2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Variational Learning of Autoregressive Mixtures of Experts for Fully Bayesian Hybrid System Identification

Nisar Ahmed and Mark Campbell
Autonomous Systems Laboratory, Sibley School of Mechanical and Aerospace Engineering
Cornell University, Ithaca, NY 14853, USA
{nra6,mc288}@cornell.edu

***Abstract*— This paper presents a new learning method for Mixture of Expert ARX (MEARX) models and its application to identification of PieceWise ARX (PWARX) hybrid systems models. While accurate deterministically-switched PWARX models are obtainable from probabilistically-switched MEARX models, important issues such as model structure selection (i.e. estimation of the number of modes and ARX lag orders) and estimation with sparse/noisy data remain open. This paper addresses these issues through a new variational Bayesian MEARX learning approximation. This not only permits computationally efficient estimates for MEARX/PWARX regressor weights and mode boundary parameters, but also allows for theoretically sound Bayesian model structure selection. Numerical hybrid system ID examples from the literature demonstrate the proposed approach.**

I. Introduction

The important and challenging problem of hybrid system identification has attracted much recent attention (e.g. see [1], [2]). Of particular interest is the problem of identifying *PieceWise AutoRegressive eXogenuous* (PWARX) models for nonlinear systems that can be described/approximated via non-Markov switching linear dynamics. As discussed in [3], PWARX models are a special case of the highly flexible piecewise affine (PWA) model class and are equivalent to other popular hybrid system model classes, for which sophisticated control design techniques have been recently developed and succesfully applied.

As discussed in [4], while many good methods exist for PWARX ID, many important problems still remain open. Chief among these is the need for reliable and theoretically sound solutions to the difficult problem of model structure selection, i.e. estimating the appropriate number of discrete system modes and the corresponding discrete ARX regression lag orders. Although some model selection procedures exist, these are sensitive to process noise, sparse data records, and tuning parameters that are often difficult to interpret physically [4], [5]. Refs. [6], [7] introduced Bayesian PWARX ID methods that use probability distributions to incorporate prior knowledge and extract model parameter confidence bounds, both of which are especially useful for sparse and noisy data. However, their reliance on computationally intensive Monte Carlo sampling procedures makes these Bayesian methods difficult to implement when the model structure is unknown. Ref. [4] further notes that the use of deterministic classifiers for discrete mode estimation in most PWARX ID methods increases computational cost and can create difficulties in estimating discrete mode boundaries. For instance, incomplete modal partitions or 'holes' may form in some cases, while post-hoc adjustments may be needed in other cases to correct large prediction errors. Ref. [8] recently proposed an elegant solution to this last problem, in which a *probabilistic* 'soft mode' generalization of PWARX known as Mixtures of Expert of ARX (MEARX) [9] is used to simultaneously estimate the ARX and boundary parameters for each discrete mode. However, the important issues of model selection and estimation with sparse/noisy data have so far remained open for this approach.

This paper addresses these issues through a new variational Bayesian (VB) learning approach for MEARX models. The proposed method leads to a computationally efficient 'fully Bayesian' MEARX/PWARX ID scheme, in that Bayes' rule is used to estimate model parameters and select appropriate model structures in a theoretically sound manner using only closed form iterative updates.

II. Background

A. PWARX ID problem formulation

Consider a discrete-time system with noisy observed output $y_k \in \mathbb{R}$ and deterministic mapping $f(\cdot)$,

$$y_k = f(v_k) + e_k, \tag{1}$$

where e_k is an error term. For fixed input lag $n_a \geq 1$, fixed output lag $n_b \geq 1$, and $n = n_a + n_b$, let $v_k \in \mathbb{R}^{n\times 1}$ be

$$v_k = [y_{k-1}, ..., y_{k-n_a}, u_{k-1}, ..., u_{k-n_b}]^T, \tag{2}$$

where $u_k \in \mathbb{R}$ represents a known exogeneous input (u_k can also be a vector, without loss of generality). If $x_k \equiv [v_k^T, 1]^T$, a PWARX model has the PWA mapping

$$f(v_k) = \begin{cases} f_1(x_k) = w_1^T x_k, \text{ if } x_k \in \mathcal{V}_1 \\ \vdots \\ f_m(x_k) = w_m^T x_k, \text{ if } x_k \in \mathcal{V}_m \end{cases}, \tag{3}$$

where $m \geq 1$ is the number of discrete modes and $w_j \in \mathbb{R}^{(n+1)\times 1}$ is the parameter vector for mode $j \in \{1, ..., m\}$. The sets $\mathcal{V}_j$ are convex polyhedra (or 'regions') over the regressor space $\mathcal{V} \subseteq \mathbb{R}^r$, satisfying

$$\mathcal{V}_j = \{v_k \in \mathbb{R}^r : H_j x_k \preceq 0\}, \tag{4}$$

where $H_j \in \mathbb{R}^{z_j \times r}$ and '$\preceq$' denotes elementwise inequality. As per [10], the regions $\mathcal{V}_j$ form a complete partition of $\mathcal{V}$, so

that each H_j represents $z_j \leq m-1$ linear inequalities defining where mode j is 'active' in $\mathcal{V}$. The matrix of modal ARX parameters and the set of region matrices are respectively denoted by $W \equiv [w_1, ..., w_m]$ and $\mathbf{H} \equiv \{H_1, ..., H_m\}$. The set of discrete structural model parameters is denoted here as $\Omega = \{m, n_a, n_b\}$. The PWARX ID problem can be stated as follows: given N measurement data $\mathbf{Y} = \{y_1, ..., y_N\}$ and corresponding inputs $\mathbf{U} = \{u_1, ..., u_N\}$, determine Ω and the corresponding parameters W and $\mathbf{H}$.

As [4] notes, the first major challenge of PWARX ID lies in simultaneously solving a set of regression problems (estimating W) and a classification problem (estimating $\mathbf{H}$) for fixed Ω. The second major challenge lies in identifying Ω itself: m can be arbitrarily large while n_a and n_b can be different in each discrete mode, which makes estimating W and $\mathbf{H}$ more complex. Furthermore, there are no obvious metrics by which to select Ω, especially in the case of sparse and noisy data.

B. PWARX ID via MEARX ID

Ref. [8] proposed PWARX ID using MEARX models [9], which are a special case of the Mixture of Experts (ME) model from the machine learning and neural network fields [11]. A general MEARX model is defined as a *non-Markov switching* probabilistic mixture of ARX models,

$$y_k = \sum_{i=1}^{m} P(s_k = i | x_k, \Theta) f_i(x_k) + e_k, \tag{5}$$

where the modal weighting function $P(s_k = i | x_k, \Theta)$ (with parameter set $\Theta = \{\theta_1, ..., \theta_m\}$) gives the conditional probability that the discrete mode state s_k equals $i \in \{1, ..., m\}$ as a function of x_k, where $\sum_{i=1}^{m} P(s_k = i | x_k) = 1$. The modal ARX functions $f_i(x_k)$ are defined as in (3) with modal ARX parameters w_i again defining the columns of W. Since (5) is a probabilistic generalization of (1), $\Omega = \{m, n_a, n_b\}$ is also used here to describe the set of discrete MEARX model parameters. Ref. [8] defines $P(s_k = i | x_k, \Theta)$ via the *softmax function*,

$$P(s_k = i | x_k, \Theta) = \frac{\exp\left(\theta_i^T x_k\right)}{\sum_{j=1}^{m} \exp\left(\theta_j^T x_k\right)}, \tag{6}$$

where each softmax weight $\theta_i \in \mathbb{R}^{(r+1)\times 1}$ and the 'probabilistic boundaries' between the discrete modes are linear hyperplanes [11]. This last property enables straightforward transformation of a softmax-based MEARX model into a PWARX model with the same Ω. Specifically, $\mathbf{H}$ for the corresponding PWARX model can be formed from Θ by setting each $H_i \in \mathbf{H}$ as

$$H_i = [\theta_1 - \theta_i, \ \ldots \ , \theta_m - \theta_i]^T \ \text{(excluding } \theta_i - \theta_i). \tag{7}$$

Following this transformation, the modal ARX parameters $w_i \in W$ for the PWARX model are set equal to the corresponding $w_i \in W$ in the MEARX model, $\forall\ i \in \{1, ...m\}$. Ref. [8] proves that this transformation from MEARX to PWARX guarantees a complete modal partition of $\mathcal{V}$, and shows (with real and synthetic data) that accurate PWARX models are readily obtained from MEARX models that are suitably identified in terms of Ω, W, and Θ.

Direct continuous optimization methods using nonlinear least-squares (NLS) [8] or maximum likelihood (ML) [9] can be used to estimate W and Θ in MEARX models with fixed Ω. This enables simultaneous estimation of $\mathbf{H}$ and W, in contrast to most other PWARX ID methods which estimate $\mathbf{H}$ and W separately. However, NLS and ML are both sensitive to outliers and thus prone to severe overfitting [12]. For model selection, [8] compares various Ω settings using separate training and validation data to find the model structure with lowest PWARX prediction error. However, this approach does not use all available information to estimate Θ and W, and can therefore be wasteful if N is small. This approach can also be time-consuming, since multiple training-validation splits must be used for each candidate Ω to obtain unbiased error estimates.

III. Fully Bayesian Identification

A Bayesian learning approach can be used to address these issues for MEARX/PWARX ID. As discussed in [6], [7], the Bayesian system ID approach is especially useful when N is limited, since overfitting can be easily avoided and suitable statistical confidence bounds are readily obtained. Above all, the Bayesian approach leads to a principled data-driven model selection procedure via the 'Bayesian Ockham's razor' [11], which uses Bayes' rule in the space of model structures to identify the simplest model that best explains the data. To the authors' knowledge, Bayesian model selection has not yet been applied to hybrid system identification problems.

A. Model for Bayesian MEARX ID

For fixed $\Omega = \{m, n_a, n_b\}$ and corresponding regression inputs x_k for $k \in \{1, ..., N\}$, consider the following variables and their prior conditional probability distributions:

- $W = [w_1, ..., w_m]$ with $p(W|\vec{\alpha}) = \prod_{i=1}^{m} p(w_i|\alpha_i)$, where $p(w_i|\alpha_i) = \mathcal{N}_{w_i}(0, \alpha_i^{-1} I)$ is a zero mean Gaussian distribution with unknown inverse variance α_i,
- the inverse variances $\vec{\alpha} = [\alpha_1, ..., \alpha_m]$ with $p(\vec{\alpha}) = \prod_{i=1}^{m} p(\alpha_i; a_0, b_0)$, where $p(\alpha_i; a_0, b_0) = \mathcal{G}_{\alpha_i}(a_0, b_0)$ is a Gamma distribution with known shape and scale parameters a_0 and b_0,
- $\Theta = [\theta_1, ..., \theta_m]$ with $p(\Theta|\vec{\beta}) = \prod_{j=1}^{m} p(\theta_j|\beta_j)$, where $p(\theta_j|\beta_j) = \mathcal{N}_{\theta_j}(0, \beta_j^{-1} I)$ with unknown inverse variance β_j,
- inverse variances $\vec{\beta} = [\beta_1, ..., \beta_m]$ with $p(\vec{\beta}) = \prod_{j=1}^{m} p(\beta_j; c_0, d_0)$, where $p(\beta_j; c_0, d_0) = \mathcal{G}_{\beta_j}(c_0, d_0)$ is a Gamma distribution with known c_0 and d_0,
- discrete mode s_k for output y_k, where $P(s_k = j | x_k, \Theta)$ is the softmax distribution (6),
- observed output y_k at time k with *mode-conditional* pdf $p(y_k | x_k, s_k = j) = \mathcal{N}(f_j(x_k), \tau^{-1})$, where $f_j(x_k) = w_j^T x_k$ and the unknown inverse noise variance τ has the Gamma pdf $p(\tau|h_0, l_0) = \mathcal{G}_\tau(h_0, l_0)$ with known h_0 and l_0.

The pdf $p(y_k | x_k, s_k = j)$ arises if e_k is zero-mean Gaussian noise in (5), which the Central Limit Theorem justifies. The

Gaussian priors on w_i encode the belief of a smooth spectrum for each modal ARX process [12]. The α_i and β_i variables protect against any overfitting that commonly occurs with sparse data; they are treated as unobserved variables since their best values are unknown.

Next, define $S = \{s_1, ..., s_N\}$ and the unobserved $m \times N$ binary label matrix T with elements t_{jk}, where $t_{jk} = 1$ if $s_k = j$ and $t_{jk} = 0$ otherwise. Define $\Xi = [\vec{\alpha}, \vec{\beta}, \tau, W, \Theta, S]$ to be the set of all unknown variables. The joint pdf for $[\Xi, \mathbf{Y}]$ given $\mathbf{U}$ can be shown to be

$$p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega; a_0, b_0, c_0, d_0, h_0, l_0) = p(\tau|h_0, l_0) \prod_{i=1}^{m} p(\alpha_i; a_0, b_0) p(w_i|\alpha_i) p(\beta_i; c_0, d_0) p(\theta_i|\beta_i) \times \prod_{k=1}^{N} \prod_{j=1}^{m} [p(y_k|x_k, s_k = j) \cdot P(s_k = j|x_k, \Theta)]^{t_{jk}} \quad (8)$$

B. Bayesian parameter and model inference

The proposed MEARX ID approach has two applications of Bayes' rule: (i) parameter estimation for fixed Ω, and (ii) model selection over Ω. For (i), we require the posterior parameter pdf $p(\Xi|\mathbf{Y}, \mathbf{U}, \Omega)$. When the known output observations $\mathbf{Y}$ are taken into account, applying Bayes' rule to (8) yields

$$p(\Xi|\mathbf{Y}, \mathbf{U}, \Omega) = \frac{p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)}{\int p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega) d\Xi} = \frac{p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)}{p(\mathbf{Y}|\mathbf{U}, \Omega)}, \quad (9)$$

where $\int (\cdot) d\Xi$ denotes marginalization (summation/integration) over all possible states of Ξ and the Gamma distribution constants are suppressed for convenience. For (ii), the results of (i) are used to find the model posterior $P(\Omega|\mathbf{Y}, \mathbf{U})$. If a prior $P(\Omega)$ is assumed for each member of a finite set of unique MEARX model structures $\mathcal{M}$, then applying Bayes' rule again yields

$$P(\Omega|\mathbf{Y}, \mathbf{U}) = \frac{p(\Omega, \mathbf{Y}|\mathbf{U})}{\sum_{\mathcal{M}} p(\Omega, \mathbf{Y}|\mathbf{U})} = \frac{P(\Omega) p(\mathbf{Y}|\mathbf{U}, \Omega)}{\sum_{\mathcal{M}} P(\Omega) p(\mathbf{Y}|\mathbf{U}, \Omega)}. \quad (10)$$

As [11] shows, (10) enforces the 'Ockham's razor' principle over $\mathcal{M}$ via the *model likelihood* term $p(\mathbf{Y}|\mathbf{U}, \Omega)$ in the denominator of (9), which is naturally the larger for the simpler models in $\mathcal{M}$ that explain the observed data well. Hence, for a given $P(\Omega)$ (e.g. uniform, as is typical), the posterior (10) can be used to select the best model structure $\Omega^* \in \mathcal{M}$ using a MAP estimate, while the corresponding parameter posterior (9) can be used to find point estimates of W and Θ, e.g. using the MAP or MMSE criterion. Note that if the true generating model for the data is in $\mathcal{M}$, it will on average have the highest value for (10) [11].

However, the main challenge here is the marginalization of (8) to get $p(\mathbf{Y}|\mathbf{U}, \Omega)$ in (9) and (10), since the complex conditional dependencies between the elements of Ξ (given $\mathbf{Y}$ and $\mathbf{U}$) make the required integrations/summations analytically intractable. The next section shows how (9) and (10) can be efficiently approximated via a new variational Bayes (VB) learning approximation for MEARX models.

IV. Variational Bayes Learning

In VB learning, the analytically intractable posterior $p \equiv p(\Xi|\mathbf{Y}, \mathbf{U}, \mathbf{\Omega})$ is approximated by an analytically tractable distribution $q \equiv q(\Xi|\mathbf{Y}, \mathbf{U}, \Omega)$ that minimizes the Kullback-Leibler divergence (KLD) functional,

$$\mathrm{KL}(q||p) = \int q \log \frac{q}{p} d\Xi, \quad (11)$$

where $\mathrm{KL}(q||p) \geq 0$ and $\mathrm{KL}(q||p) = 0$ iff $p = q$ [11]. The KLD acts as an information-theoretic similarity measure between two pdfs, and can therefore be used to find a good 'free-form' probabilistic approximation to p, where q is restricted to a pdf family having some desired conditional independence properties for the elements of Ξ. Since p is unknown, (11) is minimized by maximizing a lower bound $\mathcal{L}$ to the constant model log-likelihood $\log p(\mathbf{Y}|\mathbf{U}, \Omega)$, where [11] shows

$$\log p(\mathbf{Y}|\mathbf{U}, \Omega) = \mathcal{L} + \mathrm{KL}(q||p), \quad (12)$$

$$\text{for} \quad \mathcal{L} = \int q(\Xi|\mathbf{Y}, \mathbf{U}, \mathbf{\Omega}) \log \frac{p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)}{q(\Xi|\mathbf{Y}, \mathbf{U}, \mathbf{\Omega})} d\Xi, \quad (13)$$

and $\mathcal{L} \leq \log p(\mathbf{Y}|\mathbf{U}, \Omega)$ follows from (12) and the non-negativity of (11). Hence, q can be obtained by analytically maximizing $\mathcal{L}$, which is defined via (8). Moreover, $\mathcal{L}$ can be used to approximate $\log p(\mathbf{Y}|\mathbf{U}, \Omega)$ in (9) and (10).

For MEARX learning, q is specified here via the typical 'mean field' approximation, where

$$q = q(\tau) \prod_{i=1}^{m} q(\alpha_i) q(w_i) q(\beta_i) q(\theta_i) \prod_{k=1}^{N} \prod_{j=1}^{m} q(S_k = j) = \prod_{z=1}^{V} q(\psi_z), \quad \text{for} \quad \psi_z \in \Xi, \quad (14)$$

and where $V = 1 + 4m + Nm$ denotes the total number of hidden variables in Ξ. Note that (14) implies conditional independence among all $\psi_z \in \Xi$. Ref. [11] shows that this choice of q leads to the following general formula for finding each factor $q(\psi_z)$ that maximizes (13),

$$\log q(\psi_z) = \mathbb{E}\left[\log p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)\right]_{q(\Xi) \setminus q(\psi_z)} + \text{const.}, \quad (15)$$

where the right-hand side denotes the expected value of the log joint pdf (8) with respect to all factors in (14) *excluding* $q(\psi_z)$, which is the factor of interest on the left-hand side. This leads to a coupled set of nonlinear equations for each $\log q(\psi_z)$ that can be solved iteratively in closed form via a generalization of the expectation-maximization (EM) algorithm, so that the resulting cyclic VB updates are often referred to as the VBEM algorithm [12]. As such, $\mathcal{L}$ is guaranteed to increase monotonically on each VB update cycle (i.e. each single pass through all elements of Ξ) until q converges to a local minimum of (11).

A. Lower bound softmax approximation

The VBEM algorithm via (15) entails integration of the log of (8) with respect to Θ, which cannot be done in closed form since the log of the denominator of (6) is not analytically

integrable. To avoid computationally expensive sampling-based or quadrature-based solutions to this problem, an analytical lower bound approximation to (6) is used instead. Ref. [13] shows that the softmax function (6) defining $p(s_k = i|x_k, \Theta)$ can be approximated by an unnormalized Gaussian function $f(s_k = i, x_k, \Theta)$,

$$f(s_k = i, x_k, \Theta) = \exp\left(\theta_i^T x_k\right) \exp(-\Phi_k) \tag{16}$$

$$\Phi_k = \gamma_k + \sum_{j=1}^{m} \frac{\theta_j^T x_k - \gamma_k - \xi_{jk}}{2} + \lambda(\xi_{jk})[(\theta_j^T x_k - \gamma_k)^2 - \xi_{jk}^2] + \log(1 + e^{\xi_{jk}}), \tag{17}$$

$$\lambda(\xi_{jk}) = \frac{1}{2\xi_{jk}} \left[\frac{1}{1 + \exp(-\xi_{jk})} - \frac{1}{2}\right], \tag{18}$$

where $f(s_k = i, x_k, \Theta) \leq P(s_k = i|x_k, \Theta)$ and the variational shape/scale parameters ξ_{jk} and γ_k control the tightness of the bound for $j \in \{1, ..., m\}$ and $k \in \{1, ..., N\}$.

Substituting (16) for (6) in (8) induces a lower bound $\tilde{p}(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)$ on the joint pdf, where $\tilde{p}(\Xi, \mathbf{Y}|\mathbf{U}, \Omega) \leq p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)$. Assuming the parameters ξ and γ are known (as described next, these can be estimated in a modified VBEM algorithm), it is easy to show that replacing $p(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)$ with $\tilde{p}(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)$ in (13) therefore also induces a new model log-likelihood lower bound $\tilde{\mathcal{L}}$, where $\tilde{\mathcal{L}} \leq \mathcal{L}$. Thus, eq. (15) becomes

$$\log q(\psi_z) = \mathbb{E}\left[\log \tilde{p}(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)\right]_{q(\Xi)\setminus q(\psi_z)} + \text{const.}. \tag{19}$$

B. Explicit VB update formulas

After substituting the required distributions into (19) and simplifying the resulting terms for each $\log q(\psi_z)$, it can be shown that the following VB factors $q(\psi_z) = \exp[\log q(\psi_z)]$ are obtained:

- $q(w_i) = \mathcal{N}_{w_i}(\mu_{w_i}, \Sigma_{w_i})$, where $\mathbb{E}[w_i] = \mu_{w_i}$ and

$$\Sigma_{w_i}^{-1} = \mathbb{E}[\alpha_i] \cdot I + \mathbb{E}[\tau] \cdot \sum_{k=1}^{N} \mathbb{E}[t_{ks}] x_k x_k^T \tag{20}$$

$$\mu_{w_i} = \Sigma_{w_i} \cdot \left(\mathbb{E}[\tau] \cdot \sum_{k=1}^{N} \mathbb{E}[t_{ik}] y_k x_k\right) \tag{21}$$

- $q(\theta_i) = \mathcal{N}_{\theta_i}(\mu_{\theta_i}, \Sigma_{\theta_i})$, where $\mathbb{E}[\theta_i] = \mu_{\theta_i}$ and

$$\Sigma_{\theta_i}^{-1} = \mathbb{E}[\beta_i] \cdot I + \sum_{k=1}^{N} 2\lambda(\xi_{sk}) x_k x_k^T, \tag{22}$$

$$\mu_{\theta_i} = \Sigma_{\theta_i} \cdot \left(\sum_{k=1}^{N} \left[\mathbb{E}[t_{ik}] - \frac{1}{2} + 2\gamma_k \lambda(\xi_{ik})\right] x_k\right) \tag{23}$$

- $\mathbb{E}[t_{ik}] = q(s_k = i) = \frac{\exp(\eta_i^T)}{\sum_{j=1}^{m} \exp(\eta_j)}$, where

$$\eta_i = \mathbb{E}[\theta_i]^T x_k - \frac{1}{2}\mathbb{E}[\tau] \cdot (y_k^2 + x_k^T \mathbb{E}[w_i w_i^T] x_k - 2y_k \mathbb{E}[w_i]^T x_k), \tag{24}$$

for $\mathbb{E}[w_i w_i^T] = \Sigma_{w_i} + \mu_{w_i} \mu_{w_i}^T$

- $q(\alpha_i) = \mathcal{G}_{\alpha_i}(a_i, b_i)$, where $\mathbb{E}[\alpha_i] = \frac{a_i}{b_i}$ and

$$a_i = a_0 + \frac{n+1}{2}, \tag{25}$$

$$b_i = b_0 + \frac{1}{2}\mathbb{E}\left[w_i^T w_i\right] = b_0 + \frac{1}{2}\left[\text{tr}\left(\Sigma_{w_i} + \mu_{w_i}^T \mu_{w_i}\right)\right] \tag{26}$$

- $q(\beta_i) = \mathcal{G}_{\beta_i}(c_i, d_i)$, where $\mathbb{E}[\alpha_i] = \frac{c_i}{d_i}$ and

$$c_i = c_0 + \frac{n+1}{2}, \tag{27}$$

$$d_i = d_0 + \frac{1}{2}\mathbb{E}\left[\theta_i^T \theta_i\right] = d_0 + \frac{1}{2}\left[\text{tr}\left(\Sigma_{\theta_i} + \mu_{\theta_i}^T \mu_{\theta_i}\right)\right] \tag{28}$$

- $q(\tau) = \mathcal{G}_{\beta_i}(h, l)$, where $\mathbb{E}[\tau] = \frac{h}{l}$ and

$$h = h_0 + \frac{N}{2}, \; l = l_0 + \frac{1}{2}(1 + \rho), \tag{29}$$

$$\rho = \sum_{k=1}^{N} \sum_{i=1}^{m} \mathbb{E}[t_{ik}] \left(y_k^2 + x_k^T \mathbb{E}[w_i w_i^T] x_k - 2y_k \mathbb{E}[w_i]^T x_k\right)$$

Ref. [13] shows that ξ_{jk} and γ_k are generally estimated as

$$\xi_{ik}^2 = x_k^T \mathbb{E}[\theta_i \theta_i^T] x_k + \gamma_k^2 - 2\gamma_k \mu_{\theta_i}^T x_k \tag{30}$$

$$\gamma_k = \frac{\frac{1}{2}(\frac{m}{2} - 1) + \sum_{i=1}^{m} \lambda(\xi_{ik}) \mu_{\theta_i}^T x_k}{\sum_{i=1}^{m} \lambda(\xi_{ik})} \tag{31}$$

Although not provided explicitly here due to limited space, it can be shown that the variational lower bound $\tilde{\mathcal{L}}$ to $p(\mathbf{Y}|\mathbf{U}, \Omega)$ can be computed in closed form via

$$\tilde{\mathcal{L}} = \mathbb{E}\left[\tilde{p}(\Xi, \mathbf{Y}|\mathbf{U}, \Omega)\right]_{q(\Xi|\mathbf{Y}, \mathbf{U}, \Omega)}, \tag{32}$$

where most of the required terms for the resulting expression are already computed in (20)-(31) on each VB update cycle.

Table I summarizes the iterative VBEM algorithm for fixed Ω. The VBEM algorithm is guaranteed to monotonically increase the value of $\tilde{\mathcal{L}}$ following steps 1 and 2, so $\tilde{\mathcal{L}}$ can be used to gauge convergence. Note that γ_k and ξ_{jk} are non-linearly coupled, so an extra inner-loop of n_{lc} steps is needed for convex convergence of these parameters with all $q(\theta_i)$ fixed ($n_{lc} \leq 15$ is often sufficent if $\mathbf{Y}$ is normalized). Since the VBEM algorithm converges to local KLD minimizers in $q(\Xi)$ [11], multiple initializations should be used to ensure convergence to the best solution.

C. VB model selection

To select the best $\Omega^* = \{m^*, n_a^*, n_b^*\}$ from a finite set of candidate models $\mathcal{M}$ (e.g. which can be constructed by specifying upper and lower bounds for the unknown elements of Ω), the VBEM algorithm in Table I can be applied to each candidate $\Omega_c \in \mathcal{M}$ to estimate a corresponding approximation $\tilde{\mathcal{L}}_{\Omega_c}$ to $\log p(\mathbf{Y}|\mathbf{U}, \Omega_c)$. Assuming no prior preference for any of the models in $\mathcal{M}$, $P(\Omega_c)$ can be set to the uniform distribution in (10), so that $\log P(\Omega_c|\mathbf{Y}, \mathbf{U}) \propto \log p(\mathbf{Y}|\mathbf{U}, \Omega_c) \approx \tilde{\mathcal{L}}_{\Omega_c}$. Hence, a MAP estimate of Ω^* can be found by choosing the $\Omega_c \in \mathcal{M}$ with the largest $\tilde{\mathcal{L}}_{\Omega_c}$. To obtain a PWARX model, one can then apply the transformation method of Section II-B to the Ω^* MEARX model using the MAP W and Θ estimates, which become $[\mu_{w_1^*}, ..., \mu_{w_{m^*}}]$ and $[\mu_{\theta_{1^*}}, ..., \mu_{\theta_{m^*}}]$, respectively.

TABLE I
VBEM ALGORITHM FOR FIXED Ω MEARX MODEL

0. *Given*: $\mathbf{Y}$, $\mathbf{U}$, fixed Ω with initial $\mu_{\theta_j}, \Sigma_{\theta_j}$, γ_k for $j \in \{1, ...m\}$ and $k \in \{1, ..., N\}$, and prior parameters $a_0, b_0, c_0, d_0, h_0, l_0$
1. **M-step**: for all $q(\theta_j)$ fixed,
   for i=1:n_{lc}
   (a) recompute each ξ_{jk} for fixed γ_k via (30)
   (b) recompute each γ_k for all fixed ξ_{jk} via (31)
   end
2. **E-step**: for all $j \in \{1, ...m\}$ and $k \in \{1, ..., N\}$ and for all fixed ξ_{jk} and γ_k,
   (a) recompute each $q(w_j)$ via (20)-(21)
   (b) recompute each $q(\theta_j)$ via (22)-(23)
   (c) recompute each $q(S_k = j)$ via (24)
   (d) recompute each $q(\alpha_j)$ via (25)-(26)
   (e) recompute each $q(\beta_j)$ via (27)-(28)
   (f) recompute $q(\tau)$ via (29)
3. Stop if $\tilde{\mathcal{L}}$ converged; else, Repeat 1 and 2.

V. NUMERICAL EXPERIMENTS

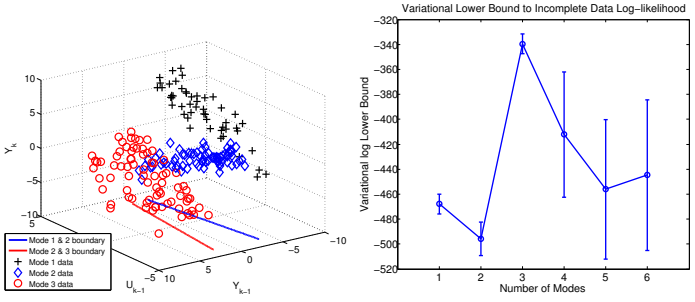
This section demonstrates the proposed VB learning procedure for MEARX-based PWARX ID using two PWARX hybrid systems taken from [10].

A. Example 1: unknown m with known n_a and n_b

Consider example 1 of [10], where the true PWARX hybrid system is given by the following W and $\mathbf{H}$ parameters for $\Omega^{\text{true}} = \{3, 1, 1\}$,

$$w_1 = [-0.4, 1, 1.5]^T,\ H_1 = \begin{bmatrix} 4 & -1 & 10 \\ 1 & 0 & 4/9 \end{bmatrix},$$
$$w_2 = [0.5, -1, -0.5]^T,\ H_2 = \begin{bmatrix} -4 & 1 & -10 \\ 5 & 1 & -6 \end{bmatrix},$$
$$w_3 = [-0.3, 0.5, -1.7]^T,\ H_3 = \begin{bmatrix} -5 & -1 & 6 \\ -1 & 0 & -4/9 \end{bmatrix},$$

Note that for this system, $v_k = [y_{k-1}, u_{k-1}]^T$ and $x_k = [v_k^T, 1]^T$ as per (3), where $u_k \sim \mathcal{U}[-4, 4]$. Figure 1 (a) shows a typical data 200 point data set for this system with two of the switching boundaries (red and blue lines). Ref. [10] identifies this system with $N = 200$ data points corrupted by bounded noise $e_k \in [-0.2, 0.2]$, where $n_a = 1$ and $n_b = 1$ are known a priori but m is unknown. The same assumptions



(a) (b)

Fig. 1. Learning results for example 1: (a) typical $N = 200$ data set, (b) estimated VB lower bounds to data log-likelihoods vs. m for $\Omega = \{m, 1, 1\}$, with clear maximum at $m = 3$ (mean and standard deviations shown).

are used here for VB MEARX learning, except that the problem is made slightly harder by assuming e_k is distributed as $\mathcal{N}(0, 0.75)$ to make the noise magnitude about four times larger. Unlike [10], the inverse noise variance $\tau = (\sigma^2)^{-1}$ is treated here as an additional estimated parameter with a continuous prior over all non-negative τ. The gamma prior scale and shape variables for the Bayesian MEARX model are all set to $a_0 = b_0 = c_0 = d_0 = h_0 = l_0 = 1$.

VB MEARX learning was applied to 20 randomly generated data sets, where $\mathcal{M} = \{\Omega \mid m \in [1, 6], n_a = 1, n_b = 1\}$. Figure 1 (b) plots the resulting $\tilde{\mathcal{L}}$ statistics for each candidate $\Omega \in \mathcal{M}$, shwoing that clear peaks are obtained at the correct number of modes $m = 3$ in all cases. The Ockham's razor effect for Bayesian model comparison is evident in Figure 1 (b): overly simple models are penalized for fitting the data poorly, while overly complex models that can fit the data as well as the true model (e.g. $m = 4$ and above) are penalized through having larger parameter spaces over which to distribute probability mass. The resulting $m = 3$ estimates for W and $\mathbf{H}$ are also accurate, e.g. the following w_i estimates are obtained with the data shown in Fig. 1,

$$\hat{w}_1 = [-0.3227, 0.9760, 2.1844]^T,$$
$$\hat{w}_2 = [0.5170, -1.0379, -0.4755]^T,$$
$$\hat{w}_3 = [-0.3219, 0.5137, -1.4665]^T,$$

More precise estimation quality measures can be obtained via the metrics proposed by Juloski, et al. [5]. The average relative weight errors $\epsilon_i = \frac{\|\hat{w}_i - w_i\|}{\|w_i\|}$ for each estimated $\hat{w}_i$ are $\epsilon_1 = 0.2124$, $\epsilon_2 = 0.1238$, and $\epsilon_3 = 0.1043$; the average number of mode misclassifications is 2.1 out of 200, with a maximum of 3 out of 200 across all trials. Hence, despite the higher noise levels imposed here, the parameter estimates obtained via VB are still quite good. Note that similar learning results were found for various other choices of the gamma prior scale and shape variables. The procedure in Table I required only 4 secs to run in Matlab for $m = 3$, and each sweep from $m = 1$ to 6 required about 78 secs.

B. Example 2: completely unknown Ω with sparse/noisy data

A more challenging PWARX ID problem from [10] considers $\Omega^{\text{true}} = \{4, 2, 2\}$ and noise bounded between $[-0.25, 0.25]$ for $N = 1000$ data points. The true columns for W are given by

$$w_1 = [-0.05, 0.76, 1.00, 0.50, -0.50]^T,$$
$$w_2 = [1.21, -0.49, -0.30, 0.90, 0]^T,$$
$$w_3 = [1.49, -0.50, 0.20, -0.45, -1.70]^T,$$
$$w_4 = [-1.20, -0.72, 0.60, -0.70, 2.000]^T. \quad (33)$$

As [10] notes, this problem is challenging since the number of parameters to be estimated for the PWARX model is relatively large for the amount data provided. However, while [10] obtains accurate parameter estimates without knowing m a priori, it was assumed that n_a and n_b were known a priori and that e_k was hard-bounded.

For VB MEARX learning, Ω is *completely* unknown, as is τ. Furthermore, four different sparsity-noise conditions are

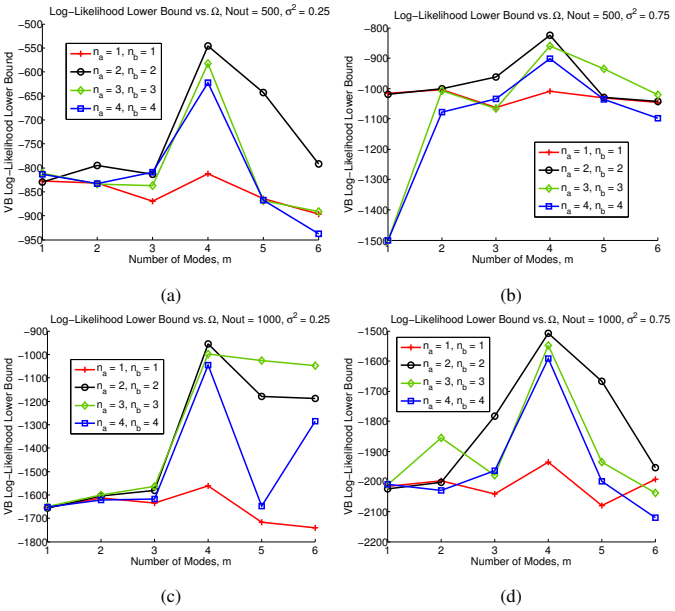


Fig. 2. $\tilde{\mathcal{L}}$ results for example 2 for different sparsity and noise levels for learning, showing clear peaks at $\Omega = \{4, 2, 2\}$ in all cases (results for $n_a = n_b$ shown only for clarity): (a) $N = 500$, $\sigma^2 = 0.25$, (b) $N = 500$, $\sigma^2 = 0.75$, (c) $N = 1000$, $\sigma^2 = 0.25$, (d) $N = 1000$, $\sigma^2 = 0.75$.

imposed by letting $\sigma^2 \in \{0.25, 0.75\}$ and $N \in \{500, 1000\}$. Hence, the problem becomes even more challenging, since Ω is unknown, the noise level can be greatly increased, and the amount of data can be reduced by half. The input $u_k \sim \mathcal{U}[-0.5, 0.5]$ and the true $\mathbf{H}$ is given by $H_1 = [\delta_2, 0, -(\delta_1 + \delta_3), 0, 0], H_2 = [\delta_1 + \delta_2, 0, \delta_1, 0, 0], H_3 = -H_1$, and $H_4 = -H_2$, where δ_i is a unit vector in $\mathbb{R}^{3\times 1}$. The number of data points per mode (N_1, N_2, N_3, N_4) for each (N, σ^2) pairing are: (186,184,67,63) for (500, 0.25); (169,161,90,80) for (500,0.75); (358,379,131,132) for (1000,0.25); and (328,337,173,162) for (1000,0.75). Note that modes 3 and 4 are not as well-observed as modes 1 and 2. All gamma shape and scale parameters are again set to 1 in each N and σ^2 cases. For model selection in each (N, σ^2) case, $\mathcal{M} = \{\Omega \mid m \in [1, 6], n_a \in [1, 4], n_b \in [1, 4]\}$. Figure 2 shows the final $\tilde{\mathcal{L}}$ values from VB MEARX learning obtained for the different candidate Ωs in each (N, σ^2) example; note that only symmetric lag models $n_a = n_b$ are shown for clarity. These plots clearly show that $\tilde{\mathcal{L}}$ peaks at Ω^{true} even with increased noise and greatly reduced data levels. Hence, even under challenging learning conditions, $\tilde{\mathcal{L}}$ still provides a reliable metric for model selection.

The relative weight error metrics $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$, classification error metrics $N_{\text{missclass}}/N$, and estimated noise error at Ω^{true} in each (N, σ^2) case are as follows: (0.0526, 0.1403, 0.2307, 0.2580), 3/500, and 0.0236 for (500,0.25); (0.0924, 0.3194, 0.1078, 0.4079), 13/500, and 0.0781 for (500,0.75); (0.0494, 0.0568, 0.1531, 0.1782), 8/1000 and 0.0167 for (1000,0.25); and (0.2098 0.1758 0.0873 0.1998), 7/1000 and 0.0483 for (1000,0.75). Thus, the PWARX estimates produced by the VB-MEARX learning procedure are reasonable for $N = 1000$ under both noise conditions, while the parameter estimates for $N = 500$ degrade noticeably when moving from the low noise condition to the high noise condition. Note that most of the errors in $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ occur in the w_3 and w_4 coefficients for the u_{k-2}. This is unsurprising, since modes 3 and 4 are not well-observed in the data and since u_k is sampled over a very narrow range, making these particular parameters more difficult to estimate. Despite this, the VB model selection procedure is still able to correctly determine all elements of Ω via $\tilde{\mathcal{L}}$.

VI. Conclusions and Future Work

This paper presented a variational Bayesian (VB) technique for MEARX learning models. The development here was motivated by the close connection between MEARX ID and PWARX hybrid system ID, for which the issues of model selection and learning with sparse/noisy data remain problematic. The VB learning procedure enables efficient parameter estimation and model selection over an unknown discrete number of modes, unknown ARX lag orders and unknown process noise magnitude. Numerical examples demonstrated the ability of the proposed method to produce useful parameter and model structure estimates under sparse and noisy data conditions. Future work will focus on improving the mean field approximation in (14), conducting formal comparisons of the VB learning method to other PWARX ID methods, and extending VB learning to full hybrid linear system identification.

References

[1] E. Cinquemani, R. Porreca, G. Ferrari-Trecate, and J. Lygeros, "Parameter identification for stochastic hybrid models of biological interaction networks," in *46th IEEE Conf. on Decision and Control*, 2007.

[2] C. Feng, C. Lagoa, and M. Sznaier, "Hybrid system identification via sparse polynomial optimization," in *ACC 2010*, pp. 160 – 165.

[3] W. Heemels, B. De Schutter, and A. Bemporad, "Equivalence of hybrid dynamical models," *Automatica*, vol. 37, no. 7, pp. 1085–1091, 2001.

[4] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems: a tutorial," *European J. of Control*, vol. 13, no. 2-3, pp. 242–260, 2007.

[5] A. Juloski, W. Heemels, G. Ferrari-Trecate, R. Vidal, S. Paoletti, and J. Niessen, "Comparison of four procedures for the identification of hybrid systems," in *HSCC 2005*.

[6] A. Juloski, S. Weiland, and W. Heemels, "A Bayesian approach to identification of hybrid systems," *IEEE Trans. on Auto. Control*, vol. 50, no. 10, pp. 1520–1533, 2005.

[7] N. Hudson and J. Burdick, "A stochastic framework for hybrid system identification with application to neurophysiological systems," in *HSCC 2007*.

[8] S. Taguchi, T. Suzuki, S. Hayakawa, and S. Inagaki, "Identification of probability weighted multiple ARX models and its application to behavior analysis," in *48th IEEE Conf. on Decision and Control*, 2009.

[9] A. Carvalho and M. Tanner, "Modeling nonlinearites with mixtures-of-experts of time series models," *Int. J. of Mathematics and Mathematical Sciences*, vol. 2006, pp. 1–22, 2006.

[10] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "A bounded-error approach to piecewise affine system identification," *IEEE Trans. on Auto. Control*, vol. 50, no. 10, pp. 1567–1580, 2005.

[11] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[12] S. Roberts and W. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Trans. on Sig. Process.*, vol. 50, no. 9, pp. 2245–2257, 2002.

[13] G. Bouchard, "Efficient bounds for the softmax function and applications to approximate inference in hybrid models," in *NIPS 2007 Workshop for Approximate Bayesian Inference in Continuous/Hybrid Systems*, Whistler, BC, Canada, 2007.