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

State Estimation Algorithms for Markov Chains Observed in Arbitrary Noise

W. P. Malcolm

***Abstract*— In this article we compute state estimation schemes for discrete-time Markov chains observed in arbitrary observation noise. Here we assume the observation noise distribution is known in advance. Appealing to a fundamental L^1 convergence result in [1] we propose to represent any practical observation noise model by a convex combination of Gaussian densities, that is, a mixture function that is itself a valid probability density function.**

To compute our state estimation schemes we use the techniques of reference probability, (see [2]). Here however, our Gaussian mixtures appear as sums in a product representation of Radon-Nikodym derivatives. The state estimation schemes we compute are; an information state recursion (filter), a general smoothing theorem, an M-ary detection scheme. A computer simulation is provided to indicate the performance of our recursive filter in a non-Gaussian observation noise scenario.

***Index Terms*— Gaussian-Mixture Distribution, Martingales, Reference Probability, Filtering, Smoothing, Detection, Viterbi Algorithms**

I. Introduction

In this work our primary aim is to relax the Gaussian noise assumption in the observation process model and substitute a suitably chosen finite Gaussian mixture model. Using this approach we compute the corresponding filters smoothers, detection schemes and state-sequence estimation schemes.

The application of a Gaussian mixture approximation in any setting, naturally raises two important questions, these are;

1) precisely how accurate is a Gaussian mixture approximation ?
2) how does one fit a Gaussian mixture to a given non-Gaussian density ?

Question 1 is readily answered by a fundamental Theorem given by Korevaar in [1], which essentially proves, (roughly speaking), that *any* practical probability density can be approximated arbitrarily closely by a finite Gaussian mixture. Korevaar's Theorem effectively provides a compelling case for Gaussian densities to be used in finite mixture approximation. Further, the second question has been answered, (in part), in [3] and [4]. Put briefly, there are several effective algorithms to determine a suitable parametric mixture.

Our motivation in this work is partly driven by a practical reality, that is, in *many* real world scenarios the observation noise processes are well known to be non-Gaussian. Some recent examples of this issue in the context of modern communications are given in the articles: [5], [6], [7] and [8].

W. P Malcolm is with the Mathematical Sciences Institute of the Australian National University, ACT 0200 Australia, Email: **malcolm@maths.anu.edu.au**

This article is organised as follows. In §I-A we describe the classes of dynamics we wish to consider, including some facts about Gaussian mixtures and the extension of reference probability to Gaussian mixture scenarios. In sections II, III and IV, we compute, respectively, Filters, Smoothers and M-ary detection schemes. Finally, in §V, we consider an example of the filter performance in a non-Gaussian noise scenario.

A. State Process Dynamics

Our state processes of interest are time-homogeneous scalar-valued Markov chains evolving in discrete-time. It is now standard to identify the state spaces of such processes with a vector-valued canonical basis of indicator functions. To this end we label our vector-valued Markov chain by $X = \{X_k, 0 \le k\}$. The state space of X is taken as the collection of unit vectors in $\mathbb{R}^n$, that is,

$$\mathcal{S} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n\} = \Big\{(1,0,\dots,0)', (0,1,0,\dots,0)', \dots, (0,0,\dots,1)'\Big\}. \quad (1.1)$$

Suppose X is defined on the probability space $(\Omega, \mathcal{F}, P)$ and write

$$a_{(j,i)} \triangleq P(X_{k+1} = \boldsymbol{e}_j \mid X_k = \boldsymbol{e}_i) = P(X_1 = \boldsymbol{e}_j | X_0 = \boldsymbol{e}_i). \quad (1.2)$$

To denote the matrix of transition probabilities for the process X, we write $A = \big[a_{(j,i)}\big]_{\substack{1\le j\le n\\ 1\le i\le n}}$. The information generated by the state process X is denoted by $\mathcal{F}_{0,k} \triangleq \sigma\{X_0, X_1, \dots, X_k\}$. Then, (see [9]),

$$X_{k+1} = AX_k + L_{k+1}. \quad (1.3)$$

Here, the stochastic process L is an increment of the $(P, \mathcal{F})$-martingale $\varphi_k = \sum_{\ell=1}^{k} L_\ell$.

B. Finite Gaussian Mixtures

Definition 1 *A scalar-valued finite Gaussian mixture is univariate function $\Psi(\xi) : \mathbb{R} \to \mathbb{R}_+$, with the form*

$$\Psi^{(m)}(\xi; \Theta) = \sum_{j=1}^{m} \alpha_j f_j\big(\xi; \{\mu_j, \sigma_j^2\}\big). \quad (1.4)$$

Here, each f_j is a Gaussian density with mean μ_j and variance σ_j^2.
We write $\Theta = \{u_1, \dots, \mu_m, \sigma_1, \dots, \sigma_m\}$. To ensure the function $\Psi(\xi)$ is itself a valid probability density function,

the following two constraints must be satisfied:

$$\alpha_j > 0, \qquad \textit{for each } j \textit{ in the set } \{1,2,\dots,m\}, \quad (1.5)$$

$$\sum_{j=1}^{m} \alpha_j = 1. \quad (1.6)$$

There are several motivations to use a Gaussian mixture approximation. Firstly Gaussian densities enjoy many useful properties. Foremost in state estimation is the fact that any linear transformation of a Gaussian random variable is again a Gaussian random variable. However, in addition to the convenience of working with Gaussian densities, a Gaussian mixture can be chosen to approximate *any* arbitrarily closely any probability density function. This important fact was established in Korevaar [1]. Korevaar's Theorem is stated below.

THEOREM 1 (KOREVAAR) *Suppose p is any practical probability density function and that we wish to approximate p by a Gaussian mixture of order m. Then, for every $\epsilon > 0$, $\exists\, m \in \mathbb{N}$, such that*

$$0 \le \int_{\mathbb{R}} |p(\xi) - \Psi^{(m)}(\xi;\Theta)| d\xi < \epsilon. \quad (1.7)$$

C. Observation Process Dynamics

In the standard formulation the Markov chain X is partially observed through a related sequence $y = \{y_0, y_1, \dots\}$, whose dynamics are given by

$$y_k = \langle X_k, \boldsymbol{g}\rangle + \langle X_k, \boldsymbol{d}\rangle V_k. \quad (1.8)$$

Here $\boldsymbol{g} = (g_1,\dots,g_n)' \in \mathbb{R}^n$, $\boldsymbol{d} = (d_1,\dots,d_n)' \in \mathbb{R}^n$ and the process $V = \{V_k\}_{k\ge 1}$, is a sequence of independent in identically distributed non-Gaussian random variables. In what follows we propose to represent the probability density for the process V by a finite weighted Gaussian mixture distribution of order $m \in \mathbb{N}$.

D. Reference Probability

Notation: Write $\Phi(\xi) : \mathbb{R} \to \mathbb{R}_+$ for the standard Gaussian density

$$\Phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\bigl(-\tfrac{1}{2}\xi^2\bigr). \quad (1.9)$$

We define a new probability measure $P^\dagger$ on the measurable space $(\Omega, \mathcal{F})$, such that, under $P^\dagger$

1) The dynamics for X remain unchanged.
2) The process y, is independently and identically distributed and is Gaussian with zero mean and unity variance.

With the measure $P^\dagger$ now defined, we construct a real world probability measure P, such that under P, the following two conditions hold.

1) The dynamics for X remain unchanged.
2) The random variables

$$V_k \triangleq \frac{y_k - \langle X_k, \boldsymbol{g}\rangle}{\langle X_k, \boldsymbol{d}\rangle}, \quad (1.10)$$

form a sequence of independently and identically distributed non-Gaussian random variables whose density function is the mixture density give at (1.4).

DEFINITION 2 *For $k = 1, 2, \dots,$*

$$\lambda_k \triangleq \sum_{j=1}^{m} \frac{\alpha_j}{\sigma_j \langle X_k, \boldsymbol{d}\rangle} \Phi\left(\frac{y_k - \bigl(\langle X_k, \boldsymbol{g}\rangle + \mu_j\bigr)}{\sigma_j \langle X_k, \boldsymbol{d}\rangle} \right) \Big/ \Phi(y_k) \quad (1.11)$$

$$\Lambda_{0,k} \triangleq \prod_{\ell=0}^{k} \lambda_\ell, \qquad \lambda_0 = 1. \quad (1.12)$$

We introduce the filtrations $\{\mathcal{Y}_{0,k}\} = \cup_{0\le\ell\le k}\mathcal{Y}_\ell$, $\{\mathcal{F}_k\} = \cup_{0\le\ell\le k}\mathcal{F}_\ell$ and $\{\mathcal{G}_k\} = \cup_{0\le\ell\le k}\mathcal{G}_\ell$, where

$$\mathcal{Y}_{0,k} \triangleq \sigma\{y_0, y_1, \dots, y_k\}, \quad (1.13)$$

$$\mathcal{F}_{0,k} \triangleq \sigma\{X_0, X_1, \dots, X_k\} \quad (1.14)$$

$$\mathcal{G}_{0,k} \triangleq \sigma\{X_0, X_1, \dots, X_k, y_0, y_1, \dots, y_k\}. \quad (1.15)$$

The 'real world' probability P, is defined in terms of the probability measure $P^\dagger$, by setting

$$\frac{dP}{dP^\dagger}\Big|_{\mathcal{G}_{0,k}} \triangleq \Lambda_{0,k}. \quad (1.16)$$

E. Bayes Rule

DEFINITION 3 *Recalling the definitions of $\mathcal{G}$ and $\mathcal{Y}$ above, we suppose that γ is any integrable $\mathcal{G}$-adapted process. Then, the abstract form of Bayes rule states*

$$E\bigl[\gamma_k \mid \mathcal{Y}_{0,k}\bigr] = \frac{E^\dagger\bigl[\Lambda_{0,k}\gamma_k \mid \mathcal{Y}_{0,k}\bigr]}{E^\dagger\bigl[\Lambda_{0,k} \mid \mathcal{Y}_{0,k}\bigr]}. \quad (1.17)$$

REMARK 1 In practice the denominator on the right hand side of equation (1.17) is never "directly calculated". This term is in effect a normalising constant and is readily determined as a basic function of the numerator. For example, if X is a Markov chain with dynamics (1.3), then we note that

$$\bigl\langle E^\dagger\bigl[\Lambda_{0,k}X_k \mid \mathcal{Y}_{0,k}\bigr], \mathbf{1}\bigr\rangle = \bigl\langle E^\dagger\bigl[\Lambda_{0,k}\langle X_k, \mathbf{1}\rangle \mid \mathcal{Y}_{0,k}\bigr] = E^\dagger\bigl[\Lambda_{0,k} \mid \mathcal{Y}_{0,k}\bigr]. \quad (1.18)$$

Here $\mathbf{1} = (1,1,\dots,1)' \in \mathbb{R}^n$.

II. FILTERING

In filtering we are interested to estimate conditional probabilities, at time k, for each of the events $X_k = \boldsymbol{e}_i$, given the information record $\mathcal{Y}_{0,k}$.
Write

$$q_k \triangleq E^\dagger\bigl[\Lambda_{0,k}X_k \mid \mathcal{Y}_{0,k}\bigr] \in \mathbb{R}^n_+. \quad (2.19)$$

THEOREM 2 (FILTER) *The recursion for the information state q, has the form:*

$$q_k = \Bigl[\sum_{j=1}^{m} B_k^j\Bigr] A q_{k-1}. \quad (2.20)$$

Here

$$B_k^j = \underset{i=1,2,\ldots,n}{diag}\,\{\gamma_k^{(i,j)}\} \in \mathbb{R}_+^{n\times n} \tag{2.21}$$

and

$$\gamma_k^{(i,j)} = \left\{ \frac{\alpha_j}{\sigma_j\langle \boldsymbol{e}_i, \boldsymbol{d}\rangle}\Phi\left(\frac{y_k - (\langle \boldsymbol{g}, \boldsymbol{e}_i\rangle + \mu_j)}{\sigma_j\langle \boldsymbol{d}, \boldsymbol{e}_i\rangle}\right) \Big/ \Phi(y_k)\right\}. \tag{2.22}$$

REMARK 2 In practice the recursion at equation (2.20) is implemented, however, one is essentially interested in the corresponding normalised probabilities of the form,

$$p\big(X_k = \boldsymbol{e}_i \mid \mathcal{Y}_{0,k}\big) = E\Big(\mathbf{1}_{\{\omega | X_k(\omega) = \boldsymbol{e}_i\}} \mid \mathcal{Y}_{0,k}\Big) = \frac{\langle q_k, \boldsymbol{e}_i\rangle}{\langle q_k, \mathbf{1}\rangle}. \tag{2.23}$$

Proof of Theorem 2.

$$\begin{aligned}
q_k &\triangleq E^\dagger\big[\Lambda_{0,k}X_k \mid \mathcal{Y}_{0,k}\big] \\
&= E^\dagger\big[\Lambda_{0,k-1}\lambda_k \sum_{i=1}^n \langle X_k, \boldsymbol{e}_i\rangle \boldsymbol{e}_i \mid \mathcal{Y}_{0,k}\big] \\
&= \sum_{i=1}^n E^\dagger\big[\Lambda_{0,k-1}\lambda_k\langle X_k, \boldsymbol{e}_i\rangle \mid \mathcal{Y}_{0,k}\big]\boldsymbol{e}_i \\
&= \sum_{i=1}^n E^\dagger\Bigg[\Lambda_{0,k-1}\Bigg\{\sum_{j=1}^m \frac{\alpha_j}{\sigma_j\langle X_k, \boldsymbol{d}\rangle}\times \\
&\qquad \Phi\left(\frac{y_k - (\langle \boldsymbol{g}, X_k\rangle + \mu_j)}{\sigma_j\langle X_k, \boldsymbol{d}\rangle}\right) \Big/ \Phi(y_k)\Bigg\}\langle X_k, \boldsymbol{e}_i\rangle \mid \mathcal{Y}_{0,k}\Bigg]\boldsymbol{e}_i \\
&= \sum_{i=1}^n\sum_{j=1}^m \left\{ \frac{\alpha_j}{\sigma_j\langle \boldsymbol{e}_i, \boldsymbol{d}\rangle}\Phi\left(\frac{y_k - (\langle \boldsymbol{g}, \boldsymbol{e}_i\rangle + \mu_j)}{\sigma_j\langle \boldsymbol{e}_i, \boldsymbol{d}\rangle}\right) \Big/ \Phi(y_k)\right\}\times \\
&\qquad E^\dagger\Big[\Lambda_{0,k-1}\langle X_k, \boldsymbol{e}_i\rangle \mid \mathcal{Y}_{0,k}\Big]\boldsymbol{e}_i \\
&= \sum_{i=1}^n\sum_{j=1}^m \gamma_k^{(i,j)} E^\dagger\Big[\Lambda_{0,k-1}\langle AX_{k-1} + L_k, \boldsymbol{e}_i\rangle \mid \mathcal{Y}_{0,k}\Big]\boldsymbol{e}_i \\
&= \sum_{i=1}^n\sum_{j=1}^m \gamma_k^{(i,j)} E^\dagger\Big[\Lambda_{0,k-1}\langle AX_{k-1}, \boldsymbol{e}_i\rangle \mid \mathcal{Y}_{0,k-1}\Big]\boldsymbol{e}_i \\
&= \sum_{i=1}^n\sum_{j=1}^m \gamma_k^{(i,j)} \langle E^\dagger\Big[\Lambda_{0,k-1}X_{k-1} \mid \mathcal{Y}_{0,k-1}\Big], A^T\boldsymbol{e}_i\rangle \boldsymbol{e}_i \\
&= \sum_{i=1}^n\sum_{j=1}^m \gamma_k^{(i,j)} \langle Aq_{k-1}, \boldsymbol{e}_i\rangle \boldsymbol{e}_i \\
&= \Big[\sum_{j=1}^m B_k^j\Big] Aq_{k-1}.
\end{aligned} \tag{2.24}$$

□

III. SMOOTHING

In our context, the term smoothing refers to computing a conditional-mean estimate of the quantity X_k, (or a MAP estimate), given the information record $\mathcal{Y}_{0,T}$, here the discrete-time index can vary in the interval $[0, T]$, that is, $0 \le k \le T$. Our approach to construct a general smoother Theorem, is to compute dynamics for the backwards, or dual process corresponding to the estimated unnormalised probability q. It is this backwards process that will introduce the information $\mathcal{Y}_{k,T}$ into the formulation of a smoothed estimate.

Appealing to the form of Bayes' Theorem given at 1.17, we note that for any $k \in \{1, 2, \ldots, T\}$

$$E\big[X_k \mid \mathcal{Y}_{0,T}\big] = \frac{E^\dagger\big[\Lambda_{0,T}X_k \mid \mathcal{Y}_{0,T}\big]}{E^\dagger\big[\Lambda_{0,T} \mid \mathcal{Y}_{0,T}\big]}. \tag{3.25}$$

Again we restrict our attention to the numerator of equation (3.25) and note from the property of repeated conditioning that

$$\begin{aligned}
&E^\dagger\big[\Lambda_{0,T}X_k \mid \mathcal{Y}_{0,T}\big] = \\
&\qquad E^\dagger\big[\Lambda_{0,k}X_k\, E^\dagger\big[\Lambda_{k+1,T} \mid \mathcal{Y}_{0,T} \vee \mathcal{F}_{0,k}\big]\mathcal{Y}_{0,T}\big].
\end{aligned} \tag{3.26}$$

It is critical here that we are able to factorise our Radon-Nikodym derivatives, that is

$$\Lambda_{k+1,T} = \prod_{\ell=k+1}^T \lambda_\ell = \lambda_{k+1}\Lambda_{k+2,T}. \tag{3.27}$$

Further, since the state process X is first-order Markov, then

$$E^\dagger\big[\Lambda_{k+1,T} \mid \mathcal{Y}_{0,T} \vee \mathcal{F}_{0,k}\big] = E^\dagger\big[\Lambda_{k+1,T} \mid \mathcal{Y}_{0,T} \vee \sigma\{X_k\}\big]. \tag{3.28}$$

We define a vector-valued process $v \in \mathbb{R}^n$, where

$$v_{k,T} \triangleq \Big(\langle v_{k,T}, \boldsymbol{e}_1\rangle, \ldots, \langle v_{k,T}, \boldsymbol{e}_n\rangle\Big)'. \tag{3.29}$$

Here

$$\langle v_{k,T}, \boldsymbol{e}_i\rangle \triangleq E^\dagger\big[\Lambda_{k+1,T} \mid \mathcal{Y}_{0,T} \vee \{X_k = \boldsymbol{e}_i\}\big]. \tag{3.30}$$

THEOREM 3 *The process v, for $k \in \{0, 1, 2, \ldots, T-1\}$, satisfies the backward recursion*

$$v_{k,T} = A^T\Big[\sum_{j=1}^m B_{k+1}^j\Big] v_{k+1,T}. \tag{3.31}$$

Here A^T denotes the matrix transpose of A and $v_{T,T} = (1, 1, \ldots, 1)'$.

Proof of Theorem 3

$$\begin{aligned}
\big\langle v_{k,T}, \boldsymbol{e}_i\big\rangle &= E^\dagger\Big[\Lambda_{k+2,T}\lambda_{k+1} \mid \mathcal{Y}_{0,T} \vee \{X_k = \boldsymbol{e}_i\}\Big] \\
&= E^\dagger\Bigg[\Lambda_{k+2,T}\Bigg\{\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_j\big\langle X_{k+1}, \boldsymbol{d}\big\rangle}\times \\
&\Phi\left(\frac{y_{k+1} - (\big\langle X_{k+1}, \boldsymbol{g}\big\rangle + \mu_j)}{\sigma_j\big\langle X_{k+1}, \boldsymbol{d}\big\rangle}\right)\Bigg/\Phi(y_{k+1})\Bigg\} \\
&\qquad\Big| \mathcal{Y}_{0,T} \vee \{X_k = \boldsymbol{e}_i\}\Bigg] \\
&= \sum_{\ell=1}^{n} E^\dagger\Bigg[\big\langle X_{k+1}, \boldsymbol{e}_\ell\big\rangle\Lambda_{k+2,T}\Bigg\{\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_j\big\langle X_{k+1}, \boldsymbol{d}\big\rangle}\times \\
&\Phi\left(\frac{y_{k+1} - (\big\langle X_{k+1}, \boldsymbol{g}\big\rangle + \mu_j)}{\sigma_j\big\langle X_{k+1}, \boldsymbol{d}\big\rangle}\right)\Bigg/\Phi(y_{k+1})\Bigg\} \\
&\qquad\Big| \mathcal{Y}_{0,T} \vee \{X_k = \boldsymbol{e}_i\}\Bigg] \\
&= \sum_{\ell=1}^{n}\sum_{j=1}^{m}\gamma_{k+1}^{(\ell,j)}E^\dagger\Big[\big\langle X_{k+1}, \boldsymbol{e}_\ell\big\rangle\times \\
&E^\dagger\big[\Lambda_{k+2,T} \mid \mathcal{Y}_{0,T} \vee \big\{X_k = \boldsymbol{e}_i\big\} \vee \big\{X_{k+1} = \boldsymbol{e}_\ell\big\}\big] \\
&\qquad\Big| \mathcal{Y}_{0,T} \vee \big\{X_k = \boldsymbol{e}_i\big\}\Big] \\
&= \sum_{\ell=1}^{n}\sum_{j=1}^{m}\gamma_{k+1}^{(\ell,j)}E^\dagger\Big[\big\langle X_{k+1}, \boldsymbol{e}_\ell\big\rangle\big\langle v_{k+1,T}, \boldsymbol{e}_\ell\big\rangle \mid \\
&\mathcal{Y}_{0,T} \vee \big\{X_k = \boldsymbol{e}_i\big\}\Big] \\
&= \sum_{\ell=1}^{n} a_{(\ell,i)}\sum_{j=1}^{m}\gamma_{k+1}^{(\ell,j)}\big\langle v_{k+1,T}, \boldsymbol{e}_\ell\big\rangle \\
&= \Big\langle A^T\Big[\sum_{j=1}^{m} B_{k+1}^{j}\Big]v_{k+1,T}, \boldsymbol{e}_i\Big\rangle.
\end{aligned} \tag{3.32}$$

□

THEOREM 4 (SMOOTHER) *The normalised smoothed estimate for the event $\big\{\omega \mid X_k(\omega) = \boldsymbol{e}_i\big\}$, given the information record $\mathcal{Y}_{0,T}$, is computed by the quotient*

$$p(X_k = \boldsymbol{e}_i \mid \mathcal{Y}_{0,T}) = \frac{\big\langle q_k, \boldsymbol{e}_i\big\rangle\big\langle v_{k,T}, \boldsymbol{e}_i\big\rangle}{\sum_{\ell=1}^{n}\big\langle q_k, \boldsymbol{e}_\ell\big\rangle\big\langle v_{k,T}, \boldsymbol{e}_\ell\big\rangle}. \tag{3.33}$$

Proof of Theorem 4.
To establish to right hand side of equation (3.33), we need only consider the numerator, that is

$$\begin{aligned}
&\big\langle E^\dagger\big[\Lambda_{0,T}X_k \mid \mathcal{Y}_{0,T}\big], \boldsymbol{e}_i\big\rangle \\
&\quad= \big\langle E^\dagger\big[\Lambda_{0,k}\Lambda_{k+1,T}X_k \mid \mathcal{Y}_{0,T}\big], \boldsymbol{e}_i\big\rangle \\
&\quad= \Big\langle E^\dagger\Big[\Lambda_{0,k}E^\dagger\big[\Lambda_{k+1,T}X_k \mid \\
&\qquad\mathcal{Y}_{0,T} \vee \big\{X_k = \boldsymbol{e}_i\big\}\big] \mid \mathcal{Y}_{0,T}\Big], \boldsymbol{e}_i\Big\rangle \\
&\quad= \Big\langle E^\dagger\Big[\Lambda_{0,k}X_kE^\dagger\big[\Lambda_{k+1,T} \mid \mathcal{Y}_{0,T} \vee \big\{X_k = \boldsymbol{e}_i\big\}\big] \mid \\
&\qquad\mathcal{Y}_{0,T}\Big], \boldsymbol{e}_i\Big\rangle \\
&\quad= \big\langle E^\dagger\big[\Lambda_{0,k}X_k\big\langle v_{k,T}, \boldsymbol{e}_i\big\rangle \mid \mathcal{Y}_{0,T}\big], \boldsymbol{e}_i\big\rangle \\
&\quad= \big\langle E^\dagger\big[\Lambda_{0,k}X_k \mid \mathcal{Y}_{0,T}\big], \boldsymbol{e}_i\big\rangle\big\langle v_{k,T}, \boldsymbol{e}_i\big\rangle \\
&\quad= \big\langle E^\dagger\big[\Lambda_{0,k}X_k \mid \mathcal{Y}_{0,k}\big], \boldsymbol{e}_i\big\rangle\big\langle v_{k,T}, \boldsymbol{e}_i\big\rangle \\
&\quad= \big\langle q_k, \boldsymbol{e}_i\big\rangle\big\langle v_{k,T}, \boldsymbol{e}_i\big\rangle.
\end{aligned} \tag{3.34}$$

□

IV. M-ARY DETECTION

The term M-ary detection is used in Electrical Engineering to describe sequential hypothesis testing for more than 2 candidate model hypotheses. Here we are interested in model-parameter hypotheses. In effect our formulation is something like a finite version of the EM algorithm, that is, rather than considering an uncountable collection of model parameter sets in the space of all admissible models, we consider a finite collection in this space. For example, write

$$H^j \triangleq \big\{A^{H_j}, \boldsymbol{g}^{H_j}, \boldsymbol{d}^{H_j}\big\}, \qquad j = 1, 2, \ldots, M. \tag{4.35}$$

Here we suppose the Gaussian mixture model for the noise process V is known. Further, it is assumed that any realisation of an observation process y is generated by a fixed model parameter set. What we would like to do is estimate the probabilities $p(H^j \mid \mathcal{Y}_{0,k})$. It will be shown that this problem separates into a pure filtering component and a pure estimation component. In the context of M-ary detection this is known as the Separation Theorem [10].

To compute our M-ary detection schemes we use a vector valued simple random variable that can assume one of M possible states corresponding to each of the M candidate model hypotheses. To this end we suppose $\boldsymbol{\alpha}$ is a vector-valued simple random variable in $\mathbb{R}^M$ and has a state space

$$\begin{aligned}
&\mathcal{A} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_M\} = \\
&\Big\{(1,0,\ldots,0)', (0,1,0,\ldots,0)', \ldots, (0,0,\ldots,1)'\Big\}.
\end{aligned} \tag{4.36}$$

Here, the statistical meaning of the state of $\boldsymbol{\alpha}$, is, $\boldsymbol{\alpha} = \boldsymbol{f}_j \Leftrightarrow H = H_j$. What we are interested to compute, are the estimated un-normalised probabilities

$$q_k^j \triangleq E^\dagger\big[\Lambda_{0,k}\big\langle\boldsymbol{\alpha}, \boldsymbol{f}_j\big\rangle \mid \mathcal{Y}_{0,k}\big]. \tag{4.37}$$

To avoid confusion between the scalar-valued probabilities defined by equation (4.37) and the vector-valued information state defined by equation (2.19), we write

$$q_k^{\text{Det}} \triangleq \big(q_k^1, q_k^2 \ldots, q_k^M\big)'. \tag{4.38}$$

Here "Det" abbreviates M-ary detection probability.

THEOREM 5 (M-ARY DETECTION FILTER) *The M-ary detection filter for the model hypothesis H_i is computed by the recursion*

$$\begin{aligned}\left\langle q_k^{Det}, \boldsymbol{f}_i \right\rangle = \sum_{\ell=1}^{M} \Bigg\{ \sum_{j=1}^{m} \frac{\alpha_j}{\sigma_j \left\langle \boldsymbol{e}_\ell, \boldsymbol{d}^{H_i} \right\rangle} \times \\ \Phi\left(\frac{y_k - \left(\left\langle \boldsymbol{e}_\ell, \boldsymbol{g}^{H_i} \right\rangle + \mu_j \right)}{\sigma_j \left\langle \boldsymbol{e}_\ell, \boldsymbol{d}^{H_i} \right\rangle} \right) \Bigg/ \Phi(y_k) \Bigg\} \\ \times \left\{ \frac{\left\langle A^{H_i} q_{k-1}^{H_i}, \boldsymbol{e}_\ell \right\rangle}{\left\langle A^{H_i} q_{k-1}^{H_i}, \mathbf{1} \right\rangle} \right\} \left\langle q_{k-1}^{Det}, \boldsymbol{f}_i \right\rangle. \end{aligned} \tag{4.39}$$

The corresponding normalised detector probability is computed by normalising over all candidate model-hypotheses, that is

$$p(H = H_i \mid \mathcal{Y}_{0,k}) = \frac{\left\langle q_k^{Det}, \boldsymbol{f}_i \right\rangle}{\left\langle q_k^{Det}, \mathbf{1} \right\rangle}. \tag{4.40}$$

REMARK 3 Theorem 5 provides a scheme to identify the most likely model hypothesis explaining a given observation. It should be noted that the most likely model hypothesis need not correspond to the true model, it will simply identify the most likely model from the collection of candidate hypotheses.

V. EXAMPLE

In this section we consider a scenario where a two-state Markov chain is observed in non-Gaussian noise taken to be a uniform distribution. The particular observation noise model we consider is uniform noise on $[-2, 2]$. Following the example in [3], we construct a Gaussian mixture approximation to the uniform density by minimising the L^1 norm. We take $M = 20$ and set all densities to equal weight, so that $\alpha_j = 1/20$ for all j. Further, all the means μ_j are regularly spaced Δ apart on $[-2, 2]$, starting from $-2+(2-(-2)/21)$. This spacing ensures matching the first moment exactly, that is

$$\tfrac{1}{2}(-2+2) = (1/20)\sum_{j=1}^{M}(-2+j\Delta) = 0. \tag{5.41}$$

We assume that the standard deviations for each of the 20 densities are the same, so the optimisation, (curve fitting), problem we wish to solve is

$$\widehat{\sigma} \triangleq \min_{\sigma \in \mathbb{R}_+} \int_{\mathbb{R}} |\tfrac{1}{4}\mathbf{1}_{[-2,2]} - \Psi^{(20)}(\xi;\Theta)| d\xi. \tag{5.42}$$

We note that since the uniform density has compact support and that $\Psi^{(20)}(\xi;\Theta)$ is a valid probability density function, the integration domain of this optimisation problem can be reduced to a compact set, that is

$$\begin{aligned}\min_{\sigma \in \mathbb{R}_+} \int_{\mathbb{R}} |\tfrac{1}{4}\mathbf{1}_{[-2,2]} - \Psi^{(20)}(\xi;\Theta)| d\xi. = \\ 1 + \min_{\sigma \in \mathbb{R}_+} \int_{[-2,2]} \Big\{ \big| \tfrac{1}{4}\mathbf{1}_{[-2,2]} - \Psi^{(20)}(\xi,\Theta) \big| - \\ \Psi^{(20)}(\xi;\Theta) \Big\} d\xi. \end{aligned} \tag{5.43}$$

Using numerical quadrature to estimate the right member of 5.43 we obtain $\widehat{\sigma} \approx 0.0901$.

The parameter values for the state process and observation process dynamics were as follows

$$A \triangleq \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix} \tag{5.44}$$

and

$$\boldsymbol{g} \triangleq (-1, 1.3)', \qquad \boldsymbol{d} \triangleq (1, 1)'. \tag{5.45}$$

The computer simulation results for our simulation scenario are shown below in Figures 1 and 2.

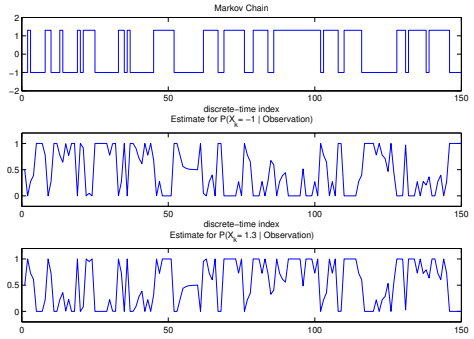

Fig. 1. In this set of three subplots we show the partially observed Markov chain, (2-state), and the estimated (filter) probabilities corresponding to each state.

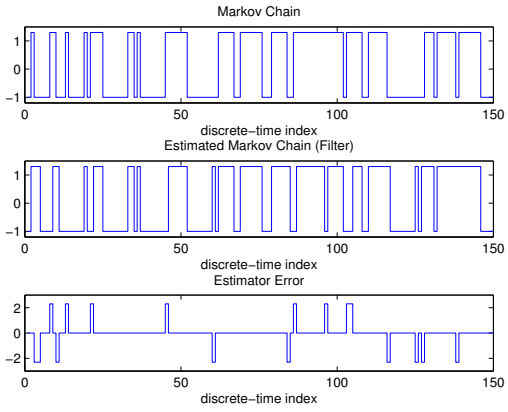

Fig. 2. Here we show a typical realisation of a MAP estimated state process and its corresponding finite state error process.

VI. Conclusion

In this article state estimation schemes for HMMs in non-Gaussian noise scenarios were developed. The techniques used to compute these schemes was the method of reference probability.

The motivation for this work was driven by the need to extend HMM estimation schemes beyond the usual Gaussian noise assumption. Moreover, our aim was to effectively consider *all* practical noise models by appealing to the efficacy of Gaussian mixture approximations of densities. The schemes we developed were: a recursive filter, its associated N-step ahead predictor, a general smoothing Theorem, and an M-ary Detection scheme. Our formulations are also in a form amenable to the usual and important analysis, such as stability analysis etc.

References

[1] J. Korevaar, *Mathematical Methods, Volume 1*. Academic Press New York, 1968.

[2] L. Aggoun and R. J. Elliott, *Measure Theory and Filtering, Introduction and Applications*. Cambridge University Press, 2004.

[3] H. W. Sorenson and D. L. Alspach, "Recursive bayesian estimation using gaussian sums," *Automatica*, vol. 7, pp. 465–479, 1971.

[4] D. L. Alspach and H. W. Sorenson, "Nonlinear bayesian estimation using gaussian sum approximations," *IEEE Trans. Automat. Contr.*, vol. AC-17, no. 4, pp. 439–448, August 1972.

[5] X. Wang and R. Chen, "Blind turbo equalization in gaussian and non-gaussian noise," *IEEE Trans. Veh. Technol.*, vol. 50, no. 4, pp. 1092–1105, July 2001.

[6] ——, "Adaptive bayesian multiuser detection for synchronous cdma with gaussian and non-gaussian noise," *IEEE Trans. Signal Processing*, vol. 48, no. 7, pp. 2013–2029, July 2000.

[7] P. Spasojevic and X. Wang, "Improved robust multiuser detection in non-gaussian channels," *IEEE Signal Processing Lett.*, vol. 8, no. 3, pp. 83–86, March 2001.

[8] X. Wang and H. V. Poor, "Robust multiuser detection in non-gaussian channels," *IEEE Trans. Signal Processing*, vol. 47, no. 2, pp. 289–305, February 1999.

[9] R. J. Elliott, L. Aggoun, and J. P. Moore, *Hidden Markov Models Estimation and Control*, ser. Applications of Mathematics. Springer Verlag, 1995, no. 29.

[10] H. V. Poor, *An introduction to signal detection and estimation*, 2nd ed. Springer Verlag, 1994.