Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

A Hidden Markov Filtering Approach to Multiple Change-point Models

Tze Leung Lai and Haipeng Xing

Abstract— **We describe a hidden Markov modeling approach to multiple change-points that has attractive computational and statistical properties. This approach yields explicit recursive filters and smoothers for estimating the piecewise constant parameters. Applications to array-CGH data analysis in genetic studies of cancer and to on-line detection, estimation and adaptive control of stochastic systems whose parameters may undergo occasional changes are given to illustrate the versatility of the proposed methodology.**

I. INTRODUCTION

The problems of change-point detection, estimation and adaptive control in stochastic systems which may undergo abrupt changes over time arise in many fields in science and engineering, including industrial quality control, automated fault detection in controlled dynamical systems, segmentation of signals, and computational molecular biology. In this paper we describe a unified approach to these problems by using a class of hidden Markov models that have tractable forward and backward filters. These filters have explicit representations, which can be further approximated by parallel recursive algorithms for on-line implementation.

The filtering approach, via hidden Markov modeling, to change-point problems dates back to Shiryaev [1,2] who formulated the problem of optimal sequential detection based on observations y_t that are independent with a common density function f_0 for $t < \nu$ and with another common density function f_1 for $t \geq \nu$, in which f_0 and f_1 are known density functions. Assuming a loss of c for each observation taken after ν and a loss of 1 for a false alarm before ν, Shiryaev used optimal stopping theory to show that the Bayes rule (corresponding to the prior geometric distribution on ν) triggers an alarm as soon as the posterior probability that a change has occurred exceeds some threshold; he also derived a continuous-time analog of this result. Shiryaev's theory assumes that there is a single change-point and that the pre- and post-distributions of y_t are known. The unobserved "state" in this theory is the change-time ν and the Bayes rule can be expressed in terms of the "filter," which is the posterior distribution of ν given the current and past observations $y_1, \ldots, y_t$.

To generalize to multiple change-points, one can use more general jump Markov models, as in [3]. Moreover, to remove the assumption of known pre- and post-change densities, an obvious way is to replace $\{f_0, f_1\}$ by a parametric family $\{f_\theta, \theta \in \Theta\}$ and ν by a sequence of positive interger-valued random variables $\nu_1, \nu_2, \ldots$ with i.i.d. increments. Thus, y_t has conditional density function f_{θ_t} given θ_t, in which the θ_t are constant between the change-times $\nu_1, \nu_2, \ldots$. The assumption of i.i.d. geometrically distributed inter-arrival times between change-points is equivalent to i.i.d. Bernoulli $I_t \triangleq 1_{\{\theta_t \neq \theta_{t-1}\}}$ (i.e., whether a parameter change occurs at time t). To complete the Bayesian specification, we assume a prior distribution G on θ so that θ_t has distribution G whenever a parameter change occurs. This is, therefore, a hidden Markov model with the following dynamics:

$$\begin{tikzcd}
\text{Bernoulli:} & & I_2\cdots \arrow[d] & I_t \arrow[d] & I_{t+1} \arrow[d] & \\
G: & \theta_1 \arrow[r] \arrow[d] & \theta_2\cdots \arrow[r] \arrow[d] & \theta_t \arrow[r] \arrow[d] & \theta_{t+1} \arrow[r] \arrow[d] & \text{Hidden states} \\
f_\theta: & y_1 \arrow[r] & y_2\cdots \arrow[r] & y_t \arrow[r] & y_{t+1} \arrow[r] & \text{Observations}
\end{tikzcd} \tag{1}$$

One such model was considered by Yao [4] who derived recursive formulas for the Bayes estimates of the means θ_t based on $y_1, \ldots, y_n$ $(1 \leq t \leq n)$, in which

$$y_t = \theta_t + \sigma\varepsilon_t, \tag{2}$$

ε_t are i.i.d. standard normal random variables, σ is a known scale parameter, and $\theta_t = \theta_{t-1}$ with probability $1-p$ and takes on a new normally distributed value with probability p, i.e., G is normal with mean μ and variance V. These recursive formulas involve the forward and backward filters of the hidden Markov model and can be computed by $O(n^3)$ operations. Earlier, Chernoff and Zacks [5] considered a similar model that assumes a $N(\mu, V)$ distribution on the change magnitude $\theta_t - \theta_{t-1}$ (rather than on θ_t as in [4]) whenever $\theta_t \neq \theta_{t-1}$. The computational complexity of the Bayes estimate of θ_t $(1 \leq t \leq n)$ in the Chernoff-Zacks model, however, is of exponential order 2^n. Note that Shiryaev's model is also a special case of the hidden Markov model (1) with $\Theta = \{0, 1\}$ and with 1 as an absorbing state of the Markov chain $\{\theta_t\}$ which is initialized at 0.

In Section II we derive recursive filters and smoothers in the hidden Markov model (1) to estimate the piecewise constant parameters θ_t. For the normal mean shift case, our recursive formulas agree with those of Yao [4] although our method to derive these formulas is different from his and is more versatile. A key ingredient in our method is the most recent change-time K_t up to time t. Our derivation also shows how the posterior distribution of K_t can be approximated by a bounded number of parallel recursions, with the bound

Tze Leung Lai is with the Department of Statistics, Stanford University, Palo Alto, CA 94305, USA lait@stanford.edu. His research was supported by the National Science Foundation under Grand DMS-0305749.

Haipeng Xing is with the Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794, USA xing@ams.sunysb.edu.

depending on p but not on n. On-line and off-line methods to estimate p and other hyperparameters are also discussed. Section III gives an overview of a variety of applications of the filters and smoothers in Section II, and some concluding remarks are given in Section IV.

II. A Hidden Markov Model Framework for Multiple Change-points

The model (1) provides a general framework for multiple change-points in complex systems. Suppose observations $y_1, y_2, \ldots$ are generated from the model f_θ with piecewise constant parameters θ_t. The prior distribution of θ_t is given recursively by $\theta_t = (1-I_t)\theta_{t-1} + I_t\eta_t$, in which η_t are i.i.d. random variables with common density function g and I_t are i.i.d. Bernoulli random variables with success probability p. We assume that θ_1 has density function g, which is tantamount to initializing the above recursive definition of θ_t with $I_t = 1$.

A. Recursive Filtering Formulas

Let $\mathscr{Y}_t = (y_1, \ldots, y_t)$ and

$$K_t = \max\{s \le t : I_s = 1\}, \tag{3}$$

i.e., K_t is the most recent change-point up to time t. The conditional density function $g_{i,t}$ of θ_t given $K_t = i$ and $\mathscr{Y}_t$ is

$$g_{i,t}(\theta) \propto g(\theta)\prod_{s=i}^{t} f_\theta(y_s), \tag{4}$$

in which the constant of proportionality is defined by that $g_{i,t}(\theta)$ has to integrate (or sum) to 1 over θ. Let $p_{i,t} = P(K_t = i|\mathscr{Y}_t)$. Then the posterior density $f_{t|t}(\cdot)$ of θ_t given $\mathscr{Y}_t$ can be expressed as

$$f_{t|t}(\theta) = \sum_{i=1}^{t} p_{i,t}g_{i,t}(\theta), \tag{5}$$

where $p_{i,t} = p^*_{i,t}/\sum_{k=i}^{t} p^*_{k,t}$ and

$$p^*_{i,t} = \begin{cases} pf(y_t|I_t = 1) & \text{if } i = t, \\ (1-p)p_{i,t-1}f(y_t|\mathscr{Y}_{t-1}, K_t = i) & \text{if } i < t, \end{cases}$$

in which we use $f(\cdot|\cdot)$ to denote conditional densities. Since

$$f(y_t|\mathscr{Y}_{t-1}, K_t = i) = \int f_\theta(y_t)g_{i,t-1}(\theta)d\theta,$$

we can use (4) to express $p^*_{i,t}$ as

$$p^*_{i,t} = \begin{cases} p\pi_{0,0}/\pi_{t,t} & \text{if } i = t, \\ (1-p)p_{i,t-1}\pi_{i,t-1}/\pi_{i,t} & \text{if } i < t, \end{cases} \tag{6}$$

in which

$$1/\pi_{0,0} = \int g(\theta)d\theta, \quad 1/\pi_{i,j} = \int \Big[\prod_{k=i}^{j} f_\theta(y_k)\Big]g(\theta)d\theta,$$

with the integral replaced by summation in the discrete case.

B. Combining Forward and Backward Filters for $\theta_t|\mathscr{Y}_n$

To find the posterior density $f_{t|n}(\cdot)$ of θ_t given $\mathscr{Y}_n$ for $1 \le t < n$, we first reverse time and note that $\widetilde{I}_s \triangleq I_{n-s+1}$ are still i.i.d. Bernoulli and that the time-reversed Markov chain $\widetilde{\theta}_s \triangleq \theta_{n-s+1}$ has the same transition probabilities as the forward chain θ_s. In other words, $\{\theta_s, 1 \le s \le n\}$ is reversible. Moreover, its stationary distribution has density function g. Let $\mathscr{Y}_{i,j} = (y_i, \ldots, y_j)$ and $\widetilde{K}_t = \min\{s > t : I_s = 1\}$. Analogous to the forward filter, the posterior density $f(\cdot|\mathscr{Y}_{t+1,n})$ of θ_t given $\mathscr{Y}_{t+1,n}$ is given by

$$f(\theta|\mathscr{Y}_{t+1,n}) = pg(\theta) + (1-p)\sum_{j=t+1}^{n} q_{j,t+1}g_{t+1,j}(\theta), \tag{7}$$

where $q_{j,t+1} = P(\widetilde{K}_t = j|\mathscr{Y}_{t+1,n}) = q^*_{j,t+1}/\sum_{i=j}^{n} q^*_{i,t+1}$ and

$$q^*_{j,s} = \begin{cases} p\pi_{0,0}/\pi_{s,s} & \text{if } j = s, \\ (1-p)q_{j,s+1}\pi_{s+1,j}/\pi_{s,j} & \text{if } j > s. \end{cases} \tag{8}$$

By Bayes' theorem, for $1 \le t < n$, the posterior density $f_{t|n}(\cdot)$ of θ_t given $\mathscr{Y}_n$ satisfies

$$f_{t|n}(\theta) \propto f(\theta|\mathscr{Y}_t)f(\theta|\mathscr{Y}_{t+1,n})/g(\theta). \tag{9}$$

Combining (9) with (5) and (7), simple algebra then yields

$$f_{t|n}(\theta) = \sum_{1\le i\le t\le j\le n} \beta_{ijt}g_{i,j}(\theta), \tag{10}$$

where $\beta_{ijt} = \beta^*_{ijt}/P_t$, $P_t = p + \sum_{1\le i\le t<j\le n}\beta^*_{ijt}$, and

$$\beta^*_{ijt} = \begin{cases} pp_{it} & \text{if } i \le t = j, \\ (1-p)p_{it}q_{j,t+1}\pi_{i,t}\pi_{t+1,j}/\pi_{i,j}\pi_{0,0} & \text{if } i \le t < j. \end{cases} \tag{11}$$

C. Exponential Family as a Special Case

The preceding formulas are particularly convenient if the prior density g is chosen from a conjugate family $\{g_\alpha\}$ so that the posterior density of θ (in the case of no change-point) given the data also belongs to the family but with $\alpha(\mathscr{Y}_n)$ in place of α. For example, for the normal mean shift model (2), the normal prior distribution of θ_t at change-times is a conjugate family. One extension of the normal mean shift model (2) is the exponential family

$$f_\theta(y) = \exp\{\theta^T y - \psi(\theta)\} \tag{12}$$

with respect to some measure m on R^d. The Bayesian model assumes a conjugate prior distribution on the natural parameter space $\Theta := \{\theta : \int e^{\theta^T y}dm(y) < \infty\}$. Specifically, the prior density g_α has the form

$$g_\alpha(\theta) \triangleq g(\theta; a_0, \mu_0) = c(a_0, \mu_0)\exp\{a_0\mu_0^T\theta - a_0\psi(\theta)\}, \tag{13}$$

where $\alpha = (a_0, \mu_0)$ and $1/c(a_0, \mu_0) = \int_\Theta \exp\{a_0\mu_0^T\theta - a_0\psi(\theta)\}d\theta$. The posterior density of θ given the observations $y_1, \ldots, y_m$ from f_θ is $g(\theta; a_m, \mu_m)$, where

$$a_m = a_0 + m, \quad \mu_m = (a_0\mu_0 + \sum_{i=1}^{m} y_i)/(a_0 + m); \tag{14}$$

see [6], which gives the following explicit formulas for $\pi_{i,j}$ in (6) and (8):

$$\pi_{0,0} = c(a_0, \mu_0), \quad \pi_{i,j} = c(a_0 + j - i + 1, \bar{y}_{i,j}),$$

where $\bar{y}_{i,j} = (a_0\mu_0 + \sum_{k=i}^{j} y_k)/(a_0 + j - i + 1)$ for $j \geq i$. Moreover, the posterior density function $g_{i,t}$ in (5) or (10) belongs to the above conjugate family.

A *bounded complexity mixture* (*BCMIX*) approximation, with $M(p)$ components, to the mixture (10) of t posterior densities can be obtained as follows. Let $\mathcal{K}_{t-1}(p)$ denote the set of indices i for which $p_{i,t-1}$ is kept at stage $t-1$; thus $\mathcal{K}_{t-1}(p) \supset \{t-1, , \cdots, t-m(p)\}$. At stage t, define $p^*_{i,t}$ as in (8) for $i \in \{t\} \cup \mathcal{K}_{t-1}(p)$ and let i_t be the index not belonging to $\{t, t-1, \cdots, t-m(p)+1\}$ such that

$$p^*_{i_t,t} = \min\{p^*_{j,t} : j \in \mathcal{K}_{t-1}(p) \quad \text{and} \quad j \leq t - m(p)\},$$

choosing i_t to be the minimizer farthest from t if the above set has two or more minimizers. Define $\mathcal{K}_t(p) = \{t\} \cup (\mathcal{K}_{t-1}(p) - \{i_t\})$ and let

$$p_{i,t} = \Big(p^*_{i,t} \Big/ \sum_{j \in \mathcal{K}_t(p)} p^*_{j,t}\Big), \quad i \in \mathcal{K}_t(p).$$

Similarly, we can obtain a BCMIX approximation to the backward filter $\theta_t | \mathcal{Y}_{t+1,n}$, and the BCMIX approximation to the smoother can be obtained by combining the forward and backward BCMIX filters via Bayes' theorem.

D. Stochastic Regression Models

Another far-reaching extension of Yao's normal mean shift model (2) is the stochastic regression model

$$y_t = \theta_t^T x_t + \sigma \varepsilon_t, \tag{15}$$

in which ε_t are i.i.d. standard normal, x_t is an observed regressor that is determined by the events up to time $t-1$, and $\theta_t \in R^k$ are piecewise constant parameters. The prior distribution of θ_t when a change occurs at time t is $N(\mu, V)$, and hence the posterior distribution of θ_t given $\mathcal{Y}_{i,j}$, $(K_t, \widetilde{K}_t) = (i, j)$ and $(x_s, i \leq s \leq j)$ is $N(\mu_{i,j}, \sigma^2 V_{i,j})$, where

$$V_{i,j} = \Big(V^{-1} + \sum_{k=i}^{j} x_k x_k^T\Big)^{-1}, \quad \mu_{i,j} = V_{i,j}\Big(V^{-1}\mu + \sum_{k=i}^{j} y_k x_k\Big). \tag{16}$$

Let $\phi_{\mu,V}$ denote the density function of $N(\mu, V)$ distribution, i.e., $\phi_{\mu,V}(y) = (2\pi)^{-k/2}|V|^{-1/2}\exp\{-\frac{1}{2}(y-\mu)^T V^{-1}(y-\mu)\}$. Then $\pi_{0,0} = \phi_{\mu,V}(0)$, $\pi_{i,j} = \phi_{\mu_{i,j},V_{i,j}}(0)$, yielding an explicit mixture of normal densities in (5) or (10). While the number of components in this normal mixture increases with n, we can approximate it by a mixture with at most M components, with M depending on p but not on n, similar to the BCMIX filters and smoothers in the preceding paragraph. Moreover, for $s < t$, $V_{s,t}$ and $\mu_{s,t}$ can be computed by standard recursions that follow from the matrix inversion lemma:

$$V_{t,t} = V - V x_t x_t^T V/(1 + x_t^T V x_t),$$
$$V_{s,t} = V_{s,t-1} - V_{s,t-1} x_t x_t^T V_{s,t-1}/(1 + x_t^T V_{s,t-1} x_t) \quad \text{if } s < t,$$
$$\mu_{s,t} = \mu_{s,t-1} + V_{s,t-1} x_t (y_t - x_t^T \mu_{s,t-1})/(1 + x_t^T V_{s,t-1} x_t).$$

E. On-line and Off-line Estimation of Hyperparameters

The preceding recursive filters and smoothers involve hyperparameters which include the relative frequency p of change-points, the parameter α of the prior density g_α, and also the error variance σ^2 in the case of the regression model (15). They can be estimated on-line by Rissanen's [7] accumulated prediction error, and off-line by maximum likelihood or variants thereof. Let Φ denote the vector of hyperparameters. For the hidden Markov model (1), the accumulated prediction error at time t is defined by

$$\text{APE}_t(\Phi) = \sum_{i=1}^{t} \{y_i - \widehat{y}_{i|i-1}(\Phi)\}^2, \tag{17}$$

where $\widehat{y}_{i|i-1}(\Phi)$ is the minimum-variance one-step-ahead predictor of y_i, or equivalently, $\widehat{y}_{i|i-1}(\Phi) = E_\Phi(y_i|\mathcal{Y}_{i-1})$. Instead of squared error, one can also use Kullback-Leibler divergence or other loss functions, depending on the applications. Details of the implementation are given in [8] and [9]. For off-line estimation of the hyperparameters, note that the likelihood function of these hyperparameters is given by the joint density function of $(y_1, \ldots, y_n)$, which is

$$\prod_{t=1}^{n} f(y_t|\mathcal{Y}_{t-1}) \propto \prod_{t=1}^{n} \Big(\sum_{i=1}^{t} p^*_{i,t}\Big), \tag{18}$$

where $p^*_{i,t} = p^*_{i,t}(\Phi)$ is given by (6) and the constant of proportionality does not depend on Φ. In Section III we describe how the maximum likelihood estimator of Φ or variants thereof can be computed in a specific application.

III. Applications to Detection, Segmentation, Estimation and Control

A. Segmentation Models for Array-CGH Data

Array-based comparative genomic hybridization (array-CGH) has become a widely used method for studying the genetics of cancer. For a given cell sample, array-CGH allows quantitative measurement of the average genomic DNA copy number at thousands of locations linearly ordered along the chromosomes. Typically, a test genomic DNA pool (e.g. genomic DNA from tumor cell sample) and a diploid reference genomic DNA pool are differentially labeled with dyes. These two dye-labeled samples are mixed and hybridized to a microarray chip, which is spotted with genomic targets that map to known locations on a global scale throughout the genome. The hybridized chip is then scanned, and the ratio of the test and reference fluorescence intensities for each genomic target is calculated. The ratio of the intensities of the dyes is a surrogate for the ratio of the abundance of the DNA sample labeled with the dyes, and the log fluorescence ratios y_t follow the model (2), in which ε_t are assumed to be independent standard normal as in Yao's [4] normal mean shift model. However, since the copy number of a homogeneous sample of normal diploid cells should be 2 on all autosomal chromosomes, giving a signal of 0, it is important to include 0 as the baseline state. Similar assumptions are needed in the fault detection applications in Section IIIC when the "in-control" state is known. Lai, Xing and Zhang [10] therefore assume the following Markovian dynamics on θ_t that has the baseline state 0. When the signal leaves the baseline, it moves to a non-zero state; when the next jump occurs, the signal may move back to the baseline or jump to another non-zero state. To describe the dynamics of θ_t, we use the transition probability matrix

$$P = \begin{pmatrix} 1-p & \frac{1}{2}p & \frac{1}{2}p \\ c & a & b \\ c & b & a \end{pmatrix}. \tag{20}$$

The matrix P specifies that, at time t, if the state θ_t is in the 0 (baseline) state, then at time $t+1$, θ_{t+1} stays in the 0 state with probability $1-p$, or jumps to a nonzero state which follows $N(\mu,V)$ with probability p. To allow the possibility of jumping from a non-zero state to a different non-zero state, we simply assume that the process can jump from the baseline state with probability $p/2$ to either of two nonzero states that have the same prior distribution $N(\mu,V)$. If $\theta_t \neq 0$, then at time $t+1$, it can stay in the last state with probability a, or jump to another nonzero state with probability b, or jump back to the baseline state with probability c.

The probability vector $\tilde{\pi} = (c/(p+c), \frac{1}{2}p/(p+c), \frac{1}{2}p/(p+c))$ satisfies $\tilde{\pi}P = \tilde{\pi}$, and therefore $\tilde{\pi}$ corresponds to the stationary distribution associated with P. Note also that

$$\tilde{\pi}(x)P(x,y) = \tilde{\pi}(y)P(y,x),$$

so the three-state Markov chain with transition probability matrix P and initialized at $\tilde{\pi}$ is reversible. This implies that the Markov chain $\{\theta_t\}$ has a stationary distribution π that assigns probability $c/(p+c)$ to the baseline value 0 and probability $p/(p+c)$ to a $N(\mu,V)$ random variable. Moreover, under the additional assumption that θ_0 is initialized at the stationary distribution, $\{\theta_t\}$ is a reversible Markov chain.

In analogy with (3), let $K_t = \max\{s \leq t : \theta_s = \cdots = \theta_t, \theta_{s-1} \neq \theta_s\}$ denote the nearest change-point at a location less than or equal to t. Define $p_t = P(\theta_{K_t} = 0|\mathscr{Y}_t) = P(\theta_t = 0|\mathscr{Y}_t)$ and $q_{i,t} = P(\theta_{K_t} \neq 0, K_t = i|\mathscr{Y}_t)$ for $1 \leq i \leq t$. Since the conditional distribution of θ_t, given $\mathscr{Y}_t$ and the event that $K_t = i$ and $\theta_{K_t} \neq 0$, is $N(\mu_{i,t}, V_{i,t})$, where $\mu_{i,j}$ and $V_{i,j}$ are given by (16) with $x_t \equiv 1$. it follows that the posterior distribution of θ_t given $\mathscr{Y}_t$ is a mixture of normal distributions and a point mass at 0:

$$\theta_t|\mathscr{Y}_t \sim p_t\delta_0 + \sum_{i=1}^{t} q_{i,t}N(\mu_{i,t}, V_{i,t}), \tag{21}$$

where δ_x denotes the probability distribution that assigns probability 1 to x. Making use of $p_t + \sum_{i=1}^t q_{i,t} = 1$ and $y_t = \theta_t + \sigma\varepsilon_t$, it is shown in [10] that the conditional probabilities p_t and $q_{i,t}$ can be determined by the recursions

$$\begin{aligned} p_t &\propto p_t^* := (1-p)p_{t-1} + cq_{t-1}, \\ q_{i,t} &\propto q_{i,t}^* := \begin{cases} (pp_{t-1} + bq_{t-1})\pi_{0,0}/\pi_{t,t}, & i = t, \\ aq_{i,t-1}\pi_{i,t-1}/\pi_{i,t}, & i < t, \end{cases} \end{aligned} \tag{23}$$

where $q_t = \sum_{i=1}^t q_{i,t} = 1 - p_t$, $\pi = \phi_{\mu,V}(0)$ and $\pi_{i,j} = \phi_{\mu_{i,j},V_{i,j}}(0)$ for $i \leq j$. Specifically, $p_t = p_t^* / \left[p_t^* + \sum_{i=1}^t q_{i,t}^*\right]$ and $q_{i,t} = q_{i,t}^* / \left[p_t^* + \sum_{i=1}^t q_{i,t}^*\right]$.

Since $\{\theta_t\}$ is a reversible Markov chain, we can reverse time and obtain a backward filter that is analogous to (21):

$$\theta_{t+1}|\mathscr{Y}_{t+1,n} \sim \widetilde{p}_{t+1}\delta_0 + \sum_{j=t+1}^{n} \widetilde{q}_{j,t+1}N(\mu_{t+1,j}, V_{t+1,j}),$$

in which the weights $\widetilde{p}_s, \widetilde{q}_{j,s}$ can be obtained by backward induction using the time-reversed counterpart of (23):

$$\begin{aligned} \widetilde{p}_s &\propto \widetilde{p}_s^* := (1-p)\widetilde{p}_{s+1} + c\widetilde{q}_{s+1}, \\ \widetilde{q}_{j,s} &\propto \widetilde{q}_{j,s}^* := \begin{cases} (p\widetilde{p}_{s+1} + b\widetilde{q}_{s+1})\pi_{0,0}/\pi_{s,s} & j = s, \\ a\widetilde{q}_{j,s+1}\pi_{s+1,j}/\pi_{s,j} & j > s, \end{cases} \end{aligned} \tag{24}$$

where $\widetilde{q}_{s+1} = \sum_{j=s+1}^n \widetilde{q}_{j,s+1} = 1 - \widetilde{p}_{s+1}$. As in (9) and (10), the forward and backward filters can be combined via Bayes' theorem to yield

$$\theta_t|\mathscr{Y}_n \sim \alpha_t\delta_0 + \sum_{1\leq i\leq t\leq j\leq n} \beta_{ijt}N(\mu_{ij}, V_{ij}), \tag{25}$$

in which

$$\begin{aligned} &\alpha_t = \alpha_t^*/A_t, \quad \beta_{ijt} = \beta_{ijt}^*/A_t, \quad A_t = \alpha_t^* + \sum_{1\leq i\leq t\leq j\leq n} \beta_{ijt}^*, \\ &\alpha_t^* = p_t[(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}]/c, \\ &\beta_{ijt}^* = \begin{cases} q_{i,t}(p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1})/p, & i \leq t = j, \\ aq_{i,t}\widetilde{q}_{j,t+1}\pi_{i,t}\pi_{t+1,j}/(p\pi_{0,0}\pi_{i,j}), & i \leq t < j. \end{cases} \end{aligned} \tag{26}$$

From (25), it follows that

$$P(\theta_t = 0|\mathscr{Y}_n) = \alpha_t, \quad E(\theta_t|\mathscr{Y}_n) = \sum_{1\leq i\leq t\leq j\leq n} \beta_{ijt}z_{i,j}.$$

Let C_{ij} be the segment $[i,j]$ in which all the copy numbers are same and $\theta_i \neq \theta_{i-1}$ and $\theta_j \neq \theta_{j+1}$, one may want to make inferences on properties of a genomic segment that is not identified by a segmentation procedure. A fundamental entity from which these inferences on genomic regions can be derived is the posterior distribution of the parameter sequence $\{\theta_t : 1 \leq t \leq n\}$ given $\mathscr{Y}_n$, which is an inhomogeneous Markov chain whose initial distribution is π and whose transition probabilities are given by

$$\theta_t|\theta_{t-1}, \mathscr{Y}_n \sim a_t\delta_0 + c_t 1_{\{\theta_{t-1}\neq 0\}}\delta_{\theta_{t-1}} + \sum_{j=t}^{n} b_{jt}N(\mu_{t,j}, V_{t,j}), \tag{27}$$

in which $a_t = a_t^*/B_t$, $c_t = c_t^*/B_t$, $b_{jt} = b_{jt}^*/B_t$ and

$$\begin{aligned} B_t &= a_t^* + c_t^* 1_{\{\theta_{t-1}\neq 0\}} + \sum_{j=t}^{n} b_{jt}^*, \\ a_t^* &= \phi_{0,\sigma^2}(y_t)\left[(1-p)1_{\{\theta_{t-1}=0\}} + c1_{\{\theta_{t-1}\neq 0\}}\right] \\ &\qquad \cdot \left[(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}\right] \Big/ c, \\ c_t^* &= a\phi_{\theta_{t-1},\sigma^2}(y_t)\Big\{(p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1}) + \\ &\qquad a\sum_{j=t+1}^{n} \widetilde{q}_{j,t+1}\frac{\phi_{\mu_{t+1,j},V_{t+1,j}}(\theta_{t-1})}{\phi_{\mu,v}(\theta_{t-1})}\Big\} \Big/ p, \\ b_{jt}^* &= \left[p1_{\{\theta_{t-1}=0\}} + b1_{\{\theta_{t-1}\neq 0\}}\right]\phi_{0,\sigma^2}(y_t)\widetilde{q}_{j,t}^*/p, \end{aligned}$$

using the same notation as that in (26).

To estimate the hyperparameters $\Phi = (p, a, b, \mu, V, \sigma^2)$, note that the likelihood function of Φ is given by the joint density function of $(y_1, \dots, y_n)$, which is

$$f(y_t|\mathscr{Y}_{t-1}) = \prod_{t=1}^{n} \Big\{ \big(p_t^* + \sum_{i=1}^{t} q_{i,t}^*\big)\phi_{0,\sigma^2}(y_t) \Big\}, \qquad (28)$$

in which $f(\cdot|\cdot)$ denotes conditional density function. Maximizing (28) over Φ yields the maximum likelihood estimate $\widehat{\Phi}$. Since Φ is a 6-dimensional vector and the functions $p_t^*(\Phi)$ and $q_{i,t}^*(\Phi)$ have to be computed recursively for $1 \le t \le n$, direct maximization of (28) may be computationally expensive due to the curse of dimensionality. An alternative approach is to use the EM algorithm which exploits the much simpler structure of the log likelihood $l_c(\Phi)$ of the complete data $\{(y_t, \theta_t), 1 \le t \le n\}$:

$$\begin{aligned} l_c(\Phi) = & -\tfrac{1}{2}\textstyle\sum_{t=1}^n \Big\{ \tfrac{(y_t-\theta_t)^2}{\sigma^2} + \log(2\pi\sigma^2) \Big\} \\ & -\tfrac{1}{2}\textstyle\sum_{t=1}^n \Big\{ \tfrac{(\theta_t-\mu)^2}{V} + \log(2\pi V) \Big\} 1_{\{0\neq\theta_t\neq\theta_{t-1}\}} \\ & +\textstyle\sum_{t=1}^n \Big\{ [\log(1-p)] 1_{\{\theta_t=\theta_{t-1}=0\}} + (\log p) 1_{\{\theta_t\neq\theta_{t-1}=0\}} \Big\} \\ & +\textstyle\sum_{t=}^n \Big\{ [\log(1-b-c)] 1_{\{\theta_t=\theta_{t-1}\neq 0\}} + (\log c) 1_{\{\theta_t=0\neq\theta_{t-1}\}} \\ & +(\log b) 1_{\{0\neq\theta_t\neq\theta_{t-1}\neq 0\}} \Big\}. \end{aligned} \qquad (29)$$

Since $l_c(\Phi)$ decomposes into normal and multinomial components, the E-step of the EM algorithm involves $E\big((\theta_t - \mu)^2|\mathscr{Y}_n\big)$, $E\big((\theta_t - y_t)^2|\mathscr{Y}_n\big)$ and the conditional probabilities

$$\begin{aligned} P(\theta_t = 0 = \theta_{t-1}|\mathscr{Y}_n) &= \frac{(1-p)p_{t-1}\alpha_t}{(1-p)p_{t-1} + cq_{t-1}}, \\ P(\theta_t = 0 \neq \theta_{t-1}|\mathscr{Y}_n) &= \frac{cq_{t-1}\alpha_t}{(1-p)p_{t-1} + cq_{t-1}}, \\ P(\theta_t \neq \theta_{t-1} = 0|\mathscr{Y}_n) &= c\widetilde{q}_t\alpha_{t-1}/\{(1-p)\widetilde{p}_t + c\widetilde{q}_t\}, \\ P(0 \neq \theta_t \neq \theta_{t-1} \neq 0|\mathscr{Y}_n) &= (\sum_{j=t}^{n} \beta_{tjt}) bq_{t-1}/\{bq_{t-1} + pp_{t-1}\}, \end{aligned}$$

together with $P(\theta_t = \theta_{t-1} \neq 0|\mathscr{Y}_n)$, which is determined by the property that those five conditional probability have to sum up to 1. Moreover, it follows from (29) that the M-step of the EM algorithm also has closed-from updating formulas.

B. Change-point Models for both Level and Variability

The array-CGH analysis in [10] focuses on BAC arrays, which use bacterial artificial chromosomes (BAC) as genomic targets. For other array-CGH platforms such as cDNA arrays, which measure copy numbers only at transcribed regions of the genome, one needs to incorporate changes in both the mean and the variance. In speech signal processing that uses autoregressive models, both the autoregressive parameters and the error variance are assumed to be piecewise constant; see [11] and [12].

To allow changes in both the regression parameters and the error variance in the stochastic regression model (15), Lai, Liu and Xing [8] have considered the special case of autoregressive models of the form

$$y_t = \alpha_t + \beta_{1,t}y_{t-1} + \cdots + \beta_{k,t}y_{t-k} + \sigma_t\varepsilon_t, \qquad t > k, \qquad (30)$$

where the ε_t are standard normal and $\theta_t \triangleq (\alpha_t, \beta_{1,t}, \cdots, \beta_{k,t}, 1/(2\sigma_t^2))^T$ satisfies $\theta_t = (1 - I_t)\theta_{t-1} + I_t(z_t^T, \gamma_t)^T$, in which the I_t are i.i.d. Bernoulli and (z_t^T, γ_t) are i.i.d. such that

$$\gamma_t \sim \text{Gamma}(g, 1/\lambda), \quad z_t|\gamma_t \sim \text{Normal}(z, V/(2\gamma_t)). \qquad (31)$$

Noting that (31) is a conjugate family and letting $\tau_t = (2\sigma_t^2)^{-1}$, [8] derives the follwing posterior distribution of θ_t given $\mathscr{Y}_{i,j}$ and $(K_t, \widetilde{K}_t) = (i, j)$:

$$\tau_t \sim \text{Gamma}(g_{i,j}, 1/\lambda_{i,j}), \quad z_t|\tau_t \sim \text{Normal}(\mu_{i,j}, V_{i.j}/(2\tau_t)),$$

in which $\mu_{i,j}$ and $V_{i,j}$ are given by (16), $g_{i,j} = g + (j - i + 1)/2$, and

$$\lambda_{i,j} = \lambda + z^T V^{-1} z + \sum_{t=i}^{j} y_t^2 - z_{i,j}^T V_{i,j}^{-1} z_{i,j}.$$

The filter $\theta_t|\mathscr{Y}_t$ and the smoother $\theta_t|\mathscr{Y}_n$ are given by (5) and (10), respectively, with

$$\pi_{0,0} = |V|^{-1/2}\lambda^g/\Gamma(g), \quad \pi_{i,j} = |V_{i,j}|^{-1/2}\lambda_{i,j}^{g_{i,j}}/\Gamma(g_{i,j}).$$

Incorporating a baseline state as in Section IIIA is important for array-CGH analysis. For speech signals and biological sequence analysis, the underlying Markov chain in the change-point HMM is more complicated than that in Section IIIA. These and other issues are currently under investigation.

C. Detection, Estimation and Control of Change-point ARX Models

There is an extensive literature on the problem of online detection of abrupt changes in stochastic systems; see [13],[14]. As pointed out in [14], there is a close connection between the theories of sequential hypothesis testing and quick detection of abrupt changes in stochastic systems subject to prespecified constraints on the false alarm rate or expected duration to false alarm. By using a change-of-measure argument and the law of large numbers for log-likelihood ratio statistics, an asymptotic lower bound for the detection delay in general stochastic systems subject to such constraints is derived in [15] and [16], where it is also shown how this lower bound can be asymptotically attained. When the pre- and post-change distributions are completely specified, this lower bound can be asymptotically attained by a likelihood-based CUSUM or moving average procedure. When the pre-change distribution is completely specified but the post-change distribution has unknown parameters, a window-limited generalized likelihood ratio (GLR) procedure, of the type introduced by Willsky and Jones [17] but with a suitably chosen window size, is shown to attain this lower bound asymptotically. Without assuming the baseline distribution to be specified, the Bayesian approach would put prior distributions on the pre- and post-change distributions. Since the change-time ν is unknown, it would also put a prior distribution on ν. A geometric prior distribution on ν would then bring us back to the hidden Markov model (1), with the obvious modification that θ_t enters an absorbing state at time ν in the classical change-detection framework that allows only a single change. By choosing p in the filter $\theta_t|\mathscr{Y}_t$ appropriately, the false alarm rate can be controlled at the prescribed level. Moreover, by restricting the number of components in the posterior mixture (5) for $\theta_t|\mathscr{Y}_t$ to $K(p)$ as indicated in the second paragraph of Section IIC, we can maintain the basic features of the window-limited GLR/mixture likelihood ratio detection rules in [15]. This is the basic idea behind the application of the hidden Markov filters to construct efficient sequential change-point detection rules in [9].

The idea of using on-line change-point detection methods to segment the data for recursive estimation and adaptive control has been considered in the literature; see e.g. [13], [18] and [19]. A major difficulty of this approach is that it does not incorporate the uncertainties in the segmentation (i.e., location of the change-points). Moreover, its performance depends on the choice of the detection rule and the trade-off between quick detection of change-points (related to estimation bias) and false alarm rate (related to the variances of the sequential estimates). For the change point autoregressive model with exogenous inputs (ARX model)

$$y_t + a_{1,t}y_{t-1} + \cdots + a_{k,t}y_{t-k} = b_{1,t}u_{t-1} + \cdots + b_{h,t}u_{t-h} + \sigma\varepsilon_t, \tag{19}$$

a commonly used approach is to use sliding windows or forgetting factors to modify the least squares estimate of $\theta_t \triangleq (a_{1,t},\ldots,a_{k,t},b_{1,t},\ldots,b_{h,t})^T$; see [20, pp. 140, 161]. Assuming the ε_t to be independent standard normal and $\theta_t = (1-I_t)\theta_{t-1} + I_t\eta_t$, with independent Bernoulli I_t and normal η_t, Chen and Lai [21] propose to use the hidden Markov model (HMM) filters (see Section IIC) as an alternative to the sliding-window or forgetting-factor least squares estimates. For on-line identification and adaptive control, they approximate the HMM filters by using a relatively small number of random samples drawn from the posterior distribution, which are sequentially generated over time by using a combination of importance sampling and resampling steps. They have applied these sequential Monte Carlo filters (also called "particle filters") to on-line identification of stable open-loop ARX systems and adaptive control of ARX models that are unstable in the open loop and have demonstrated the superiority of the HMM filters and certainty-equivalence control rules over the sliding-window or forgetting-factor least squares certainty-equivalence rules and recursive estimators.

Instead of using sequential Monte Carlo filters, [9] uses mixtures with a bounded number, $K = K(p)$, of components to approximate the HMM filter (5), which is a mixture of t components and whose computational complexity therefore increases to ∞ with t. These BCMIX filters, which are similar to those in Section IIC, are more tractable, not only computationally but also analytically, than the HMM filter (5) and its particle filter approximation. By making use of this more tractable structure of BCMIX filters, [9] develops an asymptotic theory of the APE criterion (17) for sequential determination of p and establishes the asymptotic efficiency of the corresponding BCMIX filters.

IV. Conclusion

We have shown that the hidden Markov model (1) provides a powerful approach to detection, segmentation, estimation and control in stochastic systems whose parameters may undergo occasional changes over time, with unknown change locations and magnitude. The model provides explicit recursive formulas for filters and smoothers to estimate the piecewise constant parameters. It leads to efficient segmentation schemes and gives confidence assessments of any given segmentation. Moreover, bounded-complexity approximations to the filters can be used for on-line detection and control.

References

[1] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory Probab. Appl.*, vol. 8, 1963, pp. 22-46.

[2] —— *Optimal Stopping Rules*. Springer-Verlag, NY: 1978.

[3] R. J. Elliott, F. Dufour, and W. P. Malcolm, "State and model estimation for discrete time jump Markov systems," *SIAM J. Contr. Optimiz.*, vol. 44, 2005, pp. 1081-1104.

[4] Y. Yao, "Estimation of a noisy discrete-time step functions: Bayes and empirical Bayes approach," *Ann. Statist*, vol. 12, 1984, pp. 1434-1447.

[5] H. Chernoff and S. Zacks. "Estimating the current mean of a normal distribution which is subjected to change in time," *Ann. Math. Statist.*, vol. 35, 1964, pp. 999-1018.

[6] T. L. Lai and H. Xing, "A Bayesian approach to multiple change-points," Technical Report, Department of Statistics, Stanford University, 2007.

[7] J. Rissanen, "Order estimation by accumulated prediction errors", In *Essays in Time Series and Applied Processes*, Special volume 23A of *J. Appl. Probab.*, 1986, pp. 55-61.

[8] T. L. Lai, H. Liu, and H. Xing, "Autoregressive models with piecewise constant volatility and regression parameters", *Statist. Sinica*, vol. 15, 2005, pp. 279-301.

[9] T. L. Lai and H. Xing, "Change-point detection, parameter estimation and adaptive control via bounded complexity mixture filters," Technical Report, Department of Statistics, Stanford University, 2007.

[10] T. L. Lai, H. Xing, and N. Zhang, "Stochastic segmentation models for array-based comparative genomic hybridization data analysis," *Biostatistics*, vol. 9, 2008, 290-307.

[11] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust. Speech Signal Process*, vol. 36, Jan. 1988, 29-40.

[12] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Processing*, vol. 50, 2002, 747-758.

[13] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes, Theory and Applications*. Englewood Cliffs, NJ:Prentice-Hall, 1993.

[14] T. L. Lai, "Sequential changepoint detection in quality control and dynamical systems," *J. R. Statist. Soc.* B, vol 57, 1995, pp. 613-658.

[15] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inform. Theory*, vol. 44, 1998, pp 2917-2929.

[16] T. L. Lai and J. Z. Shan, "Efficient recursive algorithms for detection of abrupt changes in signals and control systems," *IEEE Trans. Automat. Contr.*, vol. 44, 1999, pp 952-966.

[17] A. S. Willsky and H. L. Jones, "A generalized likelihood ratio approach to detection and estimation of jumps in linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-21, 1976, pp. 108-112.

[18] M. H. A. Davis, "The application of nonlinear filtering to fault detection in linear systems," *IEEE Trans. Automat. Contr.*, vol. 20, Apr. 1975, pp. 257-259.

[19] N. M. Filatov and H. Unbehauen, "Application of improved adaptive controllers to a plant with jumps in parameters," *Proc. 1998 IEEE Conference on Control Applications*, vol. 2, Sept. 1998, pp.1343-1347.

[20] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, New York: Springer-Verlag, 1987.

[21] Y. Chen and T. L. Lai, "Identification and adaptive control of change-point ARX models via Rao-Blackwellized particle filters," *IEEE Trans. Automat. Contr.*, vol. 52, Jan. 2007, pp. 67-72.