the construction of a Bayesian nonparametric model of a PWARX system. Let us first consider the parameters of a PWARX system in a single region, i.e. an ARX system. We shall assume that the vector of past inputs and outputs \( x_t \) and the corresponding output \( y_t \) are jointly Gaussian distributed according to

\[
(x_t, y_t) \sim N(\mu, \Sigma), \quad \mu = (\mu_x, \mu_y), \quad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix},
\]

(10)

where superscript \( x \) and \( xx \) denotes taking the corresponding part of the vector and the matrix, respectively. The conditional distribution of \( y_t \) given \( x_t \) is given by

\[
y_t | x_t \sim N(\mu_{y|x}, \Sigma_{y|x}),
\]

(11)

where

\[
\mu_{y|x} = \mu_y + \Sigma_{yy}^{-1}(\Sigma_{yx} - \mu_y), \quad \Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx} \Sigma_{y|x} \Sigma_{yx}.
\]

(12a)

(12b)

By letting

\[
\vartheta = \left[ \begin{array}{c} \Sigma_{yy}^{-1}(\Sigma_{yx} - \mu_y) \\ \Sigma_{yy}^{-1} - \mu_y \end{array} \right]
\]

(13)

the conditional distribution (11) can be written as

\[
y_t | x_t \sim N(\vartheta^T \varphi_t, \Sigma_{y|x}).
\]

(14)

That is, the distribution of \( y_t \) given \( x_t \) in this model is the same as in each subsystem in the PWARX system (4) with the variance of \( \varepsilon_t \) being \( \Sigma_{y|x} \).

In the Bayesian setting, we need a prior distribution over the unknown parameters. In this model, instead of working directly with \( \vartheta \), we can equivalently choose \( \vartheta \equiv (\mu, \Sigma) \) as the unknown and compute \( \vartheta \) using (13). As is very common in Bayesian statistical models (Bishop, 2006; Gelman et al., 2013), we use a conjugate prior; which in our case is the normal inverse Wishart (NIDW) distribution

\[
(\mu, \Sigma) \sim NIDW(\mu_0, \lambda, \Psi, \nu),
\]

(15)

with location parameter \( \mu_0 \in \mathbb{R}^d \), \( \lambda > 0 \), inverse scale matrix \( \Psi \in \mathbb{R}^{d \times d} \) and \( \nu > d - 1 \) where \( d \) is the dimension of \( \mu_0 \), see e.g. West and Harrison (1997) for details.

The above development results in a Bayesian ARX model or, put differently, a model for one of the subsystems in a PWARX model. To extend this to a full PWARX model, we make use of the DP construction.

Let

\[
G \sim \text{DP}(\alpha, G_0),
\]

(16)
where the base measure is given by the $\mathcal{NTW}$ prior, $\zeta_0 = \mathcal{NDW}(\mu_0, \lambda, \Psi, \nu)$. We then construct a hierarchical Bayesian PWARX model according to,

$$G \sim \text{DP}(\alpha, \zeta_0),$$

$$(\mu_i, \Sigma_i) \sim G, \quad t = 1, \ldots, T,$$

$$(x_t, y_t) \mid \mu_t, \Sigma_t \sim \mathcal{N}(\mu_t, \Sigma_t),$$

where $T$ is the number of input/output pairs. Note that we introduce a parameter $\theta_t = (\mu_t, \Sigma_t)$ for each time step $t = 1, \ldots, T$, which results in an overparameterisation of the model. However, due to the clustering property of the DP, many of these parameters will be identical, effectively resulting in a model with a small number of local subsystems. An important property of the model is that the regressor and the output will be clustered jointly, resulting in a (probabilistic) partition of the regressor space.

Since many of the parameters will be identical, the model can be reformulated in terms of the switching sequence $\{z_t\}_{t=1}^T$ and the unique parameter values. Let $\{\theta_k\}_{k=1}^K$ be the $K$ unique parameter values and define $z_t$ such that

$$z_t = k \Rightarrow \theta_t = \theta_k.$$  

The expression is similar to (9), but importantly (20) represents the clustering explicitly.

### 3.2 Computing the posterior

Let $w_t = (x_t, y_t)^T$ and let $D_T = \{w_t\}_{t=1}^T$ be an observed input/output data set with $T$ observations and let $\Theta$ denote all unknown parameters, e.g., for our model $\{(z_{t:1:T}, \{\mu_k^*, \Sigma_k^*, K\}_{k=1}^K) \text{ and } K\}$. In Bayesian estimation, the object of interest is the posterior distribution of the unknown parameters given data $p(\Theta \mid D_T)$. In some cases, such as when conjugate priors are used, this object can be computed analytically. This is the case above for the ARX model. In our PWARX model, however, this is not possible. We have chosen the base measure to be conjugate to the likelihood of the parameters in each cluster, but the Dirichlet process prior on $G$ makes exact inference impossible. Instead, approximate methods must be used.

There exists several inference algorithms that are tailored specifically for DPs (Neal, 2000; Blei and Jordan, 2006), all of which can be used for identification of the proposed PWARX model. Here, we consider a Markov chain Monte Carlo (MCMC) method. MCMC solves the inference problem by generating a sample path from an ergodic Markov chain, constructed so that its limiting distribution is equal to the posterior $p(\Theta \mid D_T)$. After an initial transient (the burn-in phase), the Markov chain produces samples from the posterior which can be used in computing estimators. More specifically, we use the Gibbs sampler given as Algorithm 3 in Neal (2000). In what follows, we will review this algorithm and derive the expression needed in implementing it for our specific model construction.

The conjugate base measure is not enough to enable exact inference in the model. However, conditionally on $z_{1:T}$, corresponding to a partitioning of the data points, the conjugacy of the base measure implies that we can compute the posterior for the parameters in each cluster analytically. Consequently, it is possible to use Rao-Blackwellise the Gibbs sampler and target the marginal distribution $p(z_{1:T} \mid D_T)$ directly by marginalizing out $\mu^*, K$ and $\Sigma^*, K$ in (19).

A Gibbs sampler constructs a Markov chain by sampling one parameter given the value of all the other parameters. To generate one sample from the posterior, the parameters are first ordered and then sampled sequentially from

$$p(z_{1} \mid z_{-1}, D_T), \quad t = 1, \ldots, n,$$

where $z_{-t}$ denotes all variables $z_{1:T}$ except $z_t$.

The complete algorithm for generating samples from the posterior $p(z_{1:T} \mid D_T)$ is summarized in Algorithm 1, where $M$ posterior samples $\{z_{1:T}[m]\}_{m=1}^M$ are generated. A good overview of this and other inference techniques for DP models is provided by Neal (2000).

**Algorithm 1 Gibbs sampler for $p(z_{1:T} \mid D_T)$**

**Require:** Starting state $z_{1:1} = 1$

1: for $m = 2$ to $M$ do
2: for $t = 1$ to $n$ do
3: Sample $z_t[m] \sim p(z_t \mid z_{1:t-1}[m], z_{t+1:T}[m-1], D_T)$
4: end for
5: end for

To be able to execute Algorithm 1, it remains to find the conditional distribution in (21), enabling the sampling step on line 3 to be performed. Rewriting (21) using Bayes’ theorem we get

$$p(z_{t} \mid z_{-t}, D_T) \propto p(z_t \mid z_{-t})p(w_t \mid z_{1:T}, D_{-t}).$$

The first factor is the prior $\eta$ over the clustering in (19). The clustering in a Dirichlet process is exchangeable, i.e. we can relabel and reorder the variables without changing the probability of a given configuration. Hence, we can use (20) to calculate the expression by imagining a rearrangement of $z_{1:T}$ such that $z_t$ comes last and takes the place of $z_T$. For the second factor, i.e. the marginal likelihood of the data item $w_t$, we note that, given the cluster assignments $z_{1:T}$, $w_t$ is conditionally independent of data items not belonging to cluster $z_t$. Hence, let $I_t = \{s : z_s = z_t\}$ be the indices of the data points belonging to the same cluster as $w_t$. We then have

$$p(w_t \mid z_{1:T}, D_{-t}) = \frac{p(D_{1:T} \mid z_{1:T})}{p(D_{1:T} \mid z_{1:T})},$$

where we use the convention $p(\emptyset) = 1$ if $|I_t| = 1$. For the conjugate model with an $\mathcal{NDW}$ prior, the marginal likelihood for all the data items in cluster $z_t$ is given by

$$p(D_{1:T} \mid z_{1:T}) = 1 \left(\frac{n^t \pi^t}{\lambda} \right)^{d/2} \frac{\Gamma_d(\nu'/2)}{\Gamma_d(\nu/2)} \left(\frac{\lambda}{\lambda'}\right)^{d/2},$$

where $d = \dim(x_t) + \dim(y_t)$, $\Gamma_d$ is the multivariate gamma function and the “primed” variables are given by

$$n' = |I_t|, \quad \nu'_0 = \lambda_0 + n' \bar{\mu}, \quad \lambda' = \lambda + n', \quad \nu' = \nu + n', \quad \Psi' = \Psi + C + \lambda n' (\bar{w} - \mu_0) (\bar{w} - \mu_0)^T, \quad \bar{w} = \frac{1}{n'} \sum_{t \in I_t} w_t, \quad C = \sum_{t \in I_t} (w_t - \bar{w})(w_t - \bar{w})^T.$$
An analogous expression holds also for \( p(D_{\mathcal{T}_k} | z_k) \) by removing the \( t^{th} \) data item, i.e. by excluding \( w_t \) from the sums and replacing \( n' \) by \( n' - 1 \). The outcome \( z_t = K + 1 \) implies that a new cluster is spawned. On the contrary, if at some point \( n_k = 0 \) a cluster is removed. This hints at our model construction’s inherent ability to automatically identify the number of modes of the PWARX model.

3.3 Making predictions

One of the main applications of an identified model is making predictions of future outcomes of the system. In this section, we derive explicit formulae for predicting a future output \( y \) given the regressor \( x \) for our model. Expressions for other likelihoods and non-conjugate base measures can be found in [9]. Let \( p(y \mid x, D_T) \) be the conditional distribution of \( y \) given \( x \) and training data \( D_T \) and let \( p(y \mid x, z_{1:T}, D_T) \) be the conditional distribution of \( y \) given \( x \), training data and a cluster assignment \( z_{1:T} \) for each training point in \( D_T \). The true conditional distribution is intractable to compute, but can be expressed as a conditional expectation

\[
p(y \mid x, D_T) = E_{z_{1:T} \mid D_T} [p(y \mid x, z_{1:T}, D_T)].
\] (25)

The expected value in (25) can not be computed explicitly, but it can be approximated by Monte Carlo integration using \( M \) posterior samples \( \{z_{1:T}[m]\}_{m=1}^M \) of \( z_{1:T} \mid D_T \) as

\[
p(y \mid x, D_T) \approx \frac{1}{M} \sum_{m=1}^M p(y \mid x, z_{1:T}[m], D_T).
\] (26)

With a slight abuse of notation, the conditional distribution of \( y \) given \( x, z_{1:T} \) and \( D_T \) can be written as

\[
p(y \mid x, z_{1:T}, D_T) = \frac{\alpha p(x, y) + \sum_{k=1}^K n_k p(x, y \mid D_{z_k})}{\alpha p(x) + \sum_{k=1}^K n_k p(x \mid D_{z_k})},
\] (27)

where \( \mathcal{I}_k = \{ t : z_t = k \} \) and the abuse of notation lies in \( p(x, y \mid D_{z_k}) \) and \( p(x \mid D_{z_k}) \). Here they represent the posterior predictive distribution of \( (x, y) \) and \( x \) given data in a standard Gaussian model (without the DP prior) with a \( \mathcal{NTW} \)-prior. The exact expressions are given by

\[
p(x, y \mid D_{\mathcal{I}_k}) = S_t \nu_{\mathcal{I}_k} \begin{pmatrix} \mu_0 + \frac{\lambda}{\lambda' + \alpha \sum_{t=1}^{K} n_k} \nu_{\mathcal{I}_k} + \Psi \end{pmatrix}
\] (28a)

\[
p(x \mid D_{\mathcal{I}_k}) = S_t \nu_{\mathcal{I}_k} \begin{pmatrix} \mu_x + \frac{\lambda}{\lambda' + \alpha \sum_{t=1}^{K} n_k} \nu_{\mathcal{I}_k} + \Psi \end{pmatrix}
\] (28b)

where the primed variables are calculated as in (24) with \( \mathcal{I}_k = \mathcal{I}_k \) and \( S_t \nu \begin{pmatrix} \mu, \Sigma \end{pmatrix} \) denotes the multivariate Student’s t-distribution with \( \nu \) degrees of freedom, mean \( \mu \) and a symmetric matrix parameter \( \Sigma \) that is related to the variance, see e.g. Roth (2013). Note that, for \( p(z_t \mid y) \) and \( p(x) \), we let \( D_{z_k} = \emptyset \). The expression (26) allows for computation of expected values with respect to the predictive density, i.e.

\[
E[h(y) \mid x, D_T] \approx \frac{1}{M} \sum_{m=1}^M E[h(y) \mid x, z_{1:T}[m], D_T].
\] (29)

From (27), the inner expectation can be written as

\[
E[h(y) \mid x, z_{1:T}, D_T] = \frac{\alpha \int h(y)p(x, y)dy + \sum_{k=1}^K n_k \int h(y)p(x, y \mid D_{z_k})dy}{\alpha p(x) + \sum_{k=1}^K n_k p(x \mid D_{z_k})}
\] (30)

In this expression, only \( \int h(y)p(x, y \mid D_{z_k})dy \) remains undetermined. We can write

\[
\int h(y)p(x, y \mid D_{z_k})dy = p(x \mid D_{z_k}) \int h(y)p(y \mid x, D_{z_k})dy
\] (31)

where \( p(x \mid D_{z_k}) \) is given in (28b). The second density \( p(y \mid x, D_{z_k}) \) is also Student’s t distributed (see e.g. Roth (2013) for expressions related to the conditional and marginal Student’s t distribution)

\[
p(y \mid x, D_{z_k}) = S_t \nu_{y_{x_k}} \begin{pmatrix} \mu_y \mid x, \Sigma_y \mid x \end{pmatrix}
\] (32a)

where

\[
\mu_{y \mid x} = \mu_y + \Psi \Psi y \begin{pmatrix} \mu y \mid x, \Sigma y \mid x \end{pmatrix}^{-1} (x - \mu x)
\] (32b)

\[
\nu_{y \mid x} = \nu' - d_y + 1,
\] (32c)

\[
\Sigma_{y \mid x} = \nu' - d + 1 + (x - \mu_x)^T \Psi y \begin{pmatrix} \mu y \mid x, \Sigma y \mid x \end{pmatrix}^{-1} (x - \mu x)^T
\] (32d)

\[
\Psi = \frac{\lambda + 1}{\lambda' + \nu' - d + 1} \Psi y
\] (32e)

\[
d_y = \dim(y)
\] (32f)

and the “primed” variables are given by (24) with \( \mathcal{I}_i = \mathcal{I}^k \).

We have given general expressions for estimation of an expected value of a general function \( h(y) \). In particular, we are interested in the expected value and the variance of \( y \) given \( x \). For the conditional expected value we get

\[
E[y \mid x, D_T] = \mu_{y \mid x}
\] (33)

and for the conditional variance

\[
\text{Var}[y \mid x] = \frac{\nu_{y \mid x}}{\nu_{y \mid x} - 2 \Sigma y_{y \mid x} + (\mu_{y \mid x} - E[y \mid x, D_T]^2)}.
\] (34)

4. EXPERIMENT AND RESULTS

4.1 One-dimensional PWARX system

Consider the following one-dimensional PWARX model (first introduced by Ferrari-Trecate et al. (2003))

\[
y_t = \begin{cases} u_{t-1} + 2 + \epsilon_t, & -4 \leq u_{t-1} \leq -1, \\ u_{t-1} + 2 + \epsilon_t, & -1 < u_{t-1} < 2, \\ u_{t-1} + 2 + \epsilon_t, & 2 \leq u_{t-1} \leq 4. 
\end{cases}
\] (35)

Data was generated from (35) by generating \( T = 50 \) samples of \( u_t \) uniformly on \([-4, 4] \) and \( \epsilon_t \sim \mathcal{N}(0, 0.05) \). The dataset \( \{(y_t, u_t)\}_{t=1}^{50} \) can be seen in Fig. 1. This example was chosen because it is small and easy to visualize. The hyper-parameters were fixed to

\[
\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \lambda = 1, \quad \Psi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \nu = 10.
\] (36)

To reduce the impact of the prior choice of the concentration parameter \( \alpha \), a hyperprior was used to identify \( \alpha \) from data. Inference was done by sampling \( \alpha \) given the cluster labels \( z_{1:T} \) in the Gibbs sampler, see West (1992) for details. The Gamma distribution (with shape \( \alpha = 1 \) and rate \( b = 1 \)) was used as hyperprior. The Gibbs sampler (Algorithm 1) was initialized with each data point in its own cluster, \( z_t = t \). After an initial 500 iterations (burn-in), \( M = 1000 \) samples were recorded. One representative sample from the Markov chain can be seen in Fig. 1 together with the empirical mean and covariance of the data points in each cluster represented as ellipses. A feature of the Bayesian model is that it does not make hard assignments concerning the locations of the boundaries between the different regions of the partition. This is evident from the figure, since there is an overlap in cluster assignment for the data points around \( u_{t-1} \approx -1 \).
3.3 Making predictions

In this expression, only the data item, i.e. by excluding the abuse of notation lies.

Finally, Fig. 3 shows the expected value and the maximum a posteriori (MAP) estimate for the prediction of $y_t$ given $u_{t-1}$. In areas where the distribution of $y_t$ is unimodal, the expected value is accurate, but in areas where the distribution on $y_t$ is bimodal, the MAP is a better estimate. As expected, the variance (visualized by the gray region) for the predictor is significantly larger in regions where the true system is discontinuous.

4.2 A Hammerstein system

Consider the system

$$y_t = -a_1 y_{t-1} - a_2 y_{t-2} + b_1 u_{t-1} + e_t,$$

where $v$ is a saturated version of $u$,

$$v_t = \begin{cases} u_{\text{max}} & \text{if } u_t > u_{\text{max}}, \\ u_t & \text{if } u_{\text{min}} \leq u_t \leq u_{\text{max}}, \\ u_{\text{min}} & \text{if } u_t < u_{\text{min}}. \end{cases}$$

The system from $u$ to $y$ is of a class of systems commonly referred to as a Hammerstein system. Since the saturation in (38) is piecewise affine, the entire system from $u$ to $y$ will also be piecewise affine. We let $a_1 = 0.5$, $a_2 = 0.1$, $b_1 = 1$, $u_{\text{max}} = 1$ and $u_{\text{min}} = -1$ (this setup was also used by Ohlsson and Ljung (2013)) and generate $T = 250$ measurements from the system with $u$ being zero mean Gaussian white noise with variance 4.

We evaluate the performance of our model by comparing it to an optimal estimator. With access to an oracle giving the correct cluster assignment, the parameters within each cluster are easily estimated using least squares. These parameters and the oracle directly gives the one step predictor. We generate 500 measurements from the system to use as test data. A scalar performance measure can be obtained by calculating the fit as

$$\text{fit}(\hat{y}_{1:T}, y_{1:T}) = 1 - \frac{|\hat{y}_{1:T} - y_{1:T}|}{|\bar{y}_{1:T} - \bar{y}|},$$

where $\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$ and $\hat{y}_{1:T}$ is an estimate of $y_{1:T}$. A fit value of 1 means perfect fit and a fit value of 0 means that the fit is no better then a straight line matching the mean of $y$. The fit for the optimal estimator was 76.5% and 74.8% for our model, indicating our model’s capability of not only finding the individual linear models, but also to automatically find a suitable partitioning.
partitioning, 2. the number of modes and 3. the parameters describing the local dynamics of each mode. The construction together with the specifically tailored Gibbs sampler allows us to reason about uncertainties over the involved partitionings in a principled way. For example, since the result is packaged as a posterior distribution rather than via explicit point estimates, our approach has no problems in dealing with the multimodal distributions, that often arise in PWARX models. This was indeed useful and important already in a very simple example. The so called sequential Monte Carlo methods (e.g. particle filters) can straightforwardly work with the model representations we have introduced in this work. In the future we foresee interesting developments, where model constructions of the type introduced in this work are used to solve challenging tasks where today’s standard methods fail.

Fig. 4. Validation output for the pick-and-place machine. Simulated output (solid red) plus minus two standard deviations (gray area) and measured system output (dashed black).

4.3 Pick-and-place machine

In this example, we study a pick-and-place machine. The pick-and-place machine is used to place electronic components on a circuit board. At least two modes can be distinguished, the free mode and the impact mode. In the free mode, the machine is carrying an electronic component, but is not in contact with the circuit board. In the impact mode, the electronic component is in contact with the circuit board. For a detailed description of the system, see Juloski et al. (2004).

The data is from a real physical process and it has previously been studied in Bemporad et al. (2005a); Juloski et al. (2003, 2005b, 2006); Ohlsson and Ljung (2013). It consists of a 15 s recording of the voltage input to the motor of the mounting head (input) of the pick-and-place machine and the vertical position of the mounting head (output). Both the input and output was sampled at 50 Hz and standardized by subtracting the mean and dividing by the standard deviation. The first 8 s was used for identifying the model and the last 7 s for validation.

The order of the PWARX model was set to $n_a = 2$ and $n_b = 2$. The hyperparameters of the base measure was set to: $\mu_0 = 0$, $\lambda = 0.6$, $\Psi = I$, $\nu = 40$ and for the gamma hyperprior over $\alpha$, shape parameter $a = 1$ and rate parameter $b = 1$ was used.

As in Ohlsson and Ljung (2013) and Juloski et al. (2006), we evaluate the model by simulation instead of one-step prediction. To simulate the model, 1 000 samples of the clustering $z \sim D_T$ was generated with the Gibbs sampler. For each of the clustering sample, one trajectory was simulated resulting in 1 000 simulated trajectories. The mean and the variance from these samples are reported in Fig. 4. The fit of the mean on the validation data was 77.7%. This is slightly lower than the 78.6% reported in Ohlsson and Ljung (2013). However, our method is able to correctly reports higher uncertainty in regions where the expected value is less accurate.

5. CONCLUSION

We have constructed a Bayesian nonparametric model for identification of PWARX models, enabling simultaneous and automatic identification of: 1. the regressor space

REFERENCES

Bemporad, A., Garulli, A., Paoletti, S., and Vicino, A. (2005a). A bounded-error approach to piecewise affine system identification. IEEE Transactions on Automatic Control, 50(10), 1567–1580.
Bemporad, A., Garulli, A., Paoletti, S., and Vicino, A. (2005b). A bounded-error approach to piecewise affine system identification. IEEE Transactions on Automatic Control, 50(10), 1567–1580.
Bishop, C.M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.
Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Polya urn schemes. The annals of statistics, 1(2), 355–357.
Blei, D.M. and Jordan, M.I. (2006). Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1(1), 121–144.
Breiman, L. (1993). Hinging hyperplanes for regression, classification, and function approximation. IEEE Transactions on Information Theory, 39(3), 999–1013.
Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics (3), 237–266.
Ferrari-Trecate, G., Muselli, M., Liberati, D., and Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. Automatica, 39(2), 205–217.
Fox, E., Sudderth, E.B., Jordan, M.I., and Willsky, A. (2011). Bayesian nonparametric inference of switching dynamic linear models. IEEE Transactions on Signal Processing, 59(4), 1569–1585.
Garulli, A., Paoletti, S., and Vicino, A. (2012). A survey on switched and piecewise affine system identification. In Proceedings of the 16th IFAC Symposium on System Identification (SYSID), volume 16, 344–355. Brussels, Belgium.
Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.
Hannah, L., Blei, D.M., and Powell, W.B. (2011). Dirichlet process mixtures of generalized linear models. Journal of Machine Learning Research, 12, 1923–1953.
Jouloski, A.J., Weiland, S., and Heemels, W.P.M.H. (2005a). A Bayesian approach to identification of hybrid systems. IEEE Transactions on Automatic Control, 50(10), 1520–1530.
Jouloski, A.J.L., Heemels, W., and Ferrari-Trecate, G. (2003). Identification of an experimental hybrid system. In IFAC conference on the analysis and design of hybrid systems (ADHS 03), 223–228.
Jouloski, A.L., Paoletti, S., and Roll, J. (2006). Recent techniques for the identification of piecewise affine and hybrid systems. In Current trends in nonlinear systems and control, 79–99. Springer.
Jouloski, A., Heemels, W., Ferrari-Trecate, G., Vidal, R., Paoletti, S., and Niessen, J. (2005b). Comparison of four procedures for the identification of hybrid systems. In M. Morari and L. Thiele (eds.), Hybrid Systems: Computation and Control, volume 3414 of Lecture Notes in Computer Science, 354–369. Springer Berlin Heidelberg.
Jouloski, A., Heemels, W., and Ferrari-Trecate, G. (2004). Data-based hybrid modelling of the component placement process in pick-and-place machines. Control Engineering Practice, 12(10), 1241–1252.
Lin, J.N. and Unbehauen, R. (1992). Canonical piecewise-linear approximations. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 39(8), 697–707.
Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics, 9(2), 249–265.
Ninness, B. and Henriksson, S.J. (2010). Bayesian system identification via Markov chain Monte Carlo techniques. Automatica, 46(1), 40–51.
Ohlsson, H. and Ljung, L. (2013). Identification of switched linear regression models using sum-of-norms regularization. Automatica, 49(4), 1041–1050.
Petruka, V. (1981). Bayesian system identification. Automatica, 17(1), 41–53.
Roll, J., Bemporad, A., and Ljung, L. (2004). Identification of piecewise affine systems via mixed-integer programming. Automatica, 40(1), 37–50.
Roth, M. (2013). On the multivariate t distribution. Technical Report LiTH-ISY-R-3059, Division of Automatic Control, Linköping, Sweden.
Teh, Y.W. (2010). Dirichlet processes. In Encyclopedia of Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.
West, M. (1992). Hyperparameter estimation in Dirichlet process mixtures. Technical Report 92-A03, Duke University.