8th International IFAC Symposium on
Dynamics and Control of Process Systems
Preprints Vol.3, June 6-8, 2007, Cancún, Mexico

DISTURBANCE MODELING FOR PROCESS CONTROL VIA HIDDEN MARKOV MODELS

Wee Chin Wong * **Jay H. Lee** *,1

* *School of Chemical and Biomolecular Engineering*
Georgia Institute of Technology
311 Ferst Drive Nw., Atlanta GA 30332, USA.

Abstract: This work proposes a framework for modeling disturbances that exhibit time-varying characteristics typically witnessed in process industries. These include intermittent drifts and abrupt jumps. Through examples, it is shown that the use of existing linear, stationary models is limiting. It is also demonstrated how the proposed switching model may be identified in a computationally tractable way.

Keywords: Non-stationary, Disturbance signals, Markov parameters, Switching characteristics

1. INTRODUCTION

System identification plays a vital role in chemical process control. For example, critical to the successful implementation of Model Predictive Control (MPC), the de facto standard for industrial advanced process control, is the availability of an appropriate description of plant behavior. In this context, disturbance modeling is crucial for it accounts for the effect of unmeasured signals, unmodeled plant dynamics as well as unexplainable phenomena (the residuals).

In this paper, discrete [2] time linear models with additive disturbances are the main concern. The latters' characteristics typically vary in time and reveal complex non-stationary modes. Such behavior includes intermittent drifts, abrupt jumps, and outliers, all commonly witnessed patterns in process industries. For the purpose of illustration, consider, as an approximation of reality, that depicted in Fig. (1), a time series plot of a plant's output under non-perturbed manipulated variables. It clearly exhibits dual-regime behavior. Namely, there exists periods of relatively high frequency process noise with probabilistic injections of intervals revealing mean shifts. Modeling and identification of such behavior is the focus of this paper since system identification in the presence of non-stationary noise is a relatively unexplored area for the process control community. This work is a step towards providing a formal framework for more realistic description of hitherto unquantifiable signals.

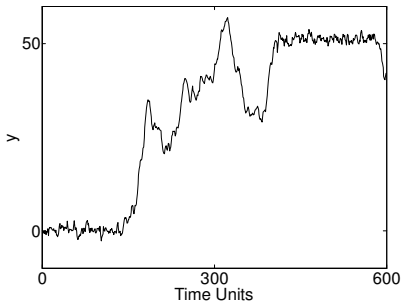

Fig. 1. White noise probabilistically interspersed with integrated white noise.

[1] The authors gratefully acknowledge the financial support from Honeywell.
[2] Single time-unit sampling intervals are assumed

1.1 Classical approaches

Faced with the aforementioned scenario, a control engineer tasked with system identification is likely to adopt a methodology developed for Eq.(1), the archetypal plant description. Here the nomenclature is standard. The interpretation is that the first term accounts for deterministic plant behavior whilst the second, the residuals, d_k. It is widely accepted that stochastic descriptions are natural for the latter. In system identification literature, d_k is commonly approximated as (stationary) filtered white noise.

$$y_k = G(q^{-1})u_k + \underbrace{H(q^{-1})n_k}_{d_k} \tag{1}$$

Eq. (1) is extended to allow for "random" walk type drifting behavior when n_k is integrated white noise, i.e. $n_k = \frac{e_k}{1-q^{-1}}$. Doing so allows one to take advantage of the wealth of system identification knowledge developed for systems experiencing low frequency drifts, which are common in chemical processes.

Regardless, the classical approach to identification is limiting in the case depicted in Fig. (1). One suboptimal method is to ignore portions of data that exhibit 'non-compliant' behavior. This is unsatisfactory (and impractical) in the face of limited data. Another recourse is to use all the data with existing methods, thereby implicitly accepting performance degradation (in the accompanying estimators and controllers).

1.2 An HMM approach

The above example reveals a need for addressing simultaneous discrete and continuous dynamics. In light of this, the potential use of Hidden Markov Models (HMMs) in providing a significant generalization of the current model form used for process control, is explored. A finite-state Markov chain is used for describing disturbance mode transitions. The Markov chain reflects probabilistic state transitions which depend only on the immediate past. Naturally, the characteristics of d_k is postulated to be dependent on the Markov state. The term 'Hidden' indicates that the latter is never known with certainty and must be inferred from available noisy measurements.

The resulting plant is termed a 'Markov Jump Linear System' (MJLS) (Costa *et al.*, 2005); where the general representation follows

$$\begin{aligned} x_k &= A_{r_k}x_{k-1} + B_{r_k}u_{k-1} + \omega_{r_k} \\ y_k &= C_{r_k}x_k + \upsilon_{r_k} \end{aligned} \tag{2}$$

where

r_k	Markov state at the k-th time sample;
x_k	continuous internal state;
ω_{r_k}	i.i.d, Gaussian noise, $\sim \mathcal{N}(0, Q_{r_k})$;
υ_{r_k}	i.i.d, Gaussian noise, $\sim \mathcal{N}(0, S_{r_k})$;
y_k	noisy measurements;
u_k	known, deterministic input signal;

$\{A_{r_k}, B_{r_k}, C_{r_k}, Q_{r_k}, S_{r_k}\}$ evolve according to the Markov chain realizations, $(r_1, r_2, \ldots, r_k, \ldots)$. Without loss of generality, $E(w_{r_k}v'_{r_k}) = 0, \forall k$. Also, $r_k \in \mathcal{J} \triangleq \{1, \ldots, J\}$, where $J \in \mathbb{Z}_+$, denotes the number of discrete Markov states.

The Markovian jumps are governed by a transition matrix $\Pi = (Pr(r_k = j | r_{k-1} = i) \triangleq p_{ij}) : \sum_j^J p_{ij} = 1 \forall i \in \mathcal{J}$. All Markov chains under consideration are ergodic. For simplicity, the Markov chain is assumed to be at steady state, satisfying $\pi = \Pi'\pi$, where π is a column vector containing the unconditional probabilities of each regime.

It is noted that the application of HMM's and/or MJLS's (and their variants) in science and engineering is not novel per se. Researchers in the fields of speech recognition (Rabiner, 1989), artificial intelligence (Murphy, 1998), econometrics (Kim and Nelson, 1999), automatic control (Bar-Shalom and Li, 1993) and other communities have employed MJLS's since the 1960s. The references in (Costa *et al.*, 2005) mark a good starting point for work on MJLS's done by the control community.

Encouraged by the successes in other fields, a restricted MJLS representation is considered for describing a wide class of commonly seen disturbances. This HMM approach has hitherto found limited use for disturbance modeling. Robertson and Lee (1998) proposed a special kind of Markovian jump disturbances for the modeling of abrupt changes but there, the focus was on state estimation.

This contribution is chiefly to demonstrate the viability of restricted MJLS's as a model structure suitable for a wide spectrum of interesting disturbance signals. Under consideration are two scenarios in which the use of stationary models (as might be done by practicing engineers during an identification experiment) results in unsatisfactory estimation. It is shown how the plant and Markov chain parameters may be obtained via Maximum Likelihood Estimation (MLE). This is facilitated by viewing $(r_1, \ldots, r_k, ...)$ as an unknown sequence to be estimated. Doing so addresses the fact that exact MLE requires a number of filters that grow exponentially with the data-length.

Section 2 presents the problem formulation. The corresponding solution methodology is delineated in Section 3. Numerical examples are presented

in Section 4. Section 5 concludes the work and comments on outstanding issues.

2. PROBLEM FORMULATION

The primary concern is disturbance modeling. For this, the representation is given by Eq. (3). The superscript n denotes noise (or equivalently disturbance).

$$\begin{aligned} x_k^n &= A^n x_{k-1}^n + \omega_{r_k} \\ y_k^n &= C^n x_k^n + \upsilon_{r_k} \end{aligned} \tag{3}$$

This system defines the noise part of the model (d_k in Eq. (1)).

The (general) identification problem is to find $\tilde{\theta} \triangleq vec\{A_i, C_i, Q_i, S_i, \Pi\}, \forall i \in \mathcal{J}$, given data of finite length, T. For notational ease, the following sequences are compactly represented viz $Y_1^T \triangleq (y_1, ..., y_T)$; $X_1^T \triangleq (x_1, ..., x_T)$ and $R_1^T \triangleq (r_1, ..., r_T)$. From now on superscripts are dropped wherever the contextual meaning is unambiguous. The general term 'system' is used to refer to either the plant dynamics or the dynamics of the disturbance model.

3. SOLUTION METHODOLOGY

Identification of Eq. (3) is achieved via MLE, i.e. $\hat{\theta}^* = \arg\max_\theta \{L_\theta \triangleq \log p_\theta(Y_1^T)\}$, where θ is $\tilde{\theta}$ concatenated with R_1^T. Direct maximization is generally difficult. However, a key observation is that maximizing the complete loglikelihood $\log p_\theta(X_1^T, Y_1^T)$, a quadratic form, is easier upon knowing X_1^T. The Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) formalizes this and is developed briefly in the following paragraphs. Readers are pointed to (Gibson and Ninness, 2005) for a treatment on EM (albeit for non-switching systems) with a control-theoretic bent.

Denote the conditional expectation operator, $\hat{\mathbb{E}}_{\hat{\theta}} \triangleq \mathbb{E}\{\cdot|Y_1^T, \hat{\theta}\}$, where the expectation is over all uncertainty and is evaluated based on $\hat{\theta}$ (an estimate of θ). Then, defining $\mathcal{Q}(\theta, \hat{\theta}) \triangleq \hat{\mathbb{E}}_{\hat{\theta}}\{\log p_\theta(X_1^T, Y_1^T)\}$, and $\mathcal{V}(\theta, \hat{\theta}) \triangleq \hat{\mathbb{E}}_{\hat{\theta}}\{\log p_\theta(X_1^T|Y_1^T)\}$, it follows from Eq. (4) that $L_\theta > L_{\hat{\theta}}$, if $\mathcal{Q}(\theta, \hat{\theta}) > \mathcal{Q}(\hat{\theta}, \hat{\theta})$, since $\mathcal{V}(\hat{\theta}, \hat{\theta}) - \mathcal{V}(\theta, \hat{\theta})$, being the Kullback-Leibler divergence, is non-negative.

$$L_\theta = \mathcal{Q}(\theta, \hat{\theta}) - \mathcal{V}(\theta, \hat{\theta}) \tag{4}$$

This observation motivates the EM algorithm, given by Eq. (5) and Eq. (6). Starting with initial parameter guesses, the E-and-M-steps making up the l-th iteration are

$$\text{compute} \quad \mathcal{Q}(\theta, \hat{\theta}_l) \triangleq \hat{\mathbb{E}}_{\hat{\theta}_l}\{\log p_\theta(X_1^T, Y_1^T)\} \tag{5}$$

$$\text{calculate} \quad \hat{\theta}_{l+1} = \arg\max_\theta \mathcal{Q}(\theta, \hat{\theta}_l) \tag{6}$$

The first argument of $\mathcal{Q}(\cdot, \cdot)$ indicates that the complete loglikelihood is to be evaluated based on θ, whereas the second argument is responsible for the conditional expectation. This generates smoothed state estimates via a Rauch-Tung-Striebel (RTS) smoother. The associated quantities are $\hat{\mathbb{E}}_{\hat{\theta}_l}\{x_k\}$, $\hat{\mathbb{E}}_{\hat{\theta}_l}\{x_k x_k'\}$, and $\hat{\mathbb{E}}_{\hat{\theta}_l}\{x_k x_{k-1}'\}$. Details of the EM algorithm are presented in Appendix A.

Without viewing R_1^T as parameters to be estimated, $\mathcal{Q}(\cdot, \cdot) = \hat{\mathbb{E}}_{\hat{\theta}_l}\{\log p_\theta(R_1^T, X_1^T, Y_1^T)\} = \mathbb{E}[\mathbb{E}\{\cdot|R_1^T, Y_1^T, \hat{\theta}_l\}]$. The second equality is from the property of the expectation operator. The outer operator is with respect to R_1^T and requires an exponentially increasing number of filters. Murphy (1998) approximated the second equality by removing the R_1^T dependence. See (Pavlovic *et al.*, 1999) for another approximate method for overcoming this intractability. Finally, it is noted that the EM algorithm is compatible with Maximum-A-Posteriori (MAP) estimation by viewing all parameters as random variables; prior knowledge, if any, can be incorporated since $\log p(\theta|Y_1^T) = \log p_\theta(Y_1^T) + \log p(\theta)$.

In the next section, several scenarios, where the presence of switching disturbance dynamics is evident from data, are considered. There, it is shown how the switching models may be obtained. Also the deficiencies of using a stationary approximation are highlighted.

4. EXAMPLES

Two examples are considered. For data-based process identification, as opposed to fundamental modeling, it may be more realistic to lump all switching dynamics into the residuals. This justifies the restriction to switching behavior occurring only in $\{Q_{r_k}, S_{r_k}\}$. The performance evaluation is based on the error of a p-step-ahead (output) predictor corresponding to an identified stationary model (N4SID (Ljung, 1999) identifies a state space model in innovations form) against that of one designed for an MJLS. The latter is identified from the same data set. Prediction is carried out over fresh data of length equivalent to the training data.

For switching models, the optimal non-linear filter grows exponentially (J^k) with time. As a sub-optimal predictor, the n^{th}-order Generalized Pseudo-Bayesian (GPBn) methodology (Bar-Shalom and Li, 1993) is favored. In the simulations, $n = 2$ was chosen. With this, only tra-

jectories whose last n terms differ are merged via moment matching into a single Gaussian. Let $\Delta_{k-n+1}^{k} \triangleq (r_{k-n+1}, \ldots, r_k)$ be a sequence of the n most recent discrete-state trajectories. For the sake of brevity, the following are the recursive equations for the 1-step ahead output predictions Eq. (7) and the corresponding error covariance, Eq. (8).

$$x_{k|k-1}(\Delta_{k-n+2}^{k}) = \sum_{r_{k-n+1}} x_{k|k-1}(\Delta_{k-n+1}^{k}) Pr(r_{k-n+1}|\Delta_{k-n+2}^{k}, Y_1^{k-1}) \tag{7}$$

$$P_{k|k-1}(\Delta_{k-n+2}^{k}) = \sum_{r_{k-n+1}} [\{x_{k|k-1}(\Delta_{k-n+2}^{k}) - x_{k|k-1}(\Delta_{k-n+1}^{k})\}\{\cdot\}' + P_{k|k-1}(\Delta_{k-n+1}^{k})] \cdot Pr(r_{k-n+1}|\Delta_{k-n+2}^{k}, Y_1^{k-1}) \tag{8}$$

The merging probabilities are obtained recursively (Bar-Shalom and Li, 1993) via Bayes rule. Subsequently, equations for p-step ahead predictors can be easily obtained.

4.1 Example 1*:*

The system description corresponding to Fig. (1) is

$$\begin{pmatrix} x \\ z \end{pmatrix}_k = \underbrace{\begin{pmatrix} 0.61 & 1 \\ 0 & 1 \end{pmatrix}}_{A} \begin{pmatrix} x \\ z \end{pmatrix}_{k-1} + \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}_{r_k}$$

$$y_k = \underbrace{\begin{pmatrix} 0.79 & 0 \end{pmatrix}}_{C} \begin{pmatrix} x \\ z \end{pmatrix}_{k-1} + \upsilon_k \tag{9}$$

The other parameters are $Q_1 = \mathrm{diag}[1, 10^{-5}], Q_2 = \mathrm{diag}[10^{-8}, 1], S = 10^{-9}, p_{11} = p_{22} = 0.995$, indicating that the probabilities of switching are relatively low. In reality, high switching frequencies would not be expected. Regime 1 corresponds to signals depicting stationary noise and regime 2, filtered integrated white noise. $T = 2000$ was used during the simulation.

Due to the integrating elements, N4SID was run in MATLAB$^{\mathrm{TM}}$, with time differenced data ($\Delta y_k = y_k - y_{k-1}$). The system order was automatically selected based on comparison of Hankel singular values. During the numerical experiment, the EM algorithm was initialized with 5% error distributed uniformly throughout the true R_1^T. This low level of error is justified on the basis that the generated signals show a clear demarcation of the regimes, justifying the usage of switching models in the first place. The guesses for the system matrices contained 15% (zero mean, unit-variance Gaussian) error on average, subject to the constraint of a stable initial system. The diagonals of Π_0 were initialized as $\mathrm{diag}[0.9, 0.9]$.

EM Alterations. Since the system matrices are regime-invariant, the following adjustments were made. In obtaining C, S, no distinction was made between data from regime 1 or 2. For A, the first row $\hat{A}(1,:)$ [3] was obtained from regime 2 data and $\hat{A}(2,:)$ from regime 1. This was found to be appropriate since state-estimation (E-Step) for the integrating mode z would generally be poorer for regime 1. Of further note is that any estimate of θ that increases $\mathcal{Q}(\cdot,\cdot)$ increases the likelihood. In light of this, it was found to be beneficial, at each iteration, to perform another E-Step prior to computing $\widehat{R_1^T}_{l+1}$. Prior knowledge was also incorporated when estimating Π. This is done by setting a Beta prior for $p_{ii}, \forall i$. This is to say, $Pr(p_{ii}) \propto (p_{ii})^{\alpha_{ii}\gamma T}(1-p_{ii})^{(1-\alpha_{ii})\gamma T}$, where α_{ii} represents the prior (or equivalently initial) estimate of p_{ii} and γ, the strength of the prior, relative to the observed training data. The net effect is that $\hat{p}_{ii}, \forall i \in \mathcal{J}$ is a convex combination of that obtained via MLE and the initial guess. A $20\% - 80\%$ split was selected, respectively. Table (1) summarizes the various performance indices and shows that suboptimal output prediction occurs when a stationary model is identified. The first three columns serve as references. Notice that N4SID does marginally better than the optimal predictor for integrated white noise (i.e. $\hat{y}_{k|k-p} = y_{k-p}$).

Table 1. Normalized sum of squared p-step-ahead prediction errors for ex.1 (average of 500 realizations)

p	GPB2 (actual model)	GPB2 (initial model)	Optimal predictor for integrated white noise	GPB2 (identified model)	N4SID
1	0.61	12.23	0.88	0.65	0.85
10	4.34	136.02	5.12	4.47	5.02

Further analysis revealed that N4SID, due to the assumption of stationarity, yielded performance inferior to GPB2 for both regimes. For example, the normalized 1-step-ahead output-prediction errors for regime 1 were 0.56 and 0.83 for a GPB2 predictor designed for the identified switching model and the filter corresponding to N4SID, respectively. For regime 2, the corresponding figures are 0.57 and 0.78.

[3] in Matlab notation

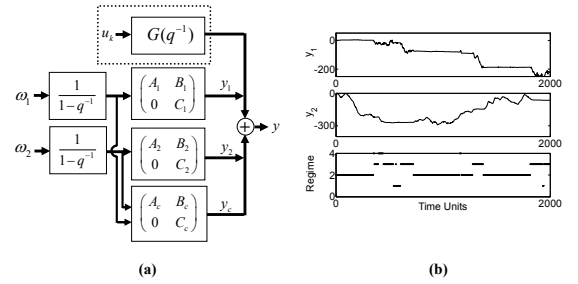

Fig. 2. (a) An inferential control setup; (b) a quad-regime scenario.

4.2 Example 2:

For the purpose of illustration, a situation where a soft-sensor approach is employed for inferential control (see the left panel of Fig. (2)) is considered. Namely, a model is constructed, so as to deduce desired but hard-to-get primary variables from the measurements of secondary ones.

In this example, Eq. (10), the secondary measurements are $y_1, y_2 \in \mathbb{R}$; $y_c \in \mathbb{R}^2$ are the primary ones. The situation reflects ω_1 and ω_2 switching between the levels 'low-low' (regime 1), 'low-hi' (regime 2), 'hi-low' (regime 3) and 'hi-hi' (regime 4), with the first and last being the most improbable ones. The chances are that either (but not both) ω_1 or ω_2 is dominant. This quad-regime scenario is depicted clearly in the right panel of Fig. (2). Implicit in this formulation is that all disturbances are added to the output channel. For simplicity, it is assumed that $u_k, \forall k$ is never perturbed and all that is observed would be disturbance patterns:

$$\begin{pmatrix} x_1 \\ x_2 \\ z_1 \\ z_2 \end{pmatrix}_k = \begin{pmatrix} 0.61 & 0 & 1 & 0 \\ 0 & 0.90 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ z_1 \\ z_2 \end{pmatrix}_{k-1} + \begin{pmatrix} 0 \\ 0 \\ \omega_1 \\ \omega_2 \end{pmatrix}_{r_k}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_{c_1} \\ y_{c_2} \end{pmatrix}_k = \begin{pmatrix} 2.0 & 0 & 0 & 0 \\ 0 & 0.95 & 0 & 0 \\ 5 & 5 & 0 & 0 \\ 7 & 7 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ z_1 \\ z_2 \end{pmatrix}_k + \upsilon_k \quad (10)$$

For self-transitions, $p_{11} = p_{44} = 0.96, p_{22} = p_{33} = 0.994$. The remaining probability is distributed uniformly over the remaining regimes, with the constraint that transitions between regimes 1 and 4 are forbidden. With this, the system is in either regime 1 or 4 for 10% of the time and in regime 2 or 3 for the other 90%. Furthermore, $Q_1 = \text{diag}[10^{-3}, 10^{-3}]$, $Q_2 = \text{diag}[10^{-3}, 1]$, $Q_3 = \text{diag}[1, 10^{-3}]$, $Q_4 = \text{diag}[1, 1]$, $S = \text{diag}[10^{-4}, 10^{-4}, 10^{-4}, 10^{-4}] \forall j$. $T = 3600$ and y_c were assumed to be available only for the sake of modeling but not for estimation after the model construction. The performance index is the p-step-ahead prediction errors of y_c.

As before, a stationary model was obtained via N4SID (without data differencing this time due to the nature of the performance index). In acquiring a switching model, the initial trajectory contained 5% error. The starting guesses for the system matrices contained 20% error on average. The diagonals of Π were $[0.90, 0.97, 0.97, 0.90]$; the other entries followed the same pattern exhibited by the true transition matrix. As is with example 1, Π was obtained via MAP estimation, assuming 20% confidence in the data and 80% confidence in the initial estimates.

The resultant models were tested on fresh data in the absence of y_c measurements. This is consistent with inferential control setups. It is noted that oftentimes, N4SID would yield an unstable predictor (i.e. the poles of $A_{N4} - K_{N4}C_{N4}$, with rows and columns corresponding to y_c removed, were outside the unit disc). A possible reason is that N4SID essentially identifies a near open-loop observer (as suggested by the previous example). The results are summarized in Table (2).

Table 2. Normalized sum of squared p-step-ahead prediction errors for ex.2 (average of 500 realizations)

p	GPB2 (actual model)	GPB2 (initial model)	GPB2 (identified)	N4SID
1	47.07	551.02	49.88	∞
10	3856	33470	3892	∞

5. CONCLUSIONS, LIMITATIONS & FUTURE RESEARCH

Below is a list of some possible concerns which reflect the fact that identification of MJLS is an open area for research.

- *Observability*: It has been shown (Vidal *et al.*, 2002) that for general MJLS's there exists an infinite number of combinations of discrete and continuous state cardinalities that will give the same input-output behavior. To mitigate this, scenarios depicting switching only in the noise parameters were chosen. Also, only scenarios where the existence of regimes was obvious were under consideration. The automatic selection of the system order is also another area of research.
- *Local Optima*: MLE optimization is typically non-convex; the likelihood surface is fraught with multiple optima. As such, the EM algorithm converges to the nearest stationary point. This was implicitly overcome by choosing suitably 'close' initial guesses. It is the authors' intentions to explore various initialization schemes. Deterministic annealing has

been proposed by Murphy (1998) as a possible avenue to explore.

- *Consistency*: It was previously pointed out (Logothetis and Krishnamurthy, 1999) that $(\hat{r_1}, \ldots, \hat{r_T})$ estimates are generally inconsistent since they are time-varying.
- *Jumps, Outliers*: Abrupt step jumps can be modeled per Example 4.1 by adding a third regime with low self-transition probability where $Q_3 = \text{diag}[10^{-5}10^2]$, say. Likewise, infrequent outliers may be captured in the same framework by superimposing an additional Markov chain that corresponds to a large covariance.
- *Future work*: It is postulated that more realistic disturbance modeling would result in better identification of plant dynamics ($G(q^{-1})$ in Eq. (1)). A possible recursive algorithm would involve Output-Error identification combined with the EM framework.

It is believed that the proposed approach is a suitable mathematical framework for describing a wide class of disturbance patterns commonly found in process industries. Also, it has been shown the adoption of stationary models oftentimes results in sub-par performance, in terms of estimation and therefore model-based control. By restricting the general class of MJLS's, and using prior knowledge where appropriate, it is hoped that a suitable balance between the expressive power of the models and computational tractability, has been achieved.

REFERENCES

Bar-Shalom, Yaakov and Xiao-Rong Li (1993). *Estimation and Tracking: Principles, Techniques, and Software*. Artech House.

Costa, O.L.V., M.D. Fragoso and R.P. Marques (2005). *Discrete-Time Markov Jump Linear Systems*. Springer.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (B)* **Vol. 39**, 1–38.

Gibson, Stuart and Brett Ninness (2005). Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica* **41**, 1667–1682.

Kim, Chang-Jin and Charles R. Nelson (1999). *State-space Models with Regime Switching*. The MIT Press.

Ljung, L. (1999). *System identification: theory for the user*. Prentice Hall PTR.

Logothetis, Andrew and Vikram Krishnamurthy (1999). Expectation maximization algorithms for map estimation of jump markov linear systems. *IEEE transactions on Signal Processsing* **47**(8), 2139–2156.

Murphy, K.P. (1998). Switching kalman filters. Technical report. DEC/Compaq Cambridge Research Labs.

Pavlovic, Vladimir, James M. Rehg, Tat-Jen Cham and Kevin Patrick Murphy (1999). A dynamic bayesian network approach to figure tracking using learned dynamic models. In: *The Proceedings of the 7th IEEE International Conference on Computer Vision*. Vol. 1. Kerkyra, Greece. pp. 94–101.

Rabiner, Lawrence R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2), 257–286.

Vidal, Rene, Alessandro Chiuso and Stefano Soatto (2002). Observability and identifiability of jump linear systems. In: *Proceedings on the 41st IEEE Conference on decision and control*. Vol. 4. Las Vegas. pp. 3614–3619.

Appendix A. DETAILS OF THE EM ALGORITHM

For brevity, we only show $\hat{A}_j, \hat{Q}_j$ computations. At the l^{th} iteration, we have $\delta(\cdot,\cdot)$ as the Kronecker delta, $N_{x,j} \triangleq \sum_{k=2}^{T} \delta(r_k, j|Y_1^T, \hat{\theta}_l)$, $\hat{\mathbb{E}}_{\hat{\theta}}\{x_{k_1} x_{k_2}{}'\} \triangleq \hat{\mathbb{E}}_{\hat{\theta}}\{x_{k_1}\}\hat{\mathbb{E}}_{\hat{\theta}}{}'\{x_{k_2}\} + \hat{\mathbb{E}}_{\hat{\theta}}\{x_{k_1} - \hat{\mathbb{E}}_{\hat{\theta}}\{x_{k_1}\}\}\{x_{k_2} - \hat{\mathbb{E}}_{\hat{\theta}}\{x_{k_2}\}\}'$

Denoting and subsquent optimization in the M-Step yields,

$$\Phi_{x,j} \triangleq \frac{1}{N_{x,j}} \sum_{k=2}^{T} \hat{\mathbb{E}}_{\hat{\theta}}\{x_k x_k^T\}\delta(r_k, j|Y_1^T, \hat{\theta}_l)$$

$$\Psi_{x,j} \triangleq \frac{1}{N_{x,j}} \sum_{k=2}^{T} \hat{\mathbb{E}}_{\hat{\theta}}\{x_k x_{k-1}^T\}\delta(r_k, j|Y_1^T, \hat{\theta}_l)$$

$$\Sigma_{x,j} \triangleq \frac{1}{N_{x,j}} \sum_{k=2}^{T} \hat{\mathbb{E}}_{\hat{\theta}}\{x_{k-1} x_{k-1}^T\}\delta(r_k, j|Y_1^T, \hat{\theta}_l)$$

$$\hat{A}_j = \Psi_{x,j}\Sigma_{x,j}^{-1}$$

$$\hat{Q}_j = \Phi_{x,j} - \Psi_{x,j}\Sigma_{x,j}^{-1}\Psi_{x,j}^T$$

The transition probability matrix is estimated by a simple counting procedure. $\widehat{R_1^T}_{l+1}$ is computed via $Pr(r_k|\hat{\omega}_1, \ldots, \hat{\omega}_T)$ and taking the corresponding mode as the Markov state estimate. For computing $\log p_\theta(X_1^T, Y_1^T)$, readers are pointed to Gibson and Ninness (2005). The extension to the switching case is straightforward with the exception that the following terms be appended.

$$-2\sum_{k=2}^{T}\sum_{i,j} \delta(r_{k-1}, i; r_k, j|Y_1^T, \hat{\theta}_l) \log p_{ij}$$

$$-2\sum_{j=1}^{J} \delta(r_1, j|Y_1^T, \hat{\theta}_l) \log \pi(r_1)$$