Segmental Semi-Markov Models for Change-Point Detection with Applications to Semiconductor Manufacturing

Technical Report UCI-ICS 00-08
Department of Information and Computer Science
University of California, Irvine

Xianping Ge, Padhraic Smyth
Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
{xge,smyth}@ics.uci.edu

March, 2000

Abstract

We formulate the problem of change-point detection in a segmental semi-Markov model framework where a change-point corresponds to state switching. The semi-Markov part of the model allows us to incorporate prior knowledge about the time of change in a Bayesian manner. The segmental part of the model allows flexible modeling of the data within individual segments, e.g., as linear, quadratic, or other regression functions. This segmental semi-Markov model is an extension of the standard hidden Markov model (HMM), from which learning and inference algorithms are extended. Results on both simulated and real data from semiconductor manufacturing illustrate the flexibility and accuracy of the proposed framework.

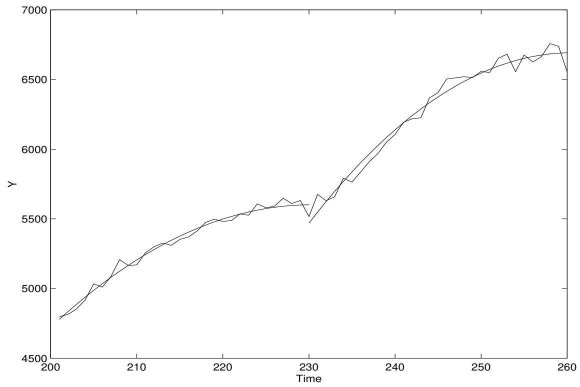

Figure 1: An illustrative example of a change-point detection problem from semiconductor manufacturing.

1 Introduction

An important problem in time-series analysis is that of *change-point detection*, namely, detecting when a time-series has "switched" in some manner. We consider our time-series to be composed of *segments* between which are *change-points*. The simplest case (and the one we focus on primarily in this paper) is that of two segments and a single change-point. As an example of a real change detection problem, Figure 1 displays a sensor channel from an industrial plasma etch process in semiconductor manufacturing. Two quadratic segments have been fitted to the data based on a manual subjective estimate of a change-point in the process. The objective is to be able to detect this change-point automatically. (We will return to this application further in Section 5). In this paper we focus on an off-line formulation of the detection problem. Future work will extend the approach to an online detector.

There is a long history of work on change detection in statistics and engineering (Basseville and Nikiforov, 1993; Lai, 1995). Many studies assume that within each segment, the distribution of individual data points does not depend on time. The simplest (and most widely used) assumption is the piecewise constant model in which each segment is a constant mean with Gaussian noise. The best known algorithm for this kind of problem is the classic CUSUM method (Page, 1954).

Of more relevance to the type of data in Figure 1 is the work on piecewise regression (Hawkins, 1976; Gustafsson, 1996), also called segmented regression (Lerman, 1980; Esterby and El-Shaarawi, 1981), or multi-phase regression (Hinkley, 1971). When the number of segments is known a priori, these techniques can be viewed simply as trying to minimize the sum of squared errors (SSE) when fitting regression functions to the segments.

In this paper, we explicitly model the temporal nature of the problem by using a state-space model (where each state corresponds to data within a segment). Specifically we propose a hidden Markov model (HMM) framework. Each segment corresponds to a state in the HMM, so a change-point corresponds to switching from one state to another. Fwu and Djuric (1996) also used a HMM for signal segmentation. However, in this paper we incorporate into our framework two extensions of the basic HMM. One is the semi-Markov model to allow an arbitrary distribution on the location of the change-point (the standard HMM restricts it to be geometric). The other extension is a segmental HMM to model the "shape" in each segment. In contrast, Fwu and Djuric (1996) dealt only with piecewise constant signals.

In the following sections of the paper, we first describe the general segmental semi-Markov model framework for change detection, then a specific change-detection algorithm is given in section 4, followed by results and evaluation in section 5.

2 Markov and semi-Markov Models for Change Detection

We begin our discussion with a standard discrete-time finite-state Markov model where each segment in the data corresponds to a state of the Markov model. Let the number of states be M. The parameters of the model include π, the initial state distribution, and A, the $M \times M$ state transition matrix. In the context of change detection, the data $\mathbf{y} = y_1 y_2 \ldots y_t \ldots y_T$ are observed, but the corresponding states $\mathbf{s} = s_1 s_2 \ldots s_t \ldots s_T$ (i.e., segment labels) are hidden. In this *hidden* Markov model, the joint distribution of the observed data sequence $\mathbf{y}$ and a state sequence $\mathbf{s}$, can be factored as:

$$p(\mathbf{y}, \mathbf{s}) = \left(\prod_{t=2}^{T} p(y_t|s_t) p(s_t|s_{t-1}) \right) p(y_1|s_1)\pi(s_1), \tag{1}$$

A straightforward way of doing change detection is to infer, given $\mathbf{y}$, the hidden states $\mathbf{s}$, and see whether (and when) there is a state transition. We will go into more detail about this in Section 4. But first let us look at some limitations of the standard Markov model.

In the standard Markov framework the distribution of the durations of the system in state i is given by

$$p_i(t_d = d) = a_{ii}^{d-1}(1 - a_{ii}) \tag{2}$$

where a_{ii} is the self-loop transition probability of state i and d is the number of time-steps spent in state i. In other words, the Markov assumption constrains the state-duration distributions to be geometric in form. In reality we may want other kinds of distributions, e.g., Gaussian. For example, in Figure 1 we have prior knowledge from the physics of the plasma etch process that a change is more likely to occur about half-way through the process, rather than at the beginning or at the end.

The problem of modifying the standard Markov model to allow for arbitrary state-durations can be addressed by the use of *semi-Markov* models (e.g., Ferguson 1980). A semi-Markov model has the following generative description:

- On entering state i a duration time t_d is drawn from a state-duration distribution $p_i(t_d)$.
- The process remains in state i for time t_d.
- At time t_d the process transitions to another state according to a transition matrix A, and the process repeats.

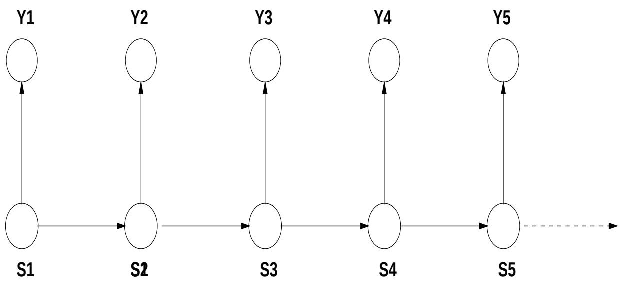

Figure 2: The graphical model for a discrete-time hidden semi-Markov model in the 'only one change" case.

The state-duration distributions, $p_i(t_d), 1 \leq i \leq M$, can be modeled using parametric distributions (such as log-normal, Gamma, etc) or non-parametrically by mixtures, kernel densities, etc. If t_d is constrained to take only integer values we get a discrete-time semi-Markov model. In a change-detection context, by including the state-duration distributions in the model, we can encode a prior on how long we expect the process to remain in each state. For the "only one change" model we need only model the distribution of change times from state 1 to state 2 (since that is the only legal transition). Such a model can be constructed from prior knowledge of the process. In applications where we have multiple "runs" of the same process, the prior can be adapted to the data over multiple runs.

The "only one change" model also permits a particularly simple representation as a graphical model in the discrete-time case (Figure 2). The semi-Markov model can be represented in this case by a non-stationary Markov model where the transition probabilities (from state 1 to state 2) are a deterministic function of the state duration distribution $p_1(t_d)$. More specifically, they can be derived directly from the cumulative distribution function corresponding to $p_1(t_d)$. A feature of this representation is that we have again the same graphical model structure corresponding to a hidden Markov model, with the attendant advantages of inference algorithms which scale linearly in the length of the sequence being analyzed (e.g., Smyth et al. 1997). Finally, note that going from a semi-Markov model to a semi-Markov HMM is straightforward; the unobserved state sequence is semi-Markov and the observed y's have the same dependency relations as for the standard HMM (Figure 2 again).

3 Segmental Observation Models

We have not yet described the functional form of the conditional densities $p(y_t|s_t)$ which relate the observed data to the hidden states. In the standard HMM framework (e.g., in speech recognition) the real-valued y_t's are often modeled as Gaussians or mixtures of Gaussians. For Gaussians, this implies a piecewise constant process with one mean μ_i per state i with additive Gaussian noise. Mixtures allow switching between multiple means per state, but still imply a constant regression process as a function of time.

There are many examples of real-world time-series where this piecewise-constant model is in-

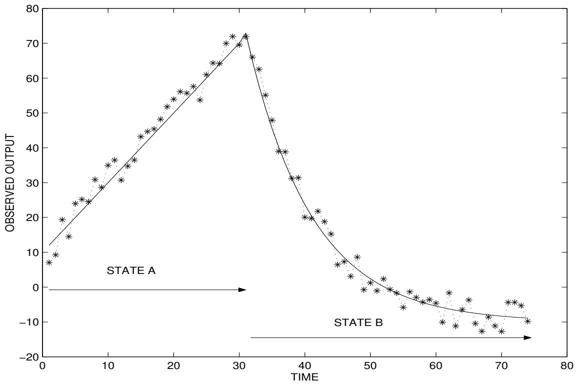

Figure 3: A simple illustration of the output of a simulated segmental Markov model. The solid lines show the underlying deterministic components of the regression models within each state, and the dotted lines show the actual noisy observations.

appropriate, e.g., the data in Figure 1. A natural generalization of the constant model is to allow each state to generate data in the form of a regression curve, i.e.,

$$y_t = f_i(t|\theta_i) + e_t \tag{3}$$

where $f_i(t|\theta_i)$ is a state-dependent regression function with parameters θ_i and e_t is additive independent noise (often assumed Gaussian, but not necessarily so). In the Gaussian noise case we get that $p(y_t|s_t = i)$ is Gaussian with a mean $f_i(t|\theta_i)$ which is a function of time and with variance σ^2. Note that conditioned on the regression parameters θ_i, the y_t's only depend on the current state s_t, as in the standard regression framework (i.e., observations are conditionally independent of everything else given the current state and state regression parameters). Thus, once again the simple graphical model structure of Figure 2 is applicable.

The segmental model (Holmes and Russell, 1999) is natural for the change detection problem since change detection is often applied to problems involving transient phenomena. Examples include linear trends, second order growth, exponential decay, etc. Figure 3 shows a simple example of the output of a simulated segmental Markov model. The process begins in state A which produces observations according to a noisy linear regression model. After 30 time steps or so it transitions to state B and produces observations according to a noisy exponential decay model.

4 A Change Detection Algorithm

Given the general framework presented above we now formulate a specific change-detection algorithm based on these ideas. We specifically focus on the "only one change" case, and give a computationally efficient solution based on the generative graphical model framework we propose.

Our model has the following specific components:

- The process is assumed to start in state 1 and can then transition to state 2 and stay there.

- A segmental hidden Markov model is used to model the "shape" in the data. We will assume that the regression functions for states 1 and 2 are linear in their parameters θ_1 and θ_2 and that the functional forms of the segment functions are known (typically from prior knowledge about the nature of the process), but not their parameters. We will also assume that the additive noise term e_t is Gaussian with zero mean and unknown variance σ^2.

- The process is semi-Markov, characterized by a state duration distribution for state 1, $p_1(t_d)$. Initially we will assume that this is specified a priori based on prior knowledge. As discussed earlier, this in turn specifies a set of non-stationary transition probabilities $p(s_t = 2|s_{t-1} = 1)$. If the prior on the state duration distribution is not available, we will assume a flat prior.

Clearly one could generalize each of these assumptions for specific applications. For example, if the functional form of the underlying regression shapes is not certain, one could fit multiple different parametric regression functions and adopt a Bayesian approach to select the best model (and change-point) which describes the data.

Under the state assumptions, given a sequence of observations $\mathbf{y}$, the change-detection problem in this context can be defined as finding the posterior probabilities of the hidden states of the model $p(\mathbf{s}|\mathbf{y})$ (inference). Also "missing" are the regression parameters for each shape. The approach we take here is to use the Expectation-Maximization (EM) algorithm to generate posterior estimates of these parameters and the state probabilities, given the data and the prior distribution on times of state changes.

The application of the EM algorithm to a segmental HMM is relatively straightforward, although slightly more complex than EM for a standard HMM. The E-step proceeds as usual to generate state probabilities given some estimates of the regression parameters θ_i, $i = 1, 2$. The only difference is that the likelihood of the observations is calculated using the segmental regression models rather than the usual "constant-mean" model. Given state probabilities, the M-step proceeds to use weighted regression (where the weights are the posterior probabilities of the hidden states) to estimate the θ_i's for each state. Weighted regression is a relatively simple extension of standard least-squares regression (Draper and Smith, 1998).

After EM converges we have point estimates of the regression parameters $\hat{\theta}_i, i = 1, 2$. To estimate whether (and when) the process changes from state 1 to state 2, we calculate the most likely state-sequence (MLSS) $\hat{\mathbf{s}}$ for the data given the model using the Viterbi algorithm (or equivalently, "the most probable explanation," MPE):

$$\hat{\mathbf{s}} = \mathrm{argmax}_{\mathbf{s}}\, p(\mathbf{s}|\mathbf{y}, \hat{\theta}). \tag{4}$$

If $\hat{\mathbf{s}}$ contains a transition from state 1 to 2, then a change occurs at the time of the transition. This is one way of estimating the location of the change time, and we call it the "MLSS" approach.

An alternative approach to detection is based on the following observation. The most likely state sequence is just one single state sequence with the highest posterior probability. Quite often there are many other state sequences with posterior probabilities *almost* as high (i.e., only slightly lower). Each of these state sequences makes a decision on the location of the change point, and the " MLSS approach" adopts the decision of $\hat{\mathbf{s}}$ only. Alternatively, we can pool together the decisions of all the state sequences $\mathbf{s}$, weighted by their posterior probabilities. We call this the *weighted* approach. Formally, the estimated change time is the weighted average

$$\hat{t}_c = \sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{y})\hat{t}_c^{(\mathbf{s})} \tag{5}$$

where $\{\mathbf{s}\}$ are the state sequences, $p(\mathbf{s}|\mathbf{y})$ is the posterior probability of $\mathbf{s}$, $\hat{t}_c^{(\mathbf{s})}$ is the change time for a particular state-sequence $\mathbf{s}$, i.e., the decision of $\mathbf{s}$ on the change time.

Equation 5 can be evaluated using dynamic programming in a manner similar to the forward-backward algorithm. Better still, we can make use of the calculated posterior probability of the hidden states at the end of the EM algorithm. The posterior probability of point i being in state 1, $p(s_i = 1|\mathbf{y})$, is related to $p(\mathbf{s}|\mathbf{y})$ by

$$p(s_i = 1|\mathbf{y}) = \sum_{\mathbf{s}:\hat{t}_c^{(\mathbf{s})}>i} p(\mathbf{s}|\mathbf{y}), \tag{6}$$

from which we have

$$\sum_{\mathbf{s}:\hat{t}_c^{(\mathbf{s})}=i} p(\mathbf{s}|\mathbf{y}) = p(s_{i-1} = 1|\mathbf{y}) - p(s_i = 1|\mathbf{y}). \tag{7}$$

This can be used to evaluate Equation 5:

$$\begin{aligned}
\hat{t}_c &= \sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{y})\hat{t}_c^{(\mathbf{s})} \\
&= \sum_i \left(i \times \sum_{\mathbf{s}:\hat{t}_c^{(\mathbf{s})}=i} p(\mathbf{s}|\mathbf{y}) \right) \\
&= \sum_i \Big(i \times (p(s_{i-1} = 1|\mathbf{y}) - p(s_i = 1|\mathbf{y})) \Big).
\end{aligned} \tag{8}$$

5 Experimental Results

5.1 Results on Simulated Data

We compared our change point detection method ("EM-Markov") with what we call the "SSE" method which minimizes the sum of squared errors (SSE) of all the fitted segments. As we mentioned in Section 1, when the number of segments is known, most existing techniques can be seen as trying to minimize the SSE when fitting regression functions to the segments.

To systematically evaluate the EM-Markov and SSE detection methods, they were tested on simulated data in the following manner:

- Data were simulated from a waveform of 100 points consisting of two linear segments with additive Gaussian noise (zero mean and variance σ^2). The slopes of two segments are $k_1 = 1$, $k_2 = 4$, respectively. See Figure 4 for an example of the simulated data.

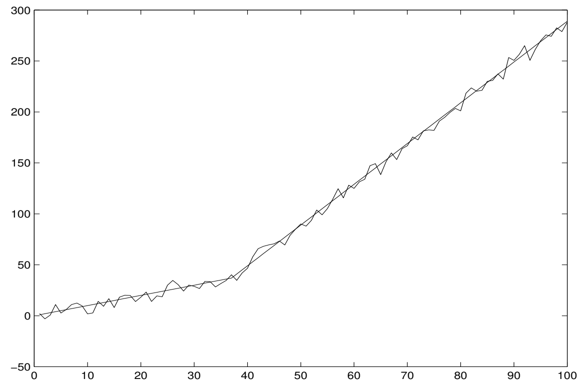

Figure 4: Synthetic data with 2 linear segments. The slopes are $k_1 = 1$, $k_2 = 2$. The noise $\sigma_y = 10$.

- The true change-point from one segment to the other is sampled from a Gaussian distribution with mean 35 and standard deviation 10. This distribution was used as the prior in the semi-Markov model. To test the sensitivity of the EM-Markov method to specification of the prior, we also looked at the case when no information on prior is available, in which case we used the flat prior, i.e., all points are equally likely to be a change point (note that this is still a semi-Markov model). Depending on whether the prior is used, and whether the "MLSS" or "weighted" approach is used to find the change point, we have four different variations of the EM-Markov method.

- 1000 random realizations of the process were generated and detections were made by the SSE method and the four variations of EM-Markov method.

- The experiment was repeated for three different noise levels: $\sigma = 5$, $\sigma = 10$, $\sigma = 15$.

Figure 5 shows a histogram of the errors of the detected change time for each of the SSE method and the EM-Markov method (with prior, "weighted") for $\sigma = 5$. The EM-Markov method is clearly superior in that the errors tend to be much smaller. Figure 6 tells a similar story for the higher noise case of $\sigma = 15$. Note that as the underlying noise increases, the detection problem is inherently more difficult (so the errors are larger for both methods). However, relative to the SSE method, the EM-Markov method is now doing even better, i.e., the reduction in error from SSE to EM-Markov is more pronounced at the higher noise level. The superiority of the EM-Markov method can also been seen in Figure 7 which shows the mean absolute errors. As the noise level increases the relative improvement from using the EM-Markov method also increases. This is as we might expect, since as the ambiguity in the data increases we would expect that a probabilistic model will be better able to deal with the ambiguities in the data compared to a non-probabilistic approach. Conversely, for low-noise situations, the detection problem will be relatively easy and we can expect relatively little difference between the two methods.

Additionally, we can see from the figure that the "weighted" variations of the EM-Markov method are much better than the "MLSS" variations, and that, not surprisingly, "with prior" is

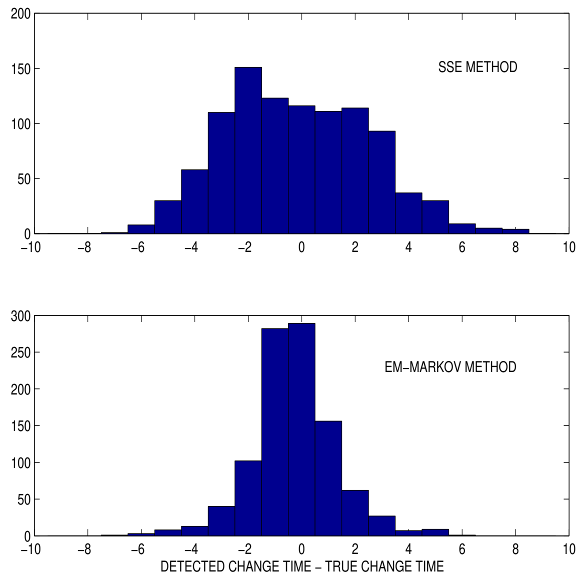

Figure 5: Histograms of the detection time errors for the SSE method and the EM-Markov method (with prior, "weighted"), $\sigma = 5$.

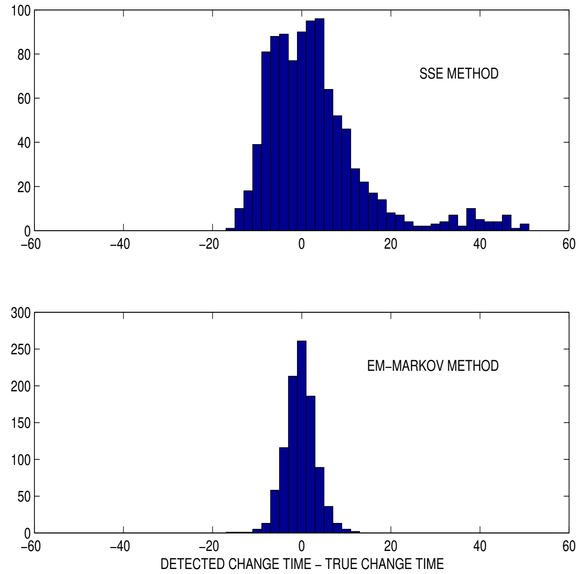

Figure 6: Histograms of the detection time errors for the SSE method and the EM-Markov method (with prior, "weighted"), $\sigma = 15$.

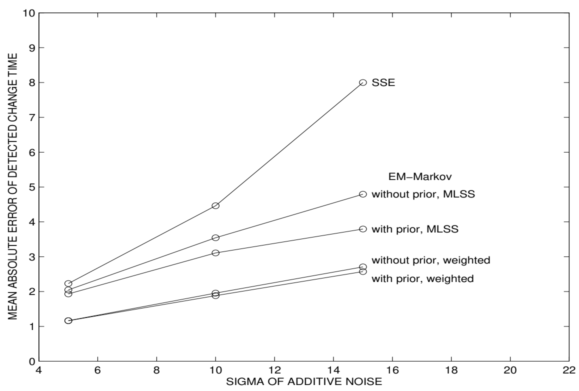

Figure 7: Comparison of the SSE method and the four variations of the EM-Markov method (with or without prior, "weighted" or MLSS), in terms of the mean absolute errors of detected change times.

better than "without prior". So the knowledge about the prior improves the performance, but it is not essential: without it, the performance is still much better than the SSE method and relatively close to performance with the prior.

5.2 Results on Plasma Etch Process Data

Plasma etch (Manos and Flamm, 1989; Williams, 1997) is a critical process in semiconductor manufacturing. A semiconductor wafer is exposed to a plasma containing various chemical components within a plasma gas chamber. The chemical composition of the plasma within the chamber is altered as a function of time to remove different layers from the wafer. Since there is no direct way of measuring when a layer on the wafer has been etched through, control of the plasma process (i.e., when to extinguish the plasma so as to stop etching) is achieved indirectly by inferring the nature of the material being etched from the spectral composition of the gas within the chamber.

Detecting the end of the etch process is quite important for reliable wafer processing. If etching is terminated too early, the desired pattern will not be etched on the wafer: if it is terminated too late, the etch may burn through to the next layer. In either case the electrical properties of the resultant wafer will not meet specifications, resulting in lower yields overall.

The data in Figure 8 show the output of a single channel spectrometer within the plasma chamber (tuned to a particular wavelength) during an actual wafer etch. The data come from a commercially available plasma etch machine (LAM 9400). Based on prior knowledge of the physics of the process, an engineer's best estimate of when the change occurred is at $t = 230$ seconds. Shown also on the figure is the change-point $t = 232$ detected by both the EM-Markov and the SSE methods. Both methods performed quite well on this particular data set. We are in the process of performing more extensive tests of the method on other wafer runs and expect (based on the simulation results) that the EM-Markov method will be significantly more reliable on average.

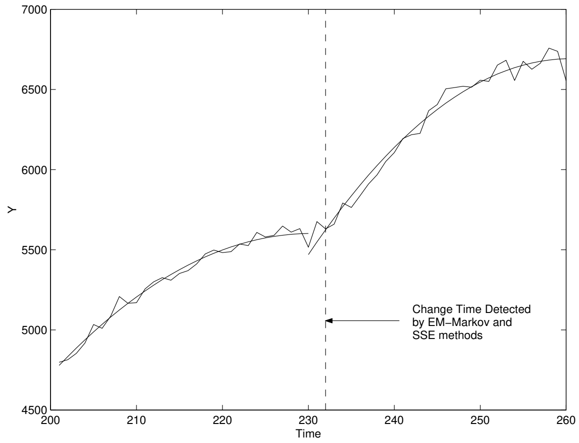

Figure 8: Detected change-points by EM-Markov and SSE methods for LAM plasma etch data.

6 Future Work and Conclusions

A natural extension of the off-line approach here is an online version of the same algorithm. For an online analysis, one essentially runs the algorithm in real-time as each new data point y_t arrives, generating inferences about the likely location of the change-point. However, in this case, one must also decide whether or not there is a change-point at all present. In theory, one can formulate this either in a hypothesis-testing or Bayesian framework in a relatively straightforward manner. Comparing different approaches empirically is a little more complex (than for the off-line case) since there now are multiple performance measures, e.g., rate of false alarms, detection delay, etc.

More generally, in an industrial setting such as plasma etch one has multiple runs over multiple wafers. In principle one can set up an adaptive Bayesian estimation methodology, where one "seeds" the system using a fairly uninformative prior and then allow the algorithm to recursively update the priors in a Bayesian fashion (where now one can have priors on regression shapes as well as expected time of change).

We conclude that the proposed segmental semi-Markov model appears to be quite a useful, flexible, and accurate framework for change-point detection. By modeling the problem within a generative model framework (including notions of state and time explicitly) one can incorporate prior knowledge in a principled manner and use the tools of probabilistic inference to infer change-points in an optimal manner.

Acknowledgements

We would like to thank Wenli Collison, Tom Ni, and David Hemker of LAM Research for providing the plasma etch data and for discussions on change-point detection in plasma etch processes. The research described in this paper was supported by NSF CAREER award IRI-9703120 and by the NIST Advanced Technology Program and KLA-Tencor.

References

M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: theory and application.* Prentice-Hall, Inc., 1993.

N. R. Draper and H. Smith. *Applied regression analysis.* John Wiley & Sons, Inc, 1998.

S. R. Esterby and A. H. El-Shaarawi. Inference about the point of change in a regression model. *Applied Statistics*, 30(3):277–285, 1981.

J. D. Ferguson. Variable duration models for speech. In *Proc. Symposium on the Application of Hidden Markov Models to Text and Speech*, pages 143–179, Oct 1980.

J.-K. Fwu and P. M. Djuric. Automatic segmentation of piecewise constant signal by hidden Markov models. In *Proceedings of 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pages 283–286, Jun 1996.

F. Gustafsson. Segmentation of signals using pieciewise constant linear regression models. Accepted for publication in IEEE Trans. on Signal Processing; available on line at `http://www.control.isy.liu.se/~fredrik/reports/TSPsegm.ps`, 1996.

D. M. Hawkins. Point estimation of the parameters of piecewise regression models. *Applied Statistics*, 25(1):51–57, 1976.

D. V. Hinkley. Inference in two-phase regression. *Journal of the American Statistical Association*, 66(336):736–743, 1971.

W. J. Holmes and M. J. Russell. Probabilistic-trajectory segmental HMMs. *Computer Speech and Language*, 13:3–37, 1999.

T. L. Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(4):613–658, 1995.

P. M. Lerman. Fitting segmented regression models by grid search. *Applied Statistics*, 29(1):77–84, 1980.

D. M. Manos and D. L. Flamm, editors. *Plasma Etching, An Introduction.* Academic Press, Inc., San Diego, 1989.

E. S. Page. Continuous inspection scheme. *Biometrika*, 41:100–115, 1954.

P. Smyth, D. Heckerman, and M. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1997.

P. F. Williams, editor. *Plasma Processing of Semiconductors.* Kuwer Academic Publishers, 1997.