Time Discretization of Continuous-Time Filters and Smoothers for HMM Parameter Estimation

Matthew R. James, *Member, IEEE*, Vikram Krishnamurthy, *Member, IEEE*, and François Le Gland, *Member, IEEE*

Abstract—**In this paper we propose algorithms for parameter estimation of fast-sampled homogeneous Markov chains observed in white Gaussian noise. Our algorithms are obtained by the robust discretization of stochastic differential equations involved in the estimation of continuous-time Hidden Markov Models (HMM's) via the EM algorithm. We present two algorithms: The first is based on the robust discretization of continuous-time filters that were recently obtained by Elliott to estimate quantities used in the EM algorithm. The second is based on the discretization of continuous-time smoothers, yielding essentially the well-known Baum–Welch re-estimation equations. The smoothing formulas for continuous-time HMM's are new, and their derivation involves two-sided stochastic integrals. The choice of discretization results in equations which are identical to those obtained by deriving the results directly in discrete time. The filter-based EM algorithm has negligible memory requirements; indeed, independent of the number of observations. In comparison the smoother-based discrete-time EM algorithm require the use of the forward–backward algorithm, which is a fixed-interval smoothing algorithm and has memory requirements proportional to the number of observations. On the other hand, the computational complexity of the filter-based EM algorithm is greater than that of the smoother-based scheme. However, the filters may be suitable for parallel implementation. Using computer simulations we compare the smoother-based and filter-based EM algorithms for HMM estimation. We provide also estimates for the discretization error.**

Index Terms— **Hidden Markov Models, robust discretization, expectation maximization algorithm, parameter estimation.**

I. Introduction

In this paper we propose algorithms for parameter estimation of fast-sampled homogeneous Markov chains observed in white Gaussian noise. The parameters estimated include transition probabilities and levels (drift coefficients) of the Markov chain, and the noise variance. Our algorithms are obtained by the robust discretization of stochastic differential equations involved in the estimation of continuous-time Hidden Markov Models (HMM's) via the EM (Expectation–Maximization) algorithm. The EM algorithm is an iterative ML (Maximum-Likelihood) parameter estimation scheme that can be used to estimate the parameters of Markov processes observed in white Gaussian noise, see [1], [3], [8], and [12].

The contributions of this paper can be outlined as follows:

1) **Parameter Estimation Algorithms**: We present two algorithms to estimate the parameters of the HMM: The first is based on the robust discretization (see below) of continuous-time *filters* that were recently obtained by Elliott [7] to estimate quantities used in the EM algorithm. We term this the *filter-based* EM scheme. The second is based on the robust discretization of continuous-time *smoothers*, yielding essentially the well-known Baum–Welch re-estimation equations. We term this the *smoother-based* EM scheme.

It turns out that the filter-based scheme has negligible memory requirements compared to the smoother-based scheme. Using a T/Δ-length noisy observation sequence of an N-state Markov chain, where Δ is the time step size, the smoother-based algorithm requires a memory of NT/Δ, whereas the filter-based algorithm requires memory independent of T/Δ. However, the computational complexity of the filter-based EM algorithm at each time instant is $O(N^4)$ (for an N-state Markov chain) compared to $O(N^2)$ for the smoother-based scheme. Despite the higher computational cost, the various filters in the filter-based scheme are decoupled and are suitable for parallel implementation on a multiprocessor system.

The continuous-time smoother based EM scheme that we present is new and its derivation involves two-sided stochastic integrals.

2) **Robust Discretization**: We perform the robust discretization mentioned above as follows: First, by using the approach due to Clark [2], we derive the *robust* versions of the differential equations that compute various quantities required in the EM algorithm. By robust, we mean that the differential equations define versions of the filters which depend *continuously* on the observation path. This is a useful property from the practical point of view, see Clark [2]. The discrete-time algorithms are obtained by discretizing the resulting robust filters. We provide also estimates for the discretization error.

3) **Probabilistic Interpretation**: We give a probabilistic interpretation to our numerical schemes. In particular, the time-discretization is chosen to yield equations which are identical to the filtering and smoothing equations for a discrete-time HMM which are obtained in Elliott [6] and Levinson, Rabiner,

Manuscript received January 12, 1994; revised January 19, 1995. This research was supported in part by the Cooperative Research Centre for Robust and Adaptive Systems.The material in this paper was presented in part at the IEEE CDC, 1992.

M. R. James is with the Department of Engineering, Faculty of Engineering and Information Technology, Australian National University, Canberra, ACT 0200, Australia.

V. Krishnamurthy is with the Department of Electrical and Electronic Engineering, University of Melborune, Parkville, VIC 3052, Australia.

F. Le Gland is with the Institut de Recherche en Informatique et Systèmes Aléatoires/Institut National de Recherche en Informatique et en Automatique, Campus de Beaulieu, 35042 Rennes Cédex, France.

Publisher Item Identifier S 0018-9448(96)01031-0.

and Sondhi [8], repectively. This is an important consistency property in that it links discretized continuous-time results with discrete-time results. We emphasize that other choices in the discretisation may lead to other recursions which do not correspond to standard discrete-time HMM formulas.

4) **Computer Simulations**: Using computer simulations we compare the filter-based and smoother-based EM algorithms. Both algorithms yielded satisfactory estimates in our simulations; however, we found that the smoother-based scheme had better numerical properties than the filter-based scheme. Important implementation aspects such as normalization are also considered.

The paper is organized as follows: In Section II we present our continuous-time robust filters. In Section III the discretized filters are derived. Also pathwise error estimates are obtained. In Sections IV and V the smoothing analogs of Sections II and III are presented. Section VI deals with important implementation issues like normalization, and in addition, the estimation of the noise variance is discussed. In Section VII, simulation examples are presented that compare the smoother-based and filter-based algorithms.

II. Continuous-Time HMM Estimation (Filtering)

In this section we first briefly review the EM algorithm (Section II-A). We then describe the continuous-time model in Section II-B and review the continuous-time filters derived in [7] in Section II-C. In Section II-D we derive robust versions of the filters.

A. Review of EM Algorithm

The basic idea behind the EM algorithm is as follows [5]. Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathcal{F})$ all absolutely continuous with respect to a fixed probability measure P_0, and let $\mathcal{Y} \subset \mathcal{F}$. The likelihood function for computing an estimate of the parameter θ based on the information available in $\mathcal{Y}$ is

$$L(\theta) = \boldsymbol{E}_0 \left[\frac{dP_\theta}{dP_0} \middle| \mathcal{Y} \right]$$

and the MLE is defined by

$$\hat{\theta} \in \operatorname*{argmax}_{\theta \in \Theta} L(\theta).$$

In general, the MLE is difficult to compute directly. The EM algorithm provides an iterative approximation method starting from an initial model estimate $\hat{\theta}_0$. Each iteration of the EM algorithm consists of two steps:

Step 1 (E-Step): Set $\theta = \hat{\theta}_p$ and compute $Q(\cdot, \theta)$, where

$$Q(\theta', \theta) = \boldsymbol{E}_\theta \left[\log \frac{dP'_\theta}{dP_\theta} \middle| \mathcal{Y} \right].$$

Step 2 (M-Step): Find

$$\hat{\theta}_{p+1} \in \operatorname*{argmax}_{\theta' \in \Theta} Q(\theta', \theta).$$

The sequence generated $\{\hat{\theta}_p, p \geq 0\}$ gives nondecreasing values of the likelihood function with equality if and only if $\hat{\theta}_{p+1} = \hat{\theta}_p$.

B. Continuous-Time Model

Let $\{X_t, t \geq 0\}$ be a continuous-time Markov chain defined on a probability space $(\Omega, \mathcal{F}, \boldsymbol{P})$ with state space $S = \{e_1, e_2, \cdots, e_N\}$. Without loss of generality, we assume that e_i is the unit column vector of $\boldsymbol{R}^N$ with 1 in the ith position. Let $\langle \cdot, \cdot \rangle$ denote the scalar product in $\boldsymbol{R}^N$. If $u = (u_1, \cdots, u_N)$, then

$$\langle X_t, u \rangle = \sum_{i=1}^{N} u_i \, 1_{(X_t = e_i)}.$$

Let π_0 be the probability distribution of X_0, and $A = (a_{ij})$ be the transition rate matrix (infinitesimal generator), i.e.

$$P(X_{t+h} = e_j | X_t = e_i) = \delta_{ij} + a_{ij} h + o(h). \tag{2.1}$$

We assume that X_t is not directly observed, instead we observe the scalar process

$$y_t = \int_0^t \langle X_r, g \rangle \, dr + w_t \tag{2.2}$$

where $\{w_t, t \geq 0\}$ is a standard Brownian motion (unit variance) on $(\Omega, \mathcal{F}, P)$, which is independent of $\{X_t, t \geq 0\}$. The case of a d-dimensional observation could also be considered, and we assume $d = 1$ only for the sake of simplicity. Also $g = (g_1, \cdots, g_N)$ are the *levels* or drift coefficients of the Markov chain. Write $\mathcal{F}_t = \sigma(X_s, y_s, 0 \leq s \leq t)$ and $\mathcal{Y}_t = \sigma(y_s, 0 \leq s \leq t)$.

The aim of the estimation problem is to obtain the MLE for the unknown parameters A and g. A filtering approach for doing so using the EM algorithm is presented in [1] and [7], see also [3], [14]. In the EM algorithm, updating the estimates of A and g requires computation of the conditional expectations of the following quantities given the observation history:

1) State of the Markov chain.
2) Occupation time of the Markov chain in state e_i until time t:
$$J_t^i = \int_0^t \langle X_s, e_i \rangle \, ds.$$
3) Number of jumps of the Markov chain from state e_i to state e_j until time t:
$$N_t^{ij} = \int_0^t \langle X_{s-}, e_i \rangle \, \langle dX_s, e_j \rangle \quad \text{for } i \neq j.$$
4) Level integral in state e_i up to time t:
$$G_t^i = \int_0^t \langle X_s, e_i \rangle \, dy_s.$$

The update from A, g to A', g' is given by

$$a'_{ij} = \frac{\boldsymbol{E}[N_T^{ij} | \mathcal{Y}_T]}{\boldsymbol{E}[J_T^i | \mathcal{Y}_T]} \quad \text{for } i \neq j,$$

and

$$g'_i = \frac{\boldsymbol{E}[G_T^i | \mathcal{Y}_T]}{\boldsymbol{E}[J_T^i | \mathcal{Y}_T]} \tag{2.3}$$

where the conditional expectations are computed using the parameters A and g. In this way, a sequence of parameter estimates is generated which gives nondecreasing values of the likelihood function.

Notation: For any $\{\mathcal{F}_t, t \geq 0\}$-adapted and integrable process $\{H_t, t \geq 0\}$, denote the unnormalized conditional expectations as

$$\sigma(H_t) = \overline{\boldsymbol{E}}[H_t \Lambda_t | \mathcal{Y}_t] \tag{2.4}$$

where $\overline{P}$ is a probability measure on $(\Omega, \mathcal{F})$ defined by setting the Radon–Nikodym derivative

$$\left.\frac{dP}{d\overline{P}}\right|_{\mathcal{F}_t} = \Lambda_t = \exp\left\{\int_0^t \langle X_s, g\rangle \, dy_s - \frac{1}{2}\int_0^t \langle X_s, g\rangle^2 \, ds\right\}.$$

By the Bayes rule we have

$$\boldsymbol{E}[H_t | \mathcal{Y}_t] = \sigma(H_t)/\sigma(1).$$

In the particular case where $H_t \equiv 1$, we use the notation $p_t = \sigma(X_t)$.

Finally, let B denote the diagonal matrix

$$B = \operatorname{diag}(g_1, \cdots, g_N).$$

Note that $Be_i = g_i e_i$.

C. Continuous-Time Filters

From (2.3) above, our objective is to compute $\sigma(H_T)$ for $H_T = N_T^{ij}, G_T^i$ or J_T^i. It is not possible in general to obtain equations directly for $\sigma(H_t)$, but it is possible to obtain equations for $\sigma(H_t X_t)$, i.e., for the N-dimensional vector whose ith component is

$$\langle \sigma(H_t X_t), e_i \rangle = \overline{\boldsymbol{E}}(1_{(X_t = e_i)} H_t \Lambda_t | \mathcal{Y}_t).$$

One would then obtain $\sigma(H_T)$ as $\sigma(H_T) = \langle \sigma(H_T X_T), \mathbf{1}\rangle$ where $\mathbf{1}$ denotes the column N-vector of ones. The equations for $\sigma(H_t X_t)$ are presented in Elliott [7]. Note that the filtering equation (2.7) for the number of jumps was first obtained in Zeitouni and Dembo [14].

State (Wonham filter) [13]:

$$p_t = \pi_0 + \int_0^t A^* p_s \, ds + \int_0^t B p_s \, dy_s. \tag{2.5}$$

Occupation Time:

$$\sigma(J_t^i X_t) = \int_0^t A^* \sigma(J_s^i X_s) \, ds + \int_0^t B\sigma(J_s^i X_s) \, dy_s + \int_0^t \langle p_s, e_i\rangle e_i \, ds. \tag{2.6}$$

Number of Jumps:

$$\sigma(N_t^{ij} X_t) = \int_0^t A^* \sigma(N_s^{ij} X_s) \, ds + \int_0^t B\sigma(N_s^{ij} X_s) \, dy_s + \int_0^t \langle p_s, e_i\rangle \langle A^* e_i, e_j\rangle e_j \, ds. \tag{2.7}$$

Note that in (2.7), $\langle A^* e_i, e_j\rangle = a_{ij}$.

Level Integrals:

$$\sigma(G_t^i X_t) = \int_0^t A^* \sigma(G_s^i X_s) \, ds + \int_0^t B\sigma(G_s^i X_s)) \, dy_s + \int_0^t \langle p_s, e_i\rangle B e_i \, ds + \int_0^t \langle p_s, e_i\rangle e_i \, dy_s. \tag{2.8}$$

Note that in (2.8), $Be_i = g_i e_i$.

Remark 2.1: The re-estimation formulas (2.3) read now

$$a'_{ij} = \frac{\langle \sigma(N_T^{ij} X_T), \mathbf{1}\rangle}{\langle \sigma(J_T^i X_T), \mathbf{1}\rangle}$$

and

$$g'_i = \frac{\langle \sigma(G_T^i X_T), \mathbf{1}\rangle}{\langle \sigma(J_T^i X_T), \mathbf{1}\rangle}.$$

D. Robust Filters

In [2], Clark introduced *robust* reformulations of the nonlinear filtering equations, and showed that the conditional probability distribution has a version which depends continuously on the observations. From a practical point of view, this continuous dependence is a desirable robustness property, and leads to robust approximations.

In this section, we follow Clark's approach and define robust versions of the filters obtained by Elliott [7]. These will be used in the next section to derive robust numerical algorithms.

We introduce the processes

$$\begin{aligned} \phi_t^i &= \exp\{g_i y_t - \tfrac{1}{2}|g_i|^2 t\} \\ \Phi_t &= \operatorname{diag}(\phi_t^1, \cdots, \phi_t^N) = \exp\{B y_t - \tfrac{1}{2} B^2 t\}. \end{aligned} \tag{2.9}$$

Note that $\Phi_t e_i = \phi_t^i e_i$. For any $\{\mathcal{F}_t, t \geq 0\}$-adapted and integrable process $\{H_t, t \geq 0\}$, we define

$$\overline{\sigma}(H_t X_t) = \Phi_t^{-1} \sigma(H_t X_t). \tag{2.10}$$

State: Using Ito's rule, one can show that $\overline{p}_t = \overline{\sigma}(X_t)$ is a process of finite variation and solves the ODE

$$\frac{d}{dt}\overline{p}_t = \Phi_t^{-1} A^* \Phi_t \overline{p}_t \tag{2.11}$$

with initial condition $\overline{p}_0 = \pi_0$. Equation (2.11) was derived in [2], where it was shown that $\overline{p}_t$ is a locally Lipschitz continuous function of $(y(s), 0 \leq s \leq t)$, and (2.11) can be used to define a version of the conditional probability distribution which enjoys also this continuity property.

In the same way, robust versions $\overline{\sigma}(H_t X_t)$ of the filters $\sigma(H_t X_t)$ for the processes $H_t = J_t^i, N_t^{ij}$ and G_t^i can be obtained. They read:

Occupation Time:

$$\frac{d}{dt}\overline{\sigma}(J_t^i X_t) = \Phi_t^{-1} A^* \Phi_t \overline{\sigma}(J_t^i X_t) + \langle \overline{p}_t, e_i\rangle e_i \tag{2.12}$$

with initial condition $\overline{\sigma}(J_0^i X_0) = 0$.

Number of Jumps:

$$\frac{d}{dt}\overline{\sigma}(N_t^{ij} X_t) = \Phi_t^{-1} A^* \Phi_t \overline{\sigma}(N_t^{ij} X_t) + \langle \overline{p}_t, e_i\rangle \langle \Phi_t^{-1} A^* \Phi_t e_i, e_j\rangle e_j \tag{2.13}$$

with initial condition $\overline{\sigma}(N_0^{ij} X_0) = 0$. Note that, in (2.13), $\langle \Phi_t^{-1} A^* \Phi_t e_i, e_j\rangle = a_{ij}\phi_t^i/\phi_t^j$.

Level Integrals:

$$d\overline{\sigma}(G_t^i X_t) = \Phi_t^{-1} A^* \Phi_t \overline{\sigma}(G_t^i X_t) \, dt + \langle \overline{p}_t, e_i\rangle e_i \, dy_t \tag{2.14}$$

with initial condition $\overline{\sigma}(G_0^i X_0) = 0$.

Note that $\overline{\sigma}(J_t^i X_t)$ and $\overline{\sigma}(N_t^{ij} X_t)$ are finite variation processes solving ODE's, while $\overline{\sigma}(G_t^i X_t)$ is not a finite variation

process, and is the solution of an SDE. However, using integration by parts, the stochastic integral in (2.14) can be written in terms of a standard integral, see (2.17) below.

The following theorem shows that these differential equations define versions of the filters which depend continuously on the observation path. Let

$$\|y\| \triangleq \sup_{0\le t\le T} |y(t)|$$

denote the sup-norm of $(y(t), 0 \le t \le T)$.

Theorem 2.2: For $H_t = J_t^i, N_t^{ij}$ or G_t^i, define $\overline{\sigma}(H_tX_t)$ via (2.11)–(2.14). Then, for all $0 \le t \le T$

$$\pi(H_tX_t) \triangleq \frac{\Phi_t\overline{\sigma}(H_tX_t)}{\langle \Phi_t\overline{p}_t, \mathbf{1}\rangle}$$

defines a locally Lipschitz continuous version of $\boldsymbol{E}[H_tX_t|\mathcal{Y}_t]$

$$|\pi(H_tX_t)[y_1] - \pi(H_tX_t)[y_2]| \le K\|y_1 - y_2\| \qquad (2.15)$$

where the constant K depends on $\|y_1\|$ and $\|y_2\|$.

Proof: For $H_t = J_t^i$, N_t^{ij}, or G_t^i, define $\sigma(H_tX_t)$ by

$$\sigma(H_tX_t) = \Phi_t\overline{\sigma}(H_tX_t)$$

where $\overline{\sigma}(H_tX_t)$ is defined by the robust equations (2.11)–(2.14). Then by Ito's rule it follows that $\sigma(H_tX_t)$ is a solution of the corresponding SDE given in Section II-D. These equations have unique solutions; namely, the corresponding unnormalized conditional expectations. Therefore, after normalizing we have $\boldsymbol{E}[H_tX_t|\mathcal{Y}_t] = \pi(H_tX_t)[y]$ a.s.

To prove the local Lipschitz continuity assertion, we follow Clark [2, Theorem 4], where the following inequalities were proven:

$$|\overline{p}_t[y_1] - \overline{p}_t[y_2]| \le K\|y_1 - y_2\|$$

for some constant K depending on $\|y_1\|$ and $\|y_2\|$, and

$$\langle \Phi_t[y]\overline{p}_t[y], \mathbf{1}\rangle \ge \gamma > 0$$

for all $t \ge 0$, where γ depends on $\|y\|$. These inequalities imply that $\pi_t = \pi(X_t)$ satisfies (2.15) (with $H_t \equiv 1$).

For $H_t = J_t^i$ or N_t^{ij}, one can use the same method as in [2] to obtain

$$|\overline{\sigma}(H_tX_t)[y_1] - \overline{\sigma}(H_tX_t)[y_2]| \le K\|y_1 - y_2\| \qquad (2.16)$$

from which (2.15) follows.

For $H_t = G_t^i$, one has to take into account the stochastic integral in (2.14). However, since $\overline{p}_t = \overline{\sigma}(X_t)$ is a finite variation process, the stochastic integral can be rewritten as

$$\int_0^t \langle \overline{p}_s, e_i\rangle\, dy_s = y_t\langle \overline{p}_t, e_i\rangle - \int_0^t y_s\langle \Phi_s^{-1}A^*\Phi_s\overline{p}_s, e_i\rangle\, ds. \qquad (2.17)$$

If we denote the right-hand side of this equation by F_t^i, then clearly

$$|F_t^i[y_1] - F_t^i[y_2]| \le K\|y_1 - y_2\|$$

where K depends on $\|y_1\|$ and $\|y_2\|$. Therefore, we can write

$$\overline{\sigma}(G_t^iX_t)[y] = \int_0^t \Phi_s^{-1}[y]A^*\Phi_s[y]\overline{\sigma}(G_s^iX_s)[y]\, ds + F_t^i[y]$$

and hence (2.15) follows by standard arguments based on the Gronwall lemma. □

III. Time Discretization of Filtering Equations

The purpose of this section is to provide computable approximation of the continuous time equations described in Section II-C, for $p_t = \sigma(X_t), \sigma(J_t^iX_t), \sigma(G_t^iX_t)$, and $\sigma(N_t^{ij}X_t)$.

Throughout the paper, a regular partition

$$0 = t_0 < t_1 < \cdots < t_{n-1} < t_n < \cdots$$

is considered, with constant time step $\Delta = t_n - t_{n-1}$. Write $M = [T/\Delta]$ for the largest integer such that $M\Delta \le T$.

Basically, two different approaches are available to obtain discrete filtering equations.

1) One approach is to sample the continuous time observations $\{y_t, t \ge 0\}$ and approximate the original continuous-time HMM: the filtering equations for the discrete-time HMM would then provide an approximation of the filtering equations for the continuous-time HMM.
2) The other approach is to directly discretize the filtering equations, or their robust versions obtained in Section II-D.

As far as state estimation is concerned, it is shown in Clark [2] that reasonable discretization schemes of the robust equation (2.11) could also provide discretization schemes for the corresponding Duncan–Mortensen–Zakai equation (2.5), and give rise to interesting probabilistic interpretation, thus linking the two above mentioned approaches.

The purpose here is to review the results of [2] for the state estimation problem, and extend these results to the HMM parameter estimation problem. Following [2], a reasonable approximation of (2.11) between sampling times t_{n-1} and t_n, would look like

$$\begin{aligned}\overline{p}_{t_n} &= \overline{p}_{t_{n-1}} + \int_{t_{n-1}}^{t_n} \Phi_s^{-1}A^*\Phi_s\overline{p}_s\, ds \\ &\simeq \overline{p}_{t_{n-1}} + \Phi_{t_n'}^{-1}A^*\Phi_{t_n'}\overline{p}_{t_n''}\Delta\end{aligned}$$

for some $t_{n-1} \le t_n', t_n'' \le t_n$. Various approximations can be obtained, for different choices of t_n' and t_n''. Indeed, we will make choices which result in the approaches 1) and 2) mentioned above coinciding, and yielding standard discrete-time formulas. Choosing $t_n' = t_n'' = t_{n-1}$ gives

$$\overline{p}_{t_n} \simeq [I + \Delta\Phi_{t_{n-1}}^{-1}A^*\Phi_{t_{n-1}}]\overline{p}_{t_{n-1}}$$

which results in the following explicit approximation:

$$\overline{p}_n = [I + \Delta\Phi_{t_{n-1}}^{-1}A^*\Phi_{t_{n-1}}]\overline{p}_{n-1}. \qquad (3.1)$$

Multiplying both sides by Φ_{t_n} gives the following approximation for the Duncan–Mortensen–Zakai equation (2.5):

$$p_n = \Phi_{t_n}\Phi_{t_{n-1}}^{-1}[I + \Delta A^*]p_{n-1} = \Psi_n[I + \Delta A^*]p_{n-1} \qquad (3.2)$$

where

$$\Psi_t^s = \Phi_t \Phi_s^{-1} = \exp\{B[y_t - y_s] - \tfrac{1}{2}B^2[t-s]\}$$

and

$$\Psi_n = \Psi_{t_n}^{t_{n-1}}.$$

Note that

$$\Psi_n = \Phi_{t_n}\Phi_{t_{n-1}}^{-1} = \operatorname{diag}(\psi_n^1, \cdots, \psi_n^N)$$

with $\psi_n^i = \phi_{t_n}^i / \phi_{t_{n-1}}^i$. Other schemes could be obtained for different choices of t_n' and t_n''.

A Discrete-Time Approximate Model

Note that, for a small enough time step $\Delta > 0$, $P = [I + \Delta A] = (\pi_{ij})$ is a stochastic matrix. Define then the following fast-sampled observations:

$$z_n^\Delta = \frac{1}{\Delta}[y_{t_n} - y_{t_{n-1}}]$$

and the following discrete-time HMM to be used throughout the paper:

$\{X_n, n \geq 0\}$ is a Markov chain with state space $S = \{e_1, \cdots, e_N\}$ and transition probability matrix P, related to the observation sequence $\{z_n^\Delta, n \geq 0\}$ through the equation

$$z_n^\Delta = g(X_n) + w_n^\Delta$$

where $\{w_n^\Delta, n \geq 0\}$ is a Gaussian white noise sequence, with covariance matrix $\Delta^{-1}I$.

It is then straightforward to check that the approximation scheme (3.2) is exactly the filtering equation (Baum's forward equation) for the state estimation in this discrete-time HMM.

Let us write

$$\mathcal{F}_n = \sigma(X_k, z_k^\Delta, 0 \leq k \leq n)$$

and

$$\mathcal{Z}_n = \sigma(z_k^\Delta, 0 \leq k \leq n).$$

In the EM algorithm for the discrete-time model above, updating the estimates of P and g requires computation of the conditional expectations of the following quantities given the observation history:

1) State of the Markov chain.
2) Number of visits of the Markov chain in state e_i until time n

$$J_n^i = \sum_{k=1}^n \langle X_{k-1}, e_i \rangle.$$

3) Number of transitions of the Markov chain from state e_i to state e_j until time n

$$N_n^{ij} = \sum_{k=1}^n \langle X_{k-1}, e_i \rangle \langle X_k, e_j \rangle.$$

4) Level sum in state e_i up to time n

$$G_n^i = \sum_{k=1}^n \langle X_{k-1}, e_i \rangle z_k^\Delta.$$

The update from P, g to P', g' is given by

$$\pi_{ij}' = \frac{\boldsymbol{E}[N_M^{ij} | \mathcal{Z}_M]}{\boldsymbol{E}[J_M^i | \mathcal{Z}_M]}$$

and

$$g_i' = \frac{\boldsymbol{E}[G_M^i | \mathcal{Z}_M]}{\boldsymbol{E}[J_M^i | \mathcal{Z}_M]} \tag{3.3}$$

where the conditional expectations are computed using the parameters P and g. In this way, a sequence of parameter estimates is generated which gives nondecreasing values of the likelihood function for the discrete-time HMM.

Remark 3.1: The main advantage of the approach adopted here for time discretization is that the sequence of parameter estimates generated by the re-estimation formulas (3.3) will automatically converge to a stationary point of the likelihood function for the discrete-time model. Therefore, provided the likelihood function for the discrete-time model is close enough to the likelihood function for the original continuous-time model, the sequence of parameter estimates generated by the re-estimation formulas (3.3) will be reasonably close to a stationary point of the likelihood function for the original continuous-time model.

In discrete time, the conditional expectations involved in the EM algorithm (3.3) are traditionally computed using smoothing (Baum–Welch re-estimation equations), rather than filtering. For the purpose of comparison, in this section we consider the use of filtering in the discrete-time EM algorithm (see [1], where such a comparison is made for diffusion processes). Time-discretized numerical schemes are obtained for the continuous-time filtering equations. Following [2], error estimates are provided using the robust filters.

B. Discrete-Time Filters

Based on the previous remarks, the approach adopted below to discretize the filtering equations for $\sigma(J_t^i X_t)$, $\sigma(N_t^{ij} X_t)$ and $\sigma(G_t^i X_t)$, is to use the corresponding filtering equations for the approximate discrete time HMM introduced above. The filtering equations for parameter estimation in discrete-time HMM are derived in Elliott [6].

State: See (3.2), to be compared with (2.5). The computational cost is $O(N^2)$ at each time instant.

Occupation Time: The filter $\sigma(J_{t_n}^i X_{t_n})$ for the occupation time in state e_i in the continuous time HMM, is approximated by Δ times the filter $\sigma(J_n^i X_n)$ for the number of visits in state e_i in the discrete time HMM. The equation for the filter $\sigma(J_n^i X_n)$ is (where p_n is defined by (3.2))

$$\sigma(J_n^i X_n) = \Psi_n P^* \sigma(J_{n-1}^i X_{n-1}) + \langle p_{n-1}, e_i \rangle \Psi_n P^* e_i \tag{3.4}$$

where $P^* = [I + \Delta A^*]$. Multiplying both sides of (3.4) by $\Phi_{t_n}^{-1}$ gives

$$\begin{aligned}\overline{\sigma}(J_n^i X_n) = {} & [I + \Delta \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}}] \overline{\sigma}(J_{n-1}^i X_{n-1}) \\ & + \langle \overline{p}_{n-1}, e_i \rangle [I + \Delta \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}}] e_i \end{aligned} \tag{3.5}$$

to be compared with (2.12). The computational cost is $O(N^3)$ at each time instant.

Number of Transitions: The filter $\sigma(N^{ij}_{t_n}X_{t_n})$ for the number of jumps from state e_i to state e_j in the continuous-time HMM, is approximated by the filter $\sigma(N^{ij}_n X_n)$ for the number of transitions from state e_i to state e_j in the discrete-time HMM. The equation for the filter $\sigma(N^{ij}_n X_n)$ is

$$\sigma(N^{ij}_n X_n) = \Psi_n P^* \sigma(N^{ij}_{n-1} X_{n-1}) + \langle p_{n-1}, e_i\rangle \langle \Psi_n P^* e_i, e_j\rangle e_j. \quad (3.6)$$

Multiplying both sides of (3.6) by $\Phi^{-1}_{t_n}$ gives

$$\overline{\sigma}(N^{ij}_n X_n) = [I + \Delta\Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}}]\overline{\sigma}(N^{ij}_{n-1}X_{n-1}) + \Delta\langle \overline{p}_{n-1}, e_i\rangle \langle \Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}} e_i, e_j\rangle e_j \quad (3.7)$$

to be compared with (2.13). The computational cost is $O(N^4)$ at each time instant.

Level Integral: The filter $\sigma(G^i_{t_n}X_{t_n})$ for the level integral in state e_i in the continuous-time HMM, is approximated by Δ times the filter $\sigma(G^i_n X_n)$ for the level sum in state e_i in the discrete-time HMM. The equation for the filter $\sigma(G^i_n X_n)$ is

$$\sigma(G^i_n X_n) = \Psi_n P^* \sigma(G^i_{n-1}X_{n-1}) + z^{\Delta}_n \langle p_{n-1}, e_i\rangle \Psi_n P^* e_i. \quad (3.8)$$

Multiplying both sides of (3.8) by $\Phi^{-1}_{t_n}$ gives

$$\overline{\sigma}(G^i_n X_n) = [I + \Delta\Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}}]\overline{\sigma}(G^i_{n-1}X_{n-1}) + z^{\Delta}_n \langle \overline{p}_{n-1}, e_i\rangle [I + \Delta\Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}}]e_i \quad (3.9)$$

to be compared with (2.14). The computational cost is $O(N^3)$ at each time instant.

Remark 3.2: The re-estimation formulas (3.3) read now

$$\pi'_{ij} = \frac{\langle \sigma(N^{ij}_M X_M), \mathbf{1}\rangle}{\langle \sigma(J^i_M X_M), \mathbf{1}\rangle}$$

and

$$g'_i = \frac{\langle \sigma(G^i_M X_M), \mathbf{1}\rangle}{\langle \sigma(J^i_M X_M), \mathbf{1}\rangle}.$$

Direct implementation of the filters for the occupation time, number of jumps, and level integrals requires, respectively, $O(N^3)$, $O(N^4)$, and $O(N^3)$ multiplications at each time instant. However, notice that the equations for $\sigma(J^i_n X_n)$, $\sigma(N^{ij}_n X_n)$, and $\sigma(G^i_n X_n)$ are all decoupled, and hence can be solved in parallel for each $i, j = 1, \cdots, N$.

Pathwise error estimates can now be obtained in a way similar to [2, Theorem 7]. This is the purpose of the remainder of this section.

C. Pathwise Error Estimates

As above, let $\|y\|$ denote the sup-norm of $(y(t), 0 \le t \le T)$, and let

$$\omega_\Delta(y) = \max\{|y(t) - y(s)|: 0 \le s, t \le T, |t - s| \le \Delta\}$$

denote the *modulus of continuity*.

Theorem 3.3: For $H_t = J^i_t, N^{ij}_t$ or G^i_t, define $\overline{\sigma}(H_t X_t)$ via (2.11)–(2.14). Similarly, for $H_n = \Delta \cdot J^i_n, N^{ij}_n$ or $\Delta \cdot G^i_n$, define $\overline{\sigma}(H_n X_n)$ via (3.1), (3.5), (3.7), and (3.9). Then, for all n, Δ such that $0 \le t_n \le T$

$$|\overline{\sigma}(H_{t_n}X_{t_n})[y] - \overline{\sigma}(H_n X_n)[y]| \le K[\Delta + \omega_\Delta(y)] \quad (3.10)$$

where the constant K depends on $\|y\|$.

Proof: The result is proved in [2, Theorem 4] for $H_t \equiv 1$ and $H_n \equiv 1$.

Let us first consider the case $H_t = J^i_t$ and $H_n = \Delta \cdot J^i_n$. We introduce the notations $\overline{\sigma}_t = \overline{\sigma}(J^i_t X_t)$, and $\overline{\sigma}_n = \Delta \cdot \overline{\sigma}(J^i_n X_n)$. From (2.12)

$$\overline{\sigma}_{t_n} = \overline{\sigma}_{t_{n-1}} + \int_{t_{n-1}}^{t_n} \Phi^{-1}_s A^* \Phi_s \overline{\sigma}_s \, ds + \int_{t_{n-1}}^{t_n} \langle \overline{p}_s, e_i\rangle e_i \, ds$$

whereas from (3.5)

$$\overline{\sigma}_n = [I + \Delta\Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}}]\overline{\sigma}_{n-1} + \Delta\langle \overline{p}_{n-1}, e_i\rangle \cdot [I + \Delta\Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}}]e_i.$$

The difference $\varepsilon_n = \overline{\sigma}_{t_n} - \overline{\sigma}_n$ satisfies the following recurrence:

$$\begin{aligned}\varepsilon_n &= [I + \Delta\Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}}]\varepsilon_{n-1} \\ &\quad + \int_{t_{n-1}}^{t_n} [\Phi^{-1}_s A^* \Phi_s \overline{\sigma}_s - \Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}} \overline{\sigma}_{t_{n-1}}] \, ds \\ &\quad + \int_{t_{n-1}}^{t_n} \langle \overline{p}_s - \overline{p}_{t_{n-1}}, e_i\rangle e_i \, ds \\ &\quad + \Delta\langle \overline{p}_{t_{n-1}} - \overline{p}_{n-1}, e_i\rangle e_i \\ &\quad + \Delta^2 \langle \overline{p}_{n-1}, e_i\rangle \Phi^{-1}_{t_{n-1}} A^* \Phi_{t_{n-1}} e_i.\end{aligned}$$

Therefore

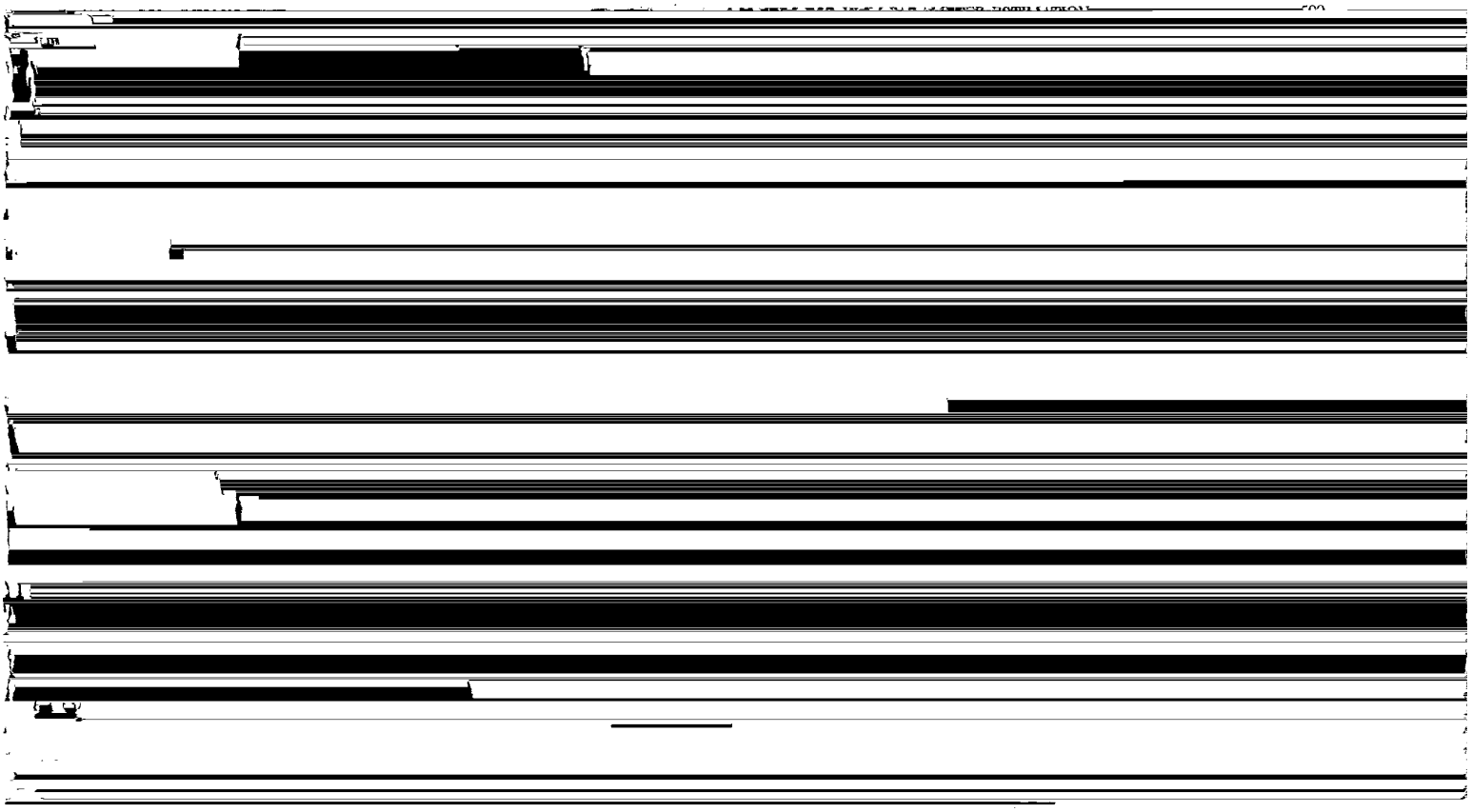
$$|\varepsilon_n[y]| \le |\varepsilon_{n-1}[y]| e^{C\Delta} + K\Delta[\Delta + \omega_\Delta(y)]$$

where the constants C and K depend on $\|y\|$, and hence (3.10) follows by the discrete Gronwall lemma.

The proof for $H_t = N^{ij}_t$ and $H_n = N^{ij}_n$ is quite similar, and is therefore omitted.

For $H_t = G^i_t$ and $H_n = \Delta \cdot G^i_n$, one has to take into account the stochastic integral in (2.14). Actually, the idea is to use an integration by parts as in the proof of Theorem 2.2. We introduce the notations $\overline{\sigma}_t = \overline{\sigma}(G^i_t X_t)$, and $\overline{\sigma}_n = \Delta \cdot \overline{\sigma}(G^i_n X_n)$. From (2.14)

$$\begin{aligned}\overline{\sigma}_{t_n} &= \overline{\sigma}_{t_{n-1}} + \int_{t_{n-1}}^{t_n} \Phi^{-1}_s A^* \Phi_s \overline{\sigma}_s \, ds + \int_{t_{n-1}}^{t_n} \langle \overline{p}_s, e_i\rangle e_i \, dy_s \\ &= \overline{\sigma}_{t_{n-1}} + \int_{t_{n-1}}^{t_n} \Phi^{-1}_s A^* \Phi_s \overline{\sigma}_s \, ds \\ &\quad + [y_{t_n} - y_{t_{n-1}}]\langle \overline{p}_{t_n}, e_i\rangle e_i \\ &\quad - \int_{t_{n-1}}^{t_n} [y_s - y_{t_{n-1}}]\langle \Phi^{-1}_s A^* \Phi_s \overline{p}_s, e_i\rangle e_i \, ds\end{aligned}$$



whereas from (3.9)

$$\begin{aligned}\overline{\sigma}_n =& [I + \Delta\Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}}]\overline{\sigma}_{n-1} + [y_{t_n} - y_{t_{n-1}}]\langle \overline{p}_{n-1}, e_i\rangle \\ & \cdot [I + \Delta\Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}}]e_i \\ =& [I + \Delta\Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}}]\overline{\sigma}_{n-1} + [y_{t_n} - y_{t_{n-1}}]\langle \overline{p}_n, e_i\rangle e_i \\ & - \Delta[y_{t_n} - y_{t_{n-1}}]\langle \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}} \overline{p}_{n-1}, e_i\rangle e_i \\ & + \Delta[y_{t_n} - y_{t_{n-1}}]\langle \overline{p}_{n-1}, e_i\rangle \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}} e_i.\end{aligned}$$

The difference $\varepsilon_n = \overline{\sigma}_{t_n} - \overline{\sigma}_n$ satisfies the following recurrence:

$$\begin{aligned}\varepsilon_n =& [I + \Delta\Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}}]\varepsilon_{n-1} \\ & + \int_{t_{n-1}}^{t_n} [\Phi_s^{-1} A^* \Phi_s \overline{\sigma}_s - \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}} \overline{\sigma}_{t_{n-1}}]\, ds \\ & + [y_{t_n} - y_{t_{n-1}}]\langle \overline{p}_{t_n} - \overline{p}_n, e_i\rangle e_i \\ & - \int_{t_{n-1}}^{t_n} [y_s - y_{t_{n-1}}]\langle \Phi_s^{-1} A^* \Phi_s \overline{p}_s, e_i\rangle e_i\, ds \\ & + \Delta[y_{t_n} - y_{t_{n-1}}]\langle \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}} \overline{p}_{n-1}, e_i\rangle e_i \\ & - \Delta[y_{t_n} - y_{t_{n-1}}]\langle \overline{p}_{n-1}, e_i\rangle \Phi_{t_{n-1}}^{-1} A^* \Phi_{t_{n-1}} e_i.\end{aligned}$$

Therefore

$$|\varepsilon_n[y]| \le |\varepsilon_{n-1}[y]| e^{C\Delta} + K\Delta[\Delta + \omega_\Delta(y)]$$

where the constants C and K depend on $\|y\|$, and hence (3.10) follows by the discrete Gronwall lemma. □

Corollary 3.4: With the definitions of Theorem 3.3, for all n, Δ such that $0 \le t_n \le T$

$$|\sigma(H_{t_n})[y] - \sigma(H_n)[y]| \le K'[\Delta + \omega_\Delta(y)] \tag{3.11}$$

where the constant K' depends on $\|y\|$.

Proof: Indeed

Consider the EM update formulas (2.3). From the definition of J_T^i we get easily

$$\boldsymbol{E}[J_T^i | \mathcal{Y}_T] = \int_0^T \boldsymbol{E}[\langle X_t, e_i\rangle | \mathcal{Y}_T]\, dt \tag{4.1}$$

where the integrand is the smoothed estimate of $\{\langle X_t, e_i\rangle, 0 \le t \le T\}$, given $\mathcal{Y}_T$, computable from the smoothing probability distribution given below. However, the computation of $\boldsymbol{E}[N_T^{ij}|\mathcal{Y}_T]$ and $\boldsymbol{E}[G_T^i|\mathcal{Y}_T]$ is not as straightforward, as we will see.

State: Following [10], we define the continuous-time smoother for state estimation, as

$$q_t = \overline{E}[X_t \Lambda_T | \mathcal{Y}_T].$$

This is an N-dimensional vector, whose ith component is

$$\langle q_t, e_i\rangle = \overline{E}[1_{(X_t = e_i)} \Lambda_T | \mathcal{Y}_T] = \langle p_t, e_i\rangle \langle v_t, e_i\rangle. \tag{4.2}$$

Here $p_t = \sigma(X_t)$ is the unnormalized filter defined in Section II-C, and v_t is the solution of the backward SDE, dual to (2.5)

$$v_t = \mathbf{1} + \int_t^T A v_s\, ds + \int_t^T B v_s\, dy_s. \tag{4.3}$$

The stochastic integral in (4.3) is a backward Ito integral. Here, duality means that $\langle q_t, \mathbf{1}\rangle = \langle p_t, v_t\rangle$ is independent of t.

Discussion 4.1: To give a hint of the proof of (4.2), we start from the following representation for the solution of (4.3):

$$\langle v_t, e_i\rangle = \overline{E}_{t,e_i}[\Lambda_T^t | \mathcal{Y}_T^t]$$

Discussion 4.2: Another way of computing $\sigma(H_T)$ for $H_T = N_T^{ij}, G_T^i$ or J_T^i, is to obtain an equation for $\langle \sigma(H_t X_t), v_t \rangle$. One would then obtain $\sigma(H_T)$ as $\sigma(H_T) = \langle \sigma(H_T X_T), v_T \rangle$, since $v_T = \mathbf{1}$.

The equations for $\langle \sigma(H_t X_t), v_t \rangle$ are derived below as in Campillo–LeGland [1], using duality arguments and the chain rule of the two-sided stochastic calculus introduced and studied in Pardoux–Protter [11]. Roughly speaking, the two-sided stochastic integral can be defined as follows: Let $\{u_t, 0 \le t \le T\}$ (resp., $\{v_t, 0 \le t \le T\}$) be the solution of a forward (resp., a backward) stochastic differential equation driven by the Brownian motion $\{y_t, 0 \le t \le T\}$ on $(\Omega, \mathcal{F}, \overline{P})$. Then, by definition

$$\int_0^T f(u_t, v_t)\, dy_t = \lim \sum_{n=1}^{M} f(u_{t_{n-1}}, v_{t_n})(y_{t_n} - y_{t_{n-1}})$$

where the limit is taken as the mesh of the partition $0 = t_0 < \cdots < t_n < \cdots < t_M = T$ goes to zero. If $f(u, v)$ does not depend on v, the two-sided integral coincides with the forward Ito integral. Similarly, if $f(u, v)$ does not depend on u, the two-sided integral coincides with the backward Ito integral. Finally (a simple form of), the chain rule for the two-sided stochastic integral reads

$$u_t v_t = u_s v_s + \int_s^t v_r\, du_r + \int_s^t u_r\, dv_r. \tag{4.5}$$

Details can be found in Pardoux–Protter [11].

Occupation Time: From (2.6) and (4.3), and using the chain rule (4.5)

$$\begin{aligned} d\langle \sigma(J_t^i X_t), v_t \rangle &= \langle A^* \sigma(J_t^i X_t), v_t \rangle\, dt + \langle B\sigma(J_t^i X_t), v_t \rangle\, dy_t \\ &\quad + \langle p_t, e_i \rangle \langle v_t, e_i \rangle\, dt - \langle \sigma(J_t^i X_t), A v_t \rangle\, dt \\ &\quad - \langle \sigma(J_t^i X_t), B v_t \rangle\, dy_t \\ &= \langle p_t, e_i \rangle \langle v_t, e_i \rangle\, dt = \langle q_t, e_i \rangle\, dt. \end{aligned}$$

Therefore

$$\sigma(J_T^i) = \langle \sigma(J_T^i X_T), v_T \rangle = \int_0^T \langle q_t, e_i \rangle\, dt \tag{4.6}$$

which is the expected result (4.5).

Number of Jumps: From (2.7) and (4.3), and using the chain rule (4.5)

$$\begin{aligned} d\langle \sigma(N_t^{ij} X_t), v_t \rangle &= \langle A^* \sigma(N_t^{ij} X_t), v_t \rangle dt + \langle B\sigma(N_t^{ij} X_t), v_t \rangle dy_t \\ &\quad + a_{ij} \langle p_t, e_i \rangle \langle v_t, e_j \rangle dt - \langle \sigma(N_t^{ij} X_t), A v_t \rangle dt \\ &\quad - \langle \sigma(N_t^{ij} X_t), B v_t \rangle dy_t \\ &= a_{ij} \langle p_t, e_i \rangle \langle v_t, e_j \rangle dt. \end{aligned}$$

Therefore

$$\sigma(N_T^{ij}) = \langle \sigma(N_T^{ij} X_T), v_T \rangle = a_{ij} \int_0^T \langle p_t, e_i \rangle \langle v_t, e_j \rangle\, dt. \tag{4.7}$$

Level Integrals: From (2.8) and (4.3), and using the chain rule (4.5)

$$\begin{aligned} d\langle \sigma(G_t^i X_t), v_t \rangle &= \langle A^* \sigma(G_t^i X_t), v_t \rangle dt + \langle B\sigma(G_t^i X_t), v_t \rangle dy_t \\ &\quad + \langle p_t, e_i \rangle \langle v_t, e_i \rangle dy_t + g_i \langle p_t, e_i \rangle \langle v_t, e_i \rangle dt \\ &\quad - \langle \sigma(G_t^i X_t), A v_t \rangle dt - \langle \sigma(G_t^i X_t), B v_t \rangle dy_t \\ &= \langle p_t, e_i \rangle \langle v_t, e_i \rangle dy_t + g_i \langle p_t, e_i \rangle \langle v_t, e_i \rangle dt \\ &= \langle q_t, e_i \rangle dy_t + g_i \langle q_t, e_i \rangle dt. \end{aligned}$$

Therefore

$$\begin{aligned} \sigma(G_T^i) &= \langle \sigma(G_T^i X_T), v_T \rangle \\ &= \int_0^T \langle q_t, e_i \rangle\, dy_t + g_i \int_0^T \langle q_t, e_i \rangle\, dt \end{aligned}$$

where the stochastic integral is a two-sided stochastic integral.

Remark 4.3: The following equivalent expressions are obtained as in Dembo and Zeitouni [4]:

$$\sigma(G_T^i) = \int_0^T \langle q_t, e_i \rangle \circ dy_t \tag{4.8}$$

where the stochastic integral is a generalized Stratonovich integral, and

$$\sigma(G_T^i) = \langle q_T, e_i \rangle y_T - \int_0^T y_t \langle \dot{q}_t, e_i \rangle\, dt$$

since $\{q_t, 0 \le t \le T\}$ is a finite variation process. Roughly speaking, the generalized Stratonovich integral of the (not necessarily adapted) process $\{u_t, 0 \le t \le T\}$ w.r.t. the Brownian motion $\{y_t, 0 \le t \le T\}$ is given by

$$\int_0^T u_t \circ dy_t = \lim \sum_{n=1}^{M} \left\{ \frac{1}{t_n - t_{n-1}} \int_{t_{n-1}}^{t_n} u_t\, dt \right\} (y_{t_n} - y_{t_{n-1}})$$

where the limit is taken as the mesh of the partition $0 = t_0 < \cdots < t_n < \cdots < t_M = T$ goes to zero. Details can be found in Dembo and Zeitouni [4].

Remark 4.4: The expressions given by (4.6)–(4.8) are the continuous-time counterparts of the expressions arising in the Baum–Welsh re-estimation formulas for the discrete-time HMM introduced in Section III-A. Indeed, the re-estimation formulas (2.3) read now

$$a'_{ij} = a_{ij} \frac{\displaystyle\int_0^T \langle p_t, e_i \rangle \langle v_t, e_j \rangle\, dt}{\displaystyle\int_0^T \langle p_t, e_i \rangle \langle v_t, e_i \rangle\, dt}$$

and

$$g'_i = \frac{\displaystyle\int_0^T \langle p_t, e_i \rangle \langle v_t, e_i \rangle \circ dy_t}{\displaystyle\int_0^T \langle p_t, e_i \rangle \langle v_t, e_i \rangle\, dt}.$$

These expressions are apparently obtained here for the first time. It should be noticed that the smoothers presented in [7] address the problem of computing $\boldsymbol{E}[H_s | \mathcal{Y}_t]$ for $0 \le s \le t$, whereas the purpose here is rather to compute $\boldsymbol{E}[H_T | \mathcal{Y}_T]$ by means of a smoother estimate for the state.

V. Time Discretization of Smoothing Equations

We consider the following approximation for the backward Duncan–Mortensen–Zakai equation (4.3)

$$v_{n-1} = [I + \Delta A]\Psi_n v_n \tag{5.1}$$

which is dual to (3.2). The resulting approximation q_n to the unnormalized smoothing probability distribution q_{t_n} is given simply by

$$\langle q_n, e_i\rangle = \langle p_n, e_i\rangle\langle v_n, e_i\rangle. \tag{5.2}$$

Because of the probabilistic interpretation associated with this approximation, the conditional expectations involved in the EM algorithm (3.3) are immediately given by the Baum–Welsh re-estimation equations, which involve the forward variable $\{p_n, 0 \le n \le M\}$ and the backward variable $\{v_n, 0 \le n \le M\}$. Our purpose here is to recover these equations as an application of duality. Indeed, another way of compute $\sigma(H_M)$ for $H_M = N_M^{ij}, G_M^i$ or J_M^i, is to obtain an equation for $\langle\sigma(H_n X_n), v_n\rangle$. One would then obtain $\sigma(H_M)$ as $\sigma(H_M) = \langle\sigma(H_M X_M), v_M\rangle$, since $v_M = \mathbf{1}$.

Recall that $M = [T/\Delta]$ denotes the largest integer such that $M\Delta \le T$.

State: See (5.2), to be compared with (4.2).

Occupation Time: From (3.4) and (5.1)

$$\begin{aligned}\langle\sigma(J_n^i X_n), v_n\rangle &= \langle\Psi_n P^*\sigma(J_{n-1}^i X_{n-1}), v_n\rangle \\ &\quad + \langle p_{n-1}, e_i\rangle\langle\Psi_n P^* e_i, v_n\rangle \\ &= \langle\sigma(J_{n-1}^i X_{n-1}), v_{n-1}\rangle \\ &\quad + \langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle.\end{aligned}$$

Therefore

$$\begin{aligned}\sigma(J_M^i) = \langle\sigma(J_M^i X_M), v_M\rangle &= \sum_{n=1}^{M}\langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle \\ &= \sum_{n=1}^{M}\langle q_{n-1}, e_i\rangle\end{aligned} \tag{5.3}$$

to be compared with (4.6).

Number of Transitions: From (3.6) and (5.1)

$$\begin{aligned}\langle\sigma(N_n^{ij} X_n), v_n\rangle &= \langle\Psi_n P^*\sigma(N_{n-1}^{ij} X_{n-1}), v_n\rangle \\ &\quad + \langle p_{n-1}, e_i\rangle\langle\Psi_n P^* e_i, e_j\rangle\langle v_n, e_j\rangle \\ &= \langle\sigma(N_{n-1}^{ij} X_{n-1}), v_{n-1}\rangle \\ &\quad + \langle p_{n-1}, e_i\rangle\langle\Psi_n P^* e_i, e_j\rangle\langle v_n, e_j\rangle.\end{aligned}$$

Note that

$$\langle\Psi_n P^* e_i, e_j\rangle = a_{ij}\Delta\phi_{t_n}^j/\phi_{t_{n-1}}^j, \quad \text{for } i \ne j.$$

Therefore, for $i \ne j$

$$\begin{aligned}\sigma(N_M^{ij}) &= \langle\sigma(N_M^{ij} X_M), v_M\rangle \\ &= a_{ij}\Delta\sum_{n=1}^{M}\langle p_{n-1}, e_i\rangle\langle v_n, e_j\rangle\phi_{t_n}^j/\phi_{t_{n-1}}^j \\ &= a_{ij}\Delta\sum_{n=1}^{M}\langle p_{n-1}, e_i\rangle\langle v_n, e_j\rangle\psi_n^j\end{aligned} \tag{5.4}$$

to be compared with (4.7).

Level Integrals: From (3.8) and (5.1)

$$\begin{aligned}\langle\sigma(G_n^i X_n), v_n\rangle &= \langle\Psi_n P^*\sigma(G_{n-1}^i X_{n-1}), v_n\rangle \\ &\quad + z_n^\Delta\langle p_{n-1}, e_i\rangle\langle\Psi_n P^* e_i, v_n\rangle \\ &= \langle\sigma(G_{n-1}^i X_{n-1}), v_{n-1}\rangle \\ &\quad + z_n^\Delta\langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle.\end{aligned}$$

Therefore

$$\begin{aligned}\sigma(G_M^i) &= \langle\sigma(G_M^i X_M), v_M\rangle \\ &= \sum_{n=1}^{M} z_n^\Delta\langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle \\ &= \sum_{n=1}^{M} z_n^\Delta\langle q_{n-1}, e_i\rangle\end{aligned} \tag{5.5}$$

to be compared with (4.8).

Remark 5.1: The re-estimation formulas (3.3) read now

$$\pi_{ij}' = \pi_{ij}\frac{\displaystyle\sum_{n=1}^{M}\langle p_{n-1}, e_i\rangle\langle v_n, e_j\rangle\psi_n^j}{\displaystyle\sum_{n=1}^{M}\langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle}$$

and

$$g_i' = \frac{\displaystyle\sum_{n=1}^{M} z_n^\Delta\langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle}{\displaystyle\sum_{n=1}^{M}\langle p_{n-1}, e_i\rangle\langle v_{n-1}, e_i\rangle}.$$

Not surprisingly, these expressions coincide exactly with the expressions arising in the Baum–Welsh re-estimation formulas for the discrete-time HMM introduced in Section III-A.

The cost for computing the smoothed estimates of the state, occupation time, number of transitions, and level integrals at each time instant are $O(N^2)$. Because either the forward variables $p_n, n = 1, \cdots, M$ or the backward variables $v_n, n = 1, \cdots, M$ must be stored to compute the fixed-interval smoothed estimates, the memory required is $O(NM)$.

The following error estimate can be proved.

Proposition 5.2: For all n, Δ such that $0 \le t_n \le T$

$$|v_{t_n} - v_n| \le K''[\Delta + \omega_\Delta(y)]$$

where the constant K'' depends on $\|y\|$.

However, what is really important for the estimation problem is to estimate the difference

$$\sigma(H_T) - \sigma(H_M) = \langle\sigma(H_{t_M} X_{t_M}), v_{t_M}\rangle - \langle\sigma(H_M X_M), v_M\rangle$$

assuming $t_M = T$. This was already the object of the Corollary 3.4.

VI. Normalization and Variance Estimation

In this section we first consider normalization of the various filters and smoothers. Then we discuss estimation of the noise variance.

A. Normalization

To avoid numerical overflow, it is important to normalize the numerical approximation schemes, see [1] and [8]. The normalized schemes are:

Filtering

State: Denoting the normalized state estimate as π_n, we have

$$\pi_n = \Psi_n P^* \pi_{n-1}/c_n \tag{6.1}$$

with initial condition π_0, where the normalization constant c_n is defined by

$$c_n = \langle \Psi_n P^* \pi_{n-1}, \mathbf{1} \rangle.$$

For $H_n = J_n^i, N_n^{ij}$ or G_n^i, we define

$$\pi(H_n X_n) = \sigma(H_n X_n)/\gamma_n$$

where

$$\gamma_n = \langle p_n, \mathbf{1} \rangle = c_n \cdot c_{n-1} \cdots c_1.$$

The remaining normalized filters are

Occupation Time:

$$\pi(J_n^i X_n) = [\Psi_n P^* \pi(J_{n-1}^i X_{n-1}) + \langle \pi_{n-1}, e_i \rangle \Psi_n P^* e_i]/c_n$$

with initial condition $\pi(J_0^i X_0) = 0$.

Number of Transitions:

$$\pi(N_n^{ij} X_n) = [\Psi_n P^* \pi(N_{n-1}^{ij} X_{n-1}) + \langle \pi_{n-1}, e_i \rangle \langle \Psi_n P^* e_i, e_j \rangle e_j]/c_n$$

with initial condition $\pi(N_0^{ij} X_0) = 0$.

Level Integral:

$$\pi(G_n^i X_n) = [\Psi_n P^* \pi(G_{n-1}^i X_{n-1}) + z_n^{\Delta} \langle \pi_{n-1}, e_i \rangle \Psi_n P^* e_i]/c_n$$

with initial condition $\pi(G_0^i X_0) = 0$.

The re-estimation formulas (3.3) read now

$$\pi'_{ij} = \frac{\langle \pi(N_M^{ij} X_M), \mathbf{1} \rangle}{\langle \pi(J_M^i X_M), \mathbf{1} \rangle}$$

and

$$g'_i = \frac{\langle \pi(G_M^i X_M), \mathbf{1} \rangle}{\langle \pi(J_M^i X_M), \mathbf{1} \rangle}.$$

Smoothing

State: With c_n defined as above, define

$$u_{n-1} = P \Psi_n u_n / c_n \tag{6.2}$$

with initial condition (at final time) $u_M = \mathbf{1}$. Note that (6.1) and (6.2) are dual, and $u_n = v_n \gamma_n / \gamma_M$. It follows that

$$\langle \pi_n, u_n \rangle = \langle p_n, v_n \rangle / \gamma_M = \langle p_M, v_M \rangle / \gamma_M = 1.$$

The normalized approximations for the conditional expectations are easily obtained from (5.3)–(5.5), and the re-estimation formulas (3.3) read now

$$\pi'_{ij} = \pi_{ij} \frac{\sum_{n=1}^{M} \langle \pi_{n-1}, e_i \rangle \langle u_n, e_j \rangle \psi_n^j / c_n}{\sum_{n=1}^{M} \langle \pi_{n-1}, e_i \rangle \langle u_{n-1}, e_i \rangle}$$

and

$$g'_i = \frac{\sum_{n=1}^{M} z_n^{\Delta} \langle \pi_{n-1}, e_i \rangle \langle u_{n-1}, e_i \rangle}{\sum_{n=1}^{M} \langle \pi_{n-1}, e_i \rangle \langle u_{n-1}, e_i \rangle}.$$

B. Estimation of Noise Variance

So far we have assumed that the variance of the observation noise $\{w_t, t \geq 0\}$ is known, and for simplicity we assumed it to be one. When the variance of $\{w_t, t \geq 0\}$ is a known value β^2 (say), the appropriate filtering equations can be obtained by a simple scaling.

In continuous time, it is not possible to obtain an MLE of the variance of $\{w_t, t \geq 0\}$ because measures corresponding to Wiener processes with different variances are not absolutely continuous, see Liptser and Shiryayev [9]. However, in discrete time, we can appeal to an underlying Lebesgue measure and use densities with respect to this Lebesgue measure to compute Radon–Nikodym derivatives of observation processes with different noise variances. So in discrete time, the MLE estimate of the observation noise variance can be approximated using the EM approach as follows.

If the variance of $\{w_t, t \geq 0\}$ is β^2, then $\{w_n^{\Delta}, n \geq 0\}$ defined in Section III-A is a white Gaussian sequence with variance β^2/Δ. Now consider the EM update from the estimate g, β to g', β', obtained by maximizing the function $Q(\cdot, (g, \beta))$. Write P and P' for the respective probability measures. Consider the Radon–Nikodym derivative $dP'/dP|_{\mathcal{F}_M} = \Lambda_M$, where $M = [T/\Delta]$ is the largest integer such that $M\Delta \leq T$ and

$$\Lambda_M = \prod_{n=1}^{M} \frac{\frac{1}{\sqrt{2\pi\beta'^2/\Delta}} \exp\left\{ -\frac{|z_n^{\Delta} - \langle g', X_{n-1} \rangle|^2}{2\beta'^2/\Delta} \right\}}{\frac{1}{\sqrt{2\pi\beta^2/\Delta}} \exp\left\{ -\frac{|z_n^{\Delta} - \langle g, X_{n-1} \rangle|^2}{2\beta^2/\Delta} \right\}}.$$

Then we can write

$$\begin{aligned} Q(\cdot, (g, \beta)) &= \boldsymbol{E}[\log \Lambda_M | \mathcal{Z}_M] \\ &= -\frac{M}{2} \log(\beta'^2/\Delta) + \frac{\Delta}{2\beta'^2} \\ &\quad \cdot \left\{ \sum_{i=1}^{N} [\hat{J}_M^i g_i'^2 - 2\hat{G}_M^i g'_i] + \sum_{n=1}^{M} |z_n^{\Delta}|^2 \right\} \\ &\quad + f(g, \beta) \end{aligned}$$

where $f(g, \beta)$ is independent of (g', β'), $\hat{G}_M^i = \boldsymbol{E}[G_M^i | \mathcal{Z}_M]$, and $\hat{J}_M^i = \boldsymbol{E}[J_M^i | \mathcal{Z}_M]$. Maximization of $Q(\cdot, (g, \beta))$ yields

$$g'_i = \frac{\boldsymbol{E}[G_M^i | \mathcal{Z}_M]}{\boldsymbol{E}[J_M^i | \mathcal{Z}_M]}$$

	J_1	J_2	J_3	G_1	G_2	G_3	% Error
Actual	13.5474	20.2009	6.2599	−13.7292	0.1509	6.3712	—
Filters	13.2370	19.5833	6.1616	−13.2830	0.1228	6.1639	15.0350
Smoothers	13.6945	19.8642	6.4393	−13.7505	0.1189	6.4357	7.2700
	N_{12}	N_{13}	N_{21}	N_{23}	N_{31}	N_{32}	
Actual	140	87	103	46	123	10	
Filters	134.2215	94.5224	96.0915	44.3199	132.7201	6.1707	
Smoothers	134.2215	94.5224	96.0915	44.3199	132.7201	6.1707	

	J_1	J_2	J_3	G_1	G_2	G_3	% Error
Actual	13.5474	20.2009	6.2599	−13.9103	0.3019	6.4824	—
Filters	13.1473	19.7305	6.1081	−13.2535	0.3153	6.0072	30.7100
Smoothers	13.6059	20.0113	6.3809	−13.7096	0.3185	6.2814	22.3150
	N_{12}	N_{13}	N_{21}	N_{23}	N_{31}	N_{32}	
Actual	140	87	103	46	123	10	
Filters	134.4378	94.8396	98.9042	41.4689	130.3399	6.0512	
Smoothers	134.4378	94.8396	98.9042	41.4689	130.3399	6.0512	

which is the same expression as in (3.3), and

$$\beta'^2 = \frac{\Delta}{M}\left\{\sum_{n=1}^{M}|z_n^{\Delta}|^2 - \sum_{i=1}^{N} g_i' \boldsymbol{E}[G_M^i|\mathcal{Z}_M]\right\} \tag{6.3}$$

which is the EM update for the noise variance.

Remark 6.1: The right-hand side of (6.3) is related to the quadratic variation $\langle y\rangle_T = \beta^2 T$ (which is observed in principle and can be used to estimate the variance β^2) of the continuous-time observation process. Indeed

$$\begin{aligned}\frac{\Delta}{M}&\left\{\sum_{n=1}^{M}|z_n^{\Delta}|^2 - \sum_{i=1}^{N} g_i' \boldsymbol{E}[G_M^i|\mathcal{Z}_M]\right\}\\ &= \frac{1}{T}\sum_{n=1}^{M}|y_{t_n} - y_{t_{n-1}}|^2 + O(\Delta^2)\\ &\approx \frac{1}{T}\langle y\rangle_T, \quad \text{as } \Delta \downarrow 0.\end{aligned}$$

VII. Numerical Examples

In this section we present computer simulations to illustrate and compare the performance of the discretized filter-based and smoother-based EM algorithms. The normalized equations presented in Section VI-A were used.

A 3-state HMM was generated with parameters A, g given by

$$A = \begin{pmatrix} -17 & 10 & 7 \\ 5 & -7 & 2 \\ 20 & 1 & -21 \end{pmatrix} \qquad g = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}.$$

The following tables (at the top of this page) show the actual values of the occupation times, level integrals, and numbers of jumps, and the estimated values computed using the filtering and smoothing equations. The percentage errors for the state estimates are also shown.

1. Simulation of the filters and smoothers, with $\Delta = 0.002, T = 40, M = 20\,000$. Observation noise variance $\beta = 0.05$, assumed known. True values of A and g are used (see the table at the top of this page).

2. Simulation of the filters and smoothers, with $\Delta = 0.002, T = 40, M = 20\,000$. Observation noise variance $\beta = 0.1$, assumed known. True values of A and g are used (see the second table at the top of this page).

3. Simulation of the EM algorithm, using the above true parameters A, g, and with $\Delta = 0.002, T = 40, M = 20\,000$. Observation noise variance $\beta = 0.05$, assumed known. Estimated values of A and g are computed using the EM algorithm.

Fig. 1 contains graphs which show the evolution of the parameter estimates in terms of passes through the EM algorithm, as well as a graph showing the improvement of the state estimate percentage error as the EM algorithm progresses. The conditional expectations needed in the EM algorithm were computed using filters. The analogous results using smoothing are shown in Fig. 2.

4. Simulation of the EM algorithm, using the above true parameters A, g, and with $\Delta = 0.002, T = 40, M = 20\,000$. The true value of the observation noise variance was $\beta = 0.05$, assumed unknown and estimated using the method of Section VI-B. Estimated values of A and g are computed using the EM algorithm.

Fig. 3 contains graphs which show the evolution of the parameter estimates and state estimate percentage error, as well as that of the noise variance estimate. The final estimate for β is 0.0595076. The conditional expectations needed in the EM

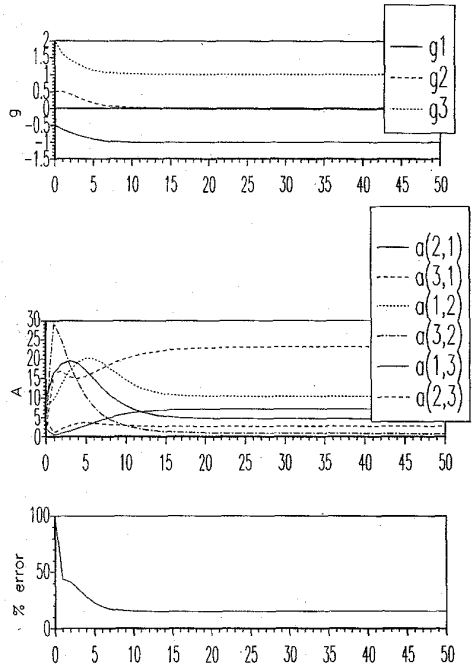


Fig. 1. Parameter estimates and state estimate error for a 3-state HMM versus EM iterations using filters.

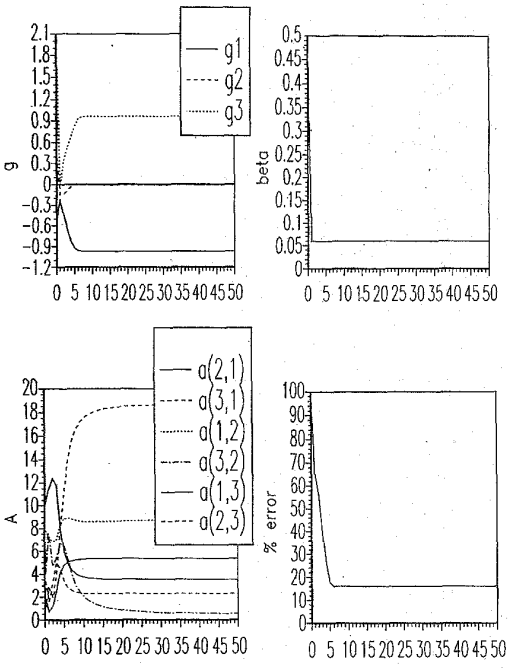
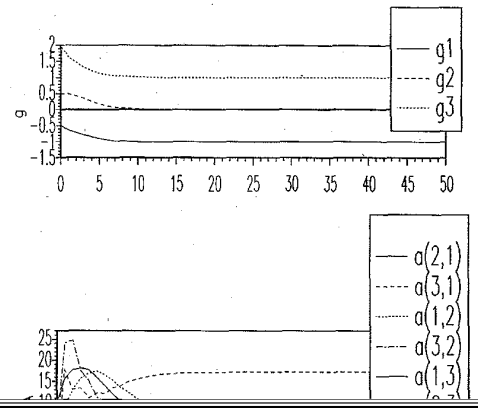
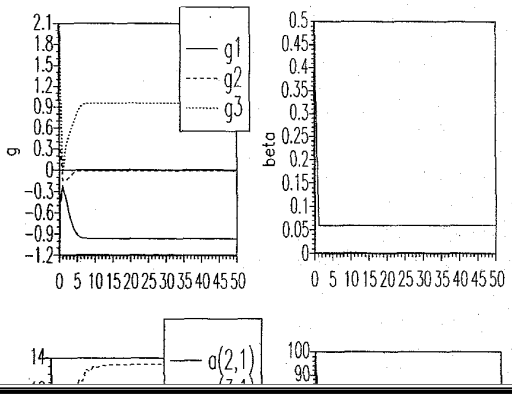


Fig. 3. Parameter and noise standard deviation estimates, and state estimate error for a 3-state HMM versus EM iterations using filters.





may incur additional round-off error, depending on how it is coded. Use of an implicit scheme may help. Also, we found that it was necessary to use double precision arithmetic to implement the algorithms, because of the exponentations needed in computing Ψ_n in the algorithms presented in Section VI-A.

REFERENCES

[1] F. Campillo and F. Le Gland, "MLE for the partially observed diffusions: Direct maximization vs. the EM algorithm," *Stochastic Processes and their Applications*, vol. 33, no. 2, pp. 245–274, 1989.

[2] J. M. C. Clark, "The design of robust approximations to the stochastic differential equations of nonlinear filtering," in J. K. Skwirzynski, Ed., *Communication Systems and Random Processes Theore, Darlington 1977.* Alphen aan den Rijn, The Netherlands: Sijthoff and Noordhoff, 1978, pp. 721–734.

[3] A. Dembo and O. Zeitouni, "Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm," *Stochastic Processes and their Applications*, vol. 23, no. 1, pp. 91–113, 1986.

[4] ——, "Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm (Corrigendum)," *Stochastic Processes and their Applications*, vol. 31, no. 1, pp. 167–169, 1989.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[6] R. J. Elliott. "Exact adaptive filters for Markov chains observed in Gaussian noise," *Automatica*, vol. 30, no. 9, pp. 1399–1408, Sept. 1994.

[7] ——, "New finite-dimensional filters and smoothers for noisily observed Markov chains," *IEEE Trans. Inform. Theory*, vol. 39, no. 1, pp. 265–271, Jan. 1993.

[8] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process in automatic speech recognition. *Bell System Tech. J.*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[9] R. Sh. Liptser and A. N. Shiryayev, *Statistics of Random Processes I. General Theory*. vol. 5 of *Applications of Mathematics*. New York: Springer-Verlag, 1977.

[10] E. Pardoux, "Equations du lissage non-linéaire," in H. Korezlioglu, G. Mazziotto, and J. Szpirglas, Eds., *Filtering and Control of Random Processes, Paris 1983*. Berlin, Germany: Springer Verlag, 1984, pp. 206–218.

[11] E. Pardoux and P. Protter, "A two-sided stochastic integral and its calculus," *Probability Theory and Related Fields*, vol. 76, no. 1, pp. 15–49, 1987.

[12] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[13] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *SIAM J. Contr. Optimiz.*, vol. 2, no. 3, pp. 347–369, 1965.

[14] O. Zeitouni and A. Dembo, "Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes," *IEEE Trans. Inform. Theory*, vol. 34, no.4, pp. 890–893, July 1988.