ESTIMATION OF STATE-SPACE MODELS WITH GAUSSIAN MIXTURE PROCESS NOISE

Sina Miran[1,2], Jonathan Z. Simon[1,2,3], Michael C. Fu[2,4], Steven I. Marcus[1,2], and Behtash Babadi[1,2]

Department of Electrical and Computer Engineering[1], Institute for Systems Research (ISR)[2],
Department of Biology[3], Robert H. Smith School of Business[4],
University of Maryland, College Park, MD, US
{smiran,jzsimon,mfu,marcus,behtash}@umd.edu

ABSTRACT

State-space models are widely used to estimate latent dynamic processes from noisy and low-dimensional observations. When applying these models to real data, it is commonly assumed that the state dynamics are governed by Gaussian statistics. However, this assumption does not hold in applications where the process noise is composed of various exogenous components with heterogeneous statistics, resulting in a multimodal distribution. In this work, we consider a state-space model with Gaussian mixture process noise to account for such multimodality. We integrate the Expectation Maximization algorithm with sequential Monte Carlo methods to jointly estimate the Gaussian mixture parameters and states from noisy and low-dimensional observations. We validate our proposed method using simulated data inspired by auditory neuroscience, which reveals significant gains in state estimation as compared to widely used techniques that assume Gaussian state dynamics.

Index Terms— state-space modeling; Gaussian mixture models; expectation maximization; particle filtering and smoothing.

1. INTRODUCTION

State-space models are among the most commonly-used frameworks for analyzing dynamical systems, with application domains including control [1], tracking [2], and most recently neuroscience [3, 4, 5, 6]. These models describe the dynamics of a latent process (i.e., the states) as well as a measurement mechanism that results in limited and noisy observations. As such, they often consist of two equations: the state (evolution) equation, and the observation equation. To model the state evolution and measurement uncertainty, additive noise terms are often considered in the state and observation equations, respectively. In most cases, domain-specific expert knowledge of the problem is used to construct these equations, and model parameters such as the noise characteristics are estimated empirically with procedures such as the Expectation Maximization (EM) algorithm [7, 8]. In the simplest case of linear dynamics with Gaussian statistics, minimum mean square error (MMSE) state estimation can be performed using the well-known Kalman filter and smoother, which have been extended to also incorporate non-linear models [9]. For non-Gaussian statistics, Gaussian sum filter/smoother [10] or Sequential Monte Carlo (SMC) [11] methods have been widely used for state estimation.

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under grant number N660011824024. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

In many applications, independent and identically distributed (i.i.d.) Gaussian statistics are assumed for the noise terms [3, 12, 13], which results in convenient closed-form solutions. This assumption is often consistent with the empirical histogram of observation noise representing measurement uncertainty. However, the empirical histogram of the process noise, computed from state estimates, heavily depends on how the latent process evolves during the course of an experiment, which in turn depends on the specific experimental condition and exogenous variables. Thus, the Gaussian assumption for process noise is often violated in real-world applications [14]. In some applications of interest, this histogram exhibits a multimodal shape, where different modes correspond to different exogenous processes that drive the states. For instance, in the context of auditory processing in the brain, presentation of auditory stimuli may cause specific components of the latent state variables representing the underlying neural processes to abruptly increase, decrease, or stay relatively constant [15]. Therefore, a multimodal process noise model would capture the state dynamics more realistically. As a result, incorporating a more accurate representation of the process noise statistics will improve the state estimates, specially in presence of excessive observation noise. However, these benefits come at the cost of jointly estimating a more complex noise model as well as the states, which can be a challenging problem when the observations lie in a noisy and low-dimensional projection of the states.

In this work, we address this problem by considering a Gaussian mixture process noise, which can, in principle, approximate any multimodal density [16]. Although state estimation under fixed Gaussian mixture process noise has been studied using either SMC methods or a Gaussian sum filter/smoother [17, 10], a framework to jointly estimate the noise parameters and states from the observed data is lacking. Inspired by the classic applications of EM in clustering literature [18], we integrate the EM and SMC frameworks and develop an algorithm to estimate the Gaussian mixture parameters from the state-space observations. We examine the performance of our proposed method using simulated data inspired by speech processing in the brain. Our results show that the proposed algorithm is capable of accurately recovering the Gaussian mixture parameters, and reveal significant performance gains over the commonly used methods that assume Gaussian process noise statistics.

2. PROBLEM FORMULATION AND PROPOSED SOLUTION

Consider the following state-space model with additive noise:

$$\begin{cases} \mathbf{x}_t = f_t(\mathbf{x}_{t-1}) + \mathbf{w}_t \\ \mathbf{y}_t = g_t(\mathbf{x}_t) + \mathbf{v}_t \end{cases} \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^p$, $\mathbf{y}_t \in \mathbb{R}^q$, and the functional forms of f_t and g_t are known for $t = 1, \ldots, T$ using domain-specific knowledge of the problem. Also, assume $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ for the observation noise. We assume $\mathbf{R}$ to be known for simplicity; it is also possible to estimate $\mathbf{R}$ within the forthcoming EM framework in a straightforward fashion [7]. We model the state dynamics over $K := T/W$ non-overlapping windows of length W each. Consider a Gaussian mixture with M components and parameter set $\Theta := \{p_{1:M}, \boldsymbol{\mu}_{1:M}, \boldsymbol{\Sigma}_{1:M}\}$ containing the mixture probabilities $p_{1:M}$, mean vectors $\boldsymbol{\mu}_{1:M}$, and covariance matrices $\boldsymbol{\Sigma}_{1:M}$. Within each window, the process noise is drawn from one of the mixture components, which we denote by $z_i \in \{1, \ldots, M\}$ for $i = 1, \ldots, K$. Thus, $\mathbf{w}_t \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$ for $t = (i-1)W + 1, \ldots, iW$, and we consider the z_i's to be i.i.d. with $\mathrm{P}(z_i = m) = p_m$ for $m = 1, \ldots, M$. In other words, z_i determines the label of the mixture component that governs the state dynamics in window i. This can also be interpreted as a jumping or switching Gaussian process noise model. Note that the special case of $W = 1$ corresponds to fitting a Gaussian mixture model to the process noise in which the labels of the mixture components can vary at the same rate as that of the observations. Thus, the resulting model could approximate any arbitrary i.i.d. process noise $\mathbf{w}_t$.

Our goal is to compute the maximum likelihood estimate (MLE) of Θ from the observations $\mathbf{y}_{1:T}$. To do so, we use the EM algorithm with latent variables $\{z_{1:K}, \mathbf{x}_{1:T}\}$. EM provides iterative updates to the parameter estimates with provable guarantees of reaching a local optimum of the log-likelihood, and it can retrieve the MLE with proper initializations [19]. In each iteration, a lower bound to the log-likelihood is computed (E-step) and then maximized (M-step). These steps are outlined as below:

E-Step: Let $\widehat{\Theta}^{(\ell)}$ denote the parameter estimates at the ℓ^{th} iteration. In the $(\ell+1)^{\text{th}}$ iteration, the surrogate function or lower bound $Q(\Theta|\widehat{\Theta}^{(\ell)})$ is computed as below:

$$\begin{aligned} Q\left(\Theta \mid \widehat{\Theta}^{(\ell)}\right) &= \mathrm{E}\left\{\log \mathrm{P}\left(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, z_{1:K} \mid \Theta\right)\right\} \\ &= \mathrm{E}\left\{\log \mathrm{P}\left(\mathbf{x}_{1:T}, z_{1:K} \mid \Theta\right)\right\} + \mathsf{c}_1 \\ &= \sum_{i=1}^{K}\sum_{m=1}^{M} \mathrm{E}\left\{\mathbb{1}_{\{z_i=m\}}\left(\log p_m + \sum_{j=1}^{W} \log \pi_{i,j,m}\right)\right\} \\ &\quad + \mathsf{c}_2, \end{aligned} \tag{2}$$

where the expectations are with respect to $\mathbf{x}_{1:T}$ and $z_{1:K}$ given $\mathbf{y}_{1:T}$ and $\widehat{\Theta}^{(\ell)}$, $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function, the variables $\pi_{i,j,m}$ are defined as

$$\pi_{i,j,m} := \mathrm{P}\left(\mathbf{x}_{(i-1)W+j} \,\middle|\, \mathbf{x}_{(i-1)W+j-1}, z_i = m, \Theta\right), \tag{3}$$

computed based on the Gaussian density for $\mathbf{w}_{(i-1)W+j}$ in Eq. (1) for $z_i = m$, and c_1 and c_2 contain all terms not depending on Θ. If we decompose the expectation in Eq. (2) into two iterated expectations with respect to $\mathbf{x}_{1:T}$ given $\mathbf{y}_{1:T}$ and $\widehat{\Theta}^{(\ell)}$, and $z_{1:K}$ given $\mathbf{x}_{1:T}$ and $\widehat{\Theta}^{(\ell)}$, this equation can be written as

$$\begin{aligned} Q\left(\Theta \mid \widehat{\Theta}^{(\ell)}\right) = &\sum_{i=1}^{K}\sum_{m=1}^{M} \mathrm{E}\left\{\widehat{\epsilon}_{i,m}^{(\ell)}\left(\log p_m + \sum_{j=1}^{W} \log \pi_{i,j,m}\right)\right\} \\ &+ \mathsf{c}_3, \end{aligned} \tag{4}$$

where the expectation is with respect to $\mathbf{x}_{(i-1)W:iW}$ given $\mathbf{y}_{1:T}$ and $\widehat{\Theta}^{(\ell)}$, c_3 is constant w.r.t. Θ, and $\widehat{\epsilon}_{i,m}^{(\ell)} := \mathrm{P}\big(z_i = m \big| \mathbf{x}_{(i-1)W:iW}, \widehat{\Theta}^{(\ell)}\big)$. Using Bayes' rule for $\mathrm{P}\Big(z_i = m \Big| \mathbf{x}_{(i-1)W:iW}, \widehat{\Theta}^{(\ell)}\Big)$, we have

$$\widehat{\epsilon}_{i,m}^{(\ell)} := \frac{\hat{p}_m^{(\ell)} \prod_{j=1}^{W} \widehat{\pi}_{i,j,m}^{(\ell)}}{\sum_{m'=1}^{M} \hat{p}_{m'}^{(\ell)} \prod_{j=1}^{W} \widehat{\pi}_{i,j,m'}^{(\ell)}}, \tag{5}$$

where $\widehat{\pi}_{i,j,m}^{(\ell)}$ is defined similarly to Eq. (3) but for $\Theta = \widehat{\Theta}^{(\ell)}$, making it a constant with respect to Θ in Eq. (4).

M-Step: In this step, we maximize the log-likelihood lower bound in Eq. (4) with respect to Θ. Differentiating Eq. (4) with respect to Θ and invoking the dominated convergence theorem to change the order of expectation and differentiation, we obtain the following parameter updates for $m = 1, \ldots, M$:

$$\hat{p}_m^{(\ell+1)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{E}\{\widehat{\epsilon}_{i,m}^{(\ell)}\}, \tag{6}$$

$$\widehat{\boldsymbol{\mu}}_m^{(\ell+1)} = \frac{\sum_{i=1}^{K} \mathrm{E}\left\{\widehat{\epsilon}_{i,m}^{(\ell)} \sum_{j=1}^{W} \mathbf{u}_{i,j}\right\}}{W \sum_{i=1}^{K} \mathrm{E}\{\widehat{\epsilon}_{i,m}^{(\ell)}\}}, \tag{7}$$

$$\widehat{\boldsymbol{\Sigma}}_m^{(\ell+1)} = \frac{\sum_{i=1}^{K} \mathrm{E}\left\{\widehat{\epsilon}_{i,m}^{(\ell)} \sum_{j=1}^{W} \left(\mathbf{u}_{i,j} - \widehat{\boldsymbol{\mu}}_m^{(\ell+1)}\right)\left(\mathbf{u}_{i,j} - \widehat{\boldsymbol{\mu}}_m^{(\ell+1)}\right)^{\top}\right\}}{W \sum_{i=1}^{K} \mathrm{E}\{\widehat{\epsilon}_{i,m}^{(\ell)}\}}, \tag{8}$$

where $\mathbf{u}_{i,j} = \mathbf{x}_{(i-1)W+j} - f\big(\mathbf{x}_{(i-1)W+j-1}\big)$.

Due to the specific dependence of $\widehat{\pi}_{i,j,m}^{(\ell)}$'s, and consequently $\widehat{\epsilon}_{i,m}^{(\ell)}$'s, on the states, the expectations in Eq. (4) and in the update equations above are intractable even if the Gaussian mixture smoothing densities are known in closed-form [16]. To approximate the expectations, which are with respect to $\mathbf{x}_{(i-1)W:iW}$ given $\mathbf{y}_{1:T}$ and $\widehat{\Theta}^{(\ell)}$, we use the SMC method. Let N be the chosen number of samples or particles. To approximate $\widehat{\epsilon}_{i,m}^{(\ell)}$'s, we need N sample paths within each small window of length W. Particle smoothing approaches can provide us with such sample paths and their corresponding weights, which we respectively denote by $\mathbf{x}_{(i-1)W:iW}^{(n)}$ and $\lambda_i^{(n)}$ for $n = 1, \ldots, N$. Define $\widehat{\epsilon}_{i,m}^{(\ell,n)}$ similarly to Eq. (5) but when $\widehat{\pi}_{i,j,m}^{(\ell)}$'s are computed using both $\mathbf{x}_{(i-1)W:iW}^{(n)}$ and $\widehat{\Theta}^{(\ell)}$ in Eq. (3). Using the SMC sample paths, the parameter update rules become

$$\hat{p}_m^{(\ell+1)} = \frac{1}{K}\sum_{i=1}^{K}\sum_{n=1}^{N} \lambda_i^{(n)} \widehat{\epsilon}_{i,m}^{(\ell,n)}, \tag{9}$$

$$\widehat{\boldsymbol{\mu}}_m^{(\ell+1)} = \frac{\sum_{i=1}^{K}\sum_{n=1}^{N} \lambda_i^{(n)} \widehat{\epsilon}_{i,m}^{(\ell,n)} \sum_{j=1}^{W} \mathbf{u}_{i,j}^{(n)}}{W \sum_{i=1}^{K}\sum_{n=1}^{N} \lambda_i^{(n)} \widehat{\epsilon}_{i,m}^{(\ell,n)}}, \tag{10}$$

$$\widehat{\boldsymbol{\Sigma}}_m^{(\ell+1)} = \frac{\sum_{i=1}^{K}\sum_{n=1}^{N} \lambda_i^{(n)} \widehat{\epsilon}_{i,m}^{(\ell,n)} \sum_{j=1}^{W} \left(\mathbf{u}_{i,j}^{(n)} - \widehat{\boldsymbol{\mu}}_m^{(\ell+1)}\right)\left(\mathbf{u}_{i,j}^{(n)} - \widehat{\boldsymbol{\mu}}_m^{(\ell+1)}\right)^{\top}}{W \sum_{i=1}^{K}\sum_{n=1}^{N} \lambda_i^{(n)} \widehat{\epsilon}_{i,m}^{(\ell,n)}}, \tag{11}$$

where $\mathbf{u}_{i,j}^{(n)} = \mathbf{x}_{(i-1)W+j}^{(n)} - f(\mathbf{x}_{(i-1)W+j-1}^{(n)})$.

The E and M steps are repeated until convergence. Two remarks on the implementation of particle smoothing and the window size W are in order:

Remark 1 *The main challenge in the implementation of particle smoothing methods is their $\mathcal{O}(N^2)$ complexity. The two common implementations are referred to as the forward-backward smoother and the two-filter smoother [20]. The former method reweights the filtering particles according to future observations, while the latter aims to sample according to the smoothing densities. The two-filter smoother, however, requires the choice of an artificial distribution, which directly impacts the sampling quality. Some approximations have been considered in [20] and [21] to reduce the complexity to $\mathcal{O}(N \log N)$ and $\mathcal{O}(N)$, respectively. One can always use the filtering particles and their weights in Eqs.* (9), (10), *and* (11) *as one such approximation. Also, an approximation can be considered for the membership probabilities $\widehat{\epsilon}_{i,m}^{(\ell,n)}$ considering the Gaussian component used to generate each sample path: for each window i and sample path n, if the $m_0{}^{\text{th}}$ Gaussian component is used to generate the sample path, we can assume $\widehat{\epsilon}_{i,m}^{(\ell,n)} = 1$ for $m = m_0$ and zero otherwise.*

Remark 2 *In our method, the window length W should be chosen small enough to ensure that the state dynamics in each window are governed by only one of the mixture components. This can often be determined using domain specific knowledge of the problem. Also, we have to make sure the observation interval $[0, T]$ includes enough windows to reliably estimate the mixture probabilities p_m's. At the same time, W should be large enough so that the parameters $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ corresponding to the dynamics within the window can be estimated reliably. It should be noted that the dimension of the sample paths grows as W increases. As a result, more sample paths would be required to represent the densities $\mathbf{x}_{(i-1)W:iW}$ given $\mathbf{y}_{1:T}$ and $\widehat{\Theta}^{(\ell)}$. This should also be taken into account when choosing W and N.*

3. AN ILLUSTRATIVE SIMULATION STUDY

To demonstrate the benefits of our approach, we consider the problem of estimating the auditory Temporal Response Function (TRF) from neural recordings of a subject listening to continuous speech [22, 15]. The TRF is a linear kernel that relates the acoustic features of the auditory stimuli to the neural response recorded via electroencephalography (EEG) or magnetoencephalography (MEG). Let $y_t \in \mathbb{R}$ denote the auditory component of the neural response extracted from M/EEG recordings at time t. Also, let $\mathbf{s}_t \in \mathbb{R}^d$ be a vector containing the d most recent samples of the auditory stimulus at time t, and $\boldsymbol{\tau}_t \in \mathbb{R}^d$ be the TRF at time t. In order to enforce smoothness in $\boldsymbol{\tau}_t$, it is common to represent the TRF over a basis spanned by the columns of a fixed dictionary $\mathbf{G} \in \mathbb{R}^{d \times p}$ [15, 23]. Since typical TRFs exhibit smooth changes in time, we use the following state-space model to capture their dynamics:

$$\begin{cases} \mathbf{x}_t = \alpha \mathbf{x}_{t-1} + \mathbf{w}_t \\ \boldsymbol{\tau}_t = \mathbf{G}\mathbf{x}_t \\ y_t = \mathbf{s}_t^\top \boldsymbol{\tau}_t + v_t \end{cases} \tag{12}$$

where α is a constant, $\mathbf{w}_t$ is the process noise, and $v_t \sim \mathcal{N}(0, \sigma^2)$ is the observation noise. The parameter σ^2 can be estimated using stimulus-free M/EEG measurements [24]; hence, we consider it to be known. Comparing to the general model in Eq. (1), we have $f_t(\mathbf{x}_{t-1}) = \alpha \mathbf{x}_{t-1}$ and $g_t(\mathbf{x}_t) = \mathbf{s}_t^\top \mathbf{G}\mathbf{x}_t$. Although the coefficient α can be estimated from the observations in the EM framework, here we fix $\alpha < 1$ close to 1 for simplicity. Thus, the goal is to estimate the parameters of the process noise $\mathbf{w}_t$ and, consequently, the basis coefficients $\mathbf{x}_t \in \mathbb{R}^p$, i.e., the states, from the neural responses y_t given the stimulus feature vectors $\mathbf{s}_t$ and the other parameters of the model.

The TRFs can then be retrieved as $\widehat{\boldsymbol{\tau}}_t = \mathbf{G}\widehat{\mathbf{x}}_t$ from the state estimates. It is worth noting that this is in general a challenging problem even at high SNRs, as the observations lie in 1-dimensional projections of a higher-dimensional state with non-Gaussian dynamics. In order to simulate such a scenario, we consider $d = 20$, $p = 3$, and $\mathbf{G}$ to be a dictionary consisting of Gaussian kernels with small variances of 2.89 and peaks normalized to one and spread uniformly along the d samples. Also, we consider i.i.d. elements for $\mathbf{s}_t$'s from $\mathcal{N}(0, 10^{-6})$. In general, the stimulus vector can include the history of different speech features such as the acoustic envelope.

Fig. 1 shows the simulated TRFs used in this study. Panel A shows the states for $T = 1920$, and panel B shows a heatmap of the TRFs, where the column at time t corresponds to $\boldsymbol{\tau}_t$. This synthetic example is inspired by the dynamic TRFs extracted from real data in real-time [15, 23], where different peaks may arise, persist, and disappear over time according to the attentional state of the listener. As we observe, the local dynamics of the states are different in each of these conditions, which renders the multimodal distributional assumption on the process noise plausible. The peaks in the TRF indicate specific stimulus lags that are most relevant to the neural response at each time. We adopt our framework to estimate a Gaussian mixture representation for $\mathbf{w}_t$ in Eq. (12). Considering the different state dynamics in the second dimension (panel A), we choose $M = 4$ mixture components for inference, and we set $W = 20$ to be small enough compared to T. In practice, these parameters could be determined by cross-validation and/or incorporating domain-specific knowledge of the problem. We assume diagonal covariance matrices $\boldsymbol{\Sigma}_m$ to reduce the dimensionality of the parameters. Also, we perform particle smoothing using the forward-backward smoother in this simulation with $N = 100$ particles [20].

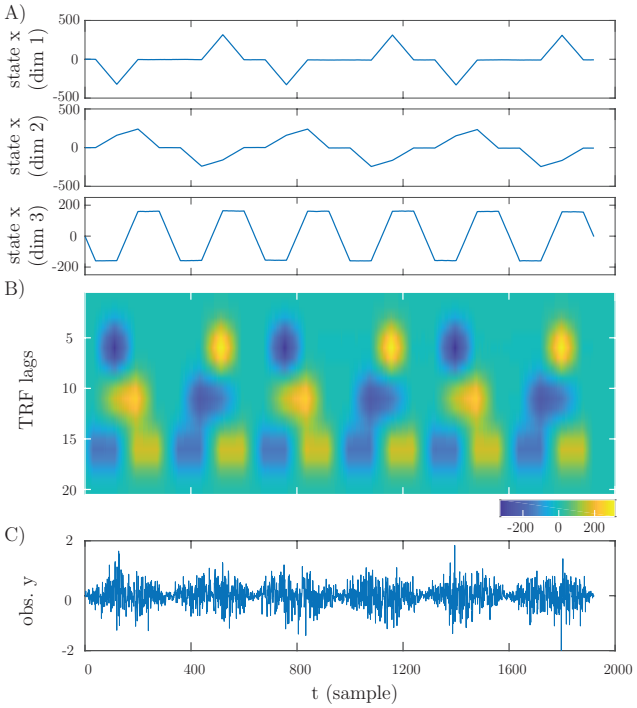

Fig. 1: Simulated TRF dynamics: (A) State evolution in time, (B) TRF heatmap vs. time, and (C) 1-dimensional observations.

Fig. 2 shows the estimated Gaussian mixtures after 1000 EM iterations (red curve) under a signal-to-noise ratio (SNR) of $\sim$ 15 dB. The EM algorithm is initialized with random means and large variances (yellow curve). As a benchmark, the histogram of the process noise $\widehat{\mathbf{w}}_t = \mathbf{x}_t - \alpha \mathbf{x}_{t-1}$ is also plotted (blue bars), which is only available to an oracle with access to the true states in Fig. 1-A. We refer to this benchmark as the oracle histogram. Visually speaking, the underlying multimodal distribution of the process noise is recovered accurately. Fig. 3 displays the EM convergence plots for the mixture parameters. Panels A and B show the convergence of p_m and $(\boldsymbol{\mu}_m)_2$ (as a representative dimension), respectively. The bold dashed lines are the parameters corresponding to a Gaussian mixture fitted to the oracle histogram in Fig. 2 (blue bars). After $\sim$200 EM iterations these parameters converge to the desired values from random initializations. Panel C displays the convergence of the variances $(\boldsymbol{\Sigma}_m)_{2,2}$ as a representative dimension. Notice that this plot is in log scale to better illustrate the convergence rate. An initial growth of the variances is observed so that the EM/SMC method can effectively probe the space of solutions, which is followed by convergence to the desired values (bold dashed lines). It takes $\sim$1000 EM iterations for the variances to converge, which is significantly higher than that needed for mixture probabilities and means. This effect was consistent in our simulations, as the mixture means and probabilities were estimated faster and more reliably than the variances.

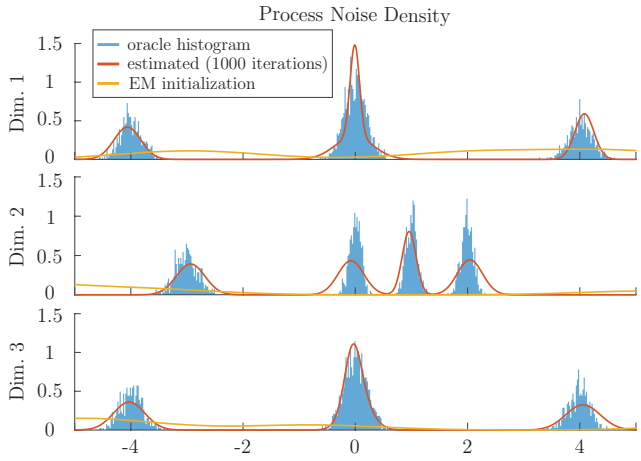

Fig. 2: Gaussian mixture estimation results for process noise in Eq. (12) corresponding to the simulated data in Fig. 1: EM initialization (yellow), estimated mixtures after 1000 EM iterations (red), and the oracle histogram of the process noise (blue) computed from the true states in Fig. 1-A.

Finally, we compare the root mean square error (RMSE) of the state estimates under different observation SNRs to that obtained from a Gaussian model in Eq. (12), where $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ and the general covariance matrix $\mathbf{Q}$ is estimated using the EM algorithm [7]. At each SNR value, the process noise parameters are estimated for the two models and then used to estimate the states for different observation realizations. State estimation in our mixture model and the Gaussian model is performed using particle and Kalman smoothing, respectively. Fig. 4 displays the normalized RMSE for the states in Fig. 1-A, averaged over 20 repeated trials with different observation noise realizations per SNR value. Although particle smoothing methods are suboptimal for parameter and state estimation, we observe that the multimodal representation of the process noise consistently helps in reducing the state estimation error at various

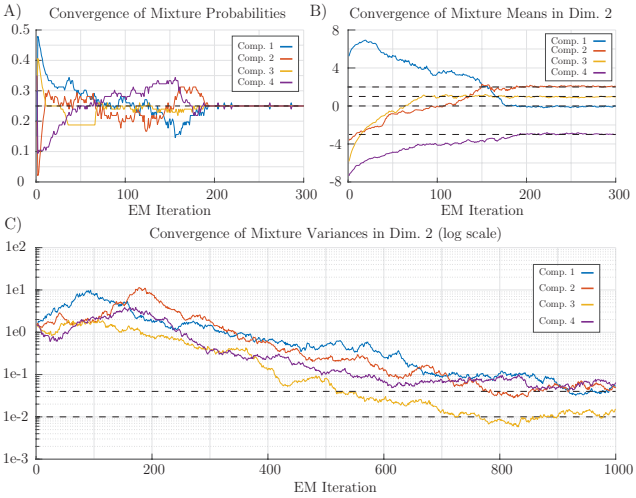

Fig. 3: Convergence plots of the parameters: (A) Mixture probabilities, (B) Projected mixture means onto the 2nd dimension, (C) Mixture variances in the 2nd dimension (log-scale).

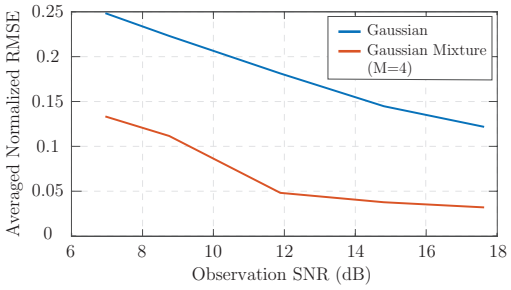

Fig. 4: Normalized RMSE vs. SNR. The RMSE values are obtained by averaging over 20 repeated trials per SNR. Blue curve corresponds to a linear Gaussian model, and red curve corresponds to a linear model with Gaussian mixture process noise. Parameters of both models are estimated at each SNR value and then used to estimate the states for different observation realizations. The richer representation of process noise results in higher state estimation accuracy.

SNR levels. The improved RMSE curve here comes at the expense of more computational complexity in model estimation.

4. CONCLUSION AND FUTURE WORK

We considered a class of state-space models with process noise following a multimodal distribution, where each mode corresponds to a separate exogenous process governing the state dynamics. We assumed a Gaussian mixture model to account for this multimodal distribution, and by integrating the EM algorithm and SMC methods, we developed an algorithm to estimate the Gaussian mixture parameters from noisy and limited observations. We illustrated the utility of our algorithm using a simulated example inspired by auditory processing in the brain. Our results show that the proposed algorithm reliably recovers the multimodal distribution of the process noise, and outperforms the commonly used methods that assume Gaussian state evolution. Future work includes exploring the possibility of closed-form alternative solutions to the particle smoothing method, which can considerably decrease the computational complexity, and applying our methodology to experimentally recorded data from auditory experiments.

5. REFERENCES

[1] Bernard Friedland, *Control system design: an introduction to state-space methods*, Courier Corporation, 2012.

[2] Hanxuan Yang, Ling Shao, Feng Zheng, Liang Wang, and Zhan Song, "Recent advances and trends in visual tracking: a review," *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, 2011.

[3] Wei Wu, Jayant E Kulkarni, Nicholas G Hatsopoulos, and Liam Paninski, "Neural decoding of hand motion using a linear state-space model with hidden states," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 4, pp. 370–378, 2009.

[4] Alireza Sheikhattar, Sina Miran, Ji Liu, Jonathan B Fritz, Shihab A Shamma, Patrick O Kanold, and Behtash Babadi, "Extracting neuronal functional network dynamics via adaptive granger causality analysis," *Proceedings of the National Academy of Sciences*, vol. 115, no. 17, pp. E3869–E3878, 2018.

[5] Ramin Bighamian, "Estimation of functional dependence in high-dimensional spike-field activity," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 2635–2638.

[6] Sina Miran, Sahar Akram, Alireza Sheikhattar, Jonathan Z Simon, Tao Zhang, and Behtash Babadi, "Real-time decoding of auditory attention from EEG via Bayesian filtering," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 25–28.

[7] Zoubin Ghahramani and Geoffrey E Hinton, "Parameter estimation for linear dynamical systems," Tech. Rep., Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science, 1996.

[8] Stuart Gibson and Brett Ninness, "Robust maximum-likelihood estimation of multivariable dynamic systems," *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.

[9] Brian DO Anderson and John B Moore, *Optimal filtering*, Courier Corporation, 2012.

[10] Genshiro Kitagawa, "The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother," *Annals of the Institute of Statistical Mathematics*, vol. 46, no. 4, pp. 605–623, 1994.

[11] Olivier Cappé, Simon J Godsill, and Eric Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[12] Hamidreza Abbaspourazad and Maryam M Shanechi, "An unsupervised learning algorithm for multiscale neural activity," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 201–204.

[13] Anne C Smith and Emery N Brown, "Estimating a state-space model from point process observations," *Neural Computation*, vol. 15, no. 5, pp. 965–991, 2003.

[14] Genshiro Kitagawa, "Non-Gaussian state—space modeling of nonstationary time series," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1032–1041, 1987.

[15] Sahar Akram, Jonathan Z Simon, and Behtash Babadi, "Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 8, pp. 1896–1905, 2017.

[16] Ba-Ngu Vo, Ba-Tuong Vo, and Ronald PS Mahler, "Closed-form solutions to forward–backward smoothing," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 2–17, 2012.

[17] Jayesh H Kotecha and Petar M Djuric, "Gaussian sum particle filtering," *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2602–2612, 2003.

[18] Rui Xu and D Wunsch II, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[19] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[20] Mike Klaas, Mark Briers, Nando De Freitas, Arnaud Doucet, Simon Maskell, and Dustin Lang, "Fast particle smoothing: If I had a million particles," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 481–488.

[21] Paul Fearnhead, David Wyncoll, and Jonathan Tawn, "A sequential smoothing algorithm with linear computational cost," *Biometrika*, vol. 97, no. 2, pp. 447–464, 2010.

[22] Nai Ding and Jonathan Z Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *American Journal of Physiology-Heart and Circulatory Physiology*, 2011.

[23] Sina Miran, Sahar Akram, Alireza Sheikhattar, Jonathan Z Simon, Tao Zhang, and Behtash Babadi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Frontiers in Neuroscience*, vol. 12, 2018.

[24] Behtash Babadi, Gabriel Obregon-Henao, Camilo Lamus, Matti S Hämäläinen, Emery N Brown, and Patrick L Purdon, "A subspace pursuit-based iterative greedy hierarchical solution to the neuromagnetic inverse problem," *NeuroImage*, vol. 87, pp. 427–443, 2014.