NON-PARAMETRIC BAYESIAN MEASUREMENT NOISE DENSITY ESTIMATION IN NON-LINEAR FILTERING

Emre Özkan[†], *Saikat Saha*[†], *Fredrik Gustafsson*[†], *and Václav Šmídl*[‡]

ABSTRACT

In this study, we investigate online Bayesian estimation of the measurement noise density of a given state space model using particle filters and Dirichlet process mixtures. Dirichlet processes are widely used in statistics for nonparametric density estimation. In the proposed method, the unknown noise is modeled as a Gaussian mixture with unknown number of components. The joint estimation of the state and the noise density is done via particle filters. Furthermore, the number of components and the noise statistics are allowed to vary in time. An extension of the method for the estimation of time varying noise characteristics is also introduced.

Index Terms— Particle Filtering, Dirichlet Process, Bayesian Estimation, Adaptive filtering, Marginalized Particle Filters.

1. INTRODUCTION

The use of particle filters under the model uncertainties has been an open problem over a decade in the field. The joint task of estimating the model unknowns and the hidden state within the same context makes the problem difficult to handle. When the uncertainties are limited to exist in the noise terms driving the model, more can be done by utilizing some statistical methods. In our previous work [1] we investigated the estimation of unknown additive Gaussian noise parameters, where the marginalization for the unknowns were made possible by using conjugate family of distributions. In this study, we aim to extend our previous work to the mixture of Gaussians for the measurement noise. Noise density estimation using finite Gaussian mixture model has been studied e.g., [2]. However in [2], the order selection problem is unresolved and the batch inference is done via Gibbs sampling. In our study we use Dirichlet Process (DP) based model which avoids the order selection problem, and do the inference with particle filters in an online fashion. Similar studies involving particle filters and DP have been addressed in [3],[4]. In [3] the system of interest is assumed to be linear. Both of these methods require explicit sampling of the mean and the covariance. In our formulation, sampling for the noise parameters is avoided. Instead we keep the sufficient statistics of the noise parameters and marginalize over them to make inference over discrete variables, which results in a more efficient scheme. It is stated in [3] that, sampling of the mean and the covariance in Rao-Blackwellised particle filter framework has the problem of moving the cluster parameters. In our method, making the inference over the discrete cluster variables, and the use of the exponential forgetting factor on the statistics help to circumvent this problem.

The rest of the paper is organized as follows. In Section 2, we introduce some necessary background information. Sections 3 and 4 give the problem definition and the methodology followed for the solution. Later the extensions of the proposed algorithm is mentioned and finally the simulation results and conclusions are given.

2. BACKGROUND

2.1. Dirichlet Process Mixtures

Dirichlet Processes are widely used in statistics for classification and mixture density estimation applications where the number of clusters or the mixture components is unknown a priori. Because of the space limitations, we will give a brief introduction to DP here. The interested readers are referred to [5][3][6]. The formulation given in [3] and [6] is relevant to our framework. Dirichlet Process Mixture(DPM) defines a Hierarchical nonparametric Bayesian model for the unknown probability distributions. The clustering property of Dirichlet Processes is well suited in the context of mixture density estimation where each cluster corresponds to a mixture component of the unknown probability density function. For an unknown distribution $F(.)$ the following nonparametric model is used.

$$F(x) = \int f(x|\theta) d\mathbb{G}(\theta) \tag{1}$$

where $\mathbb{G} \sim DP(\mathbb{G}_0, \alpha)$ and $f(x|\theta)$ is a kernel having the parameters θ. The Dirichlet Process Mixture (DPM) model defines the following hierarchical Bayesian structure.

$$\mathbb{G} \sim DP(\mathbb{G}_0, \alpha) \tag{2}$$

$$\theta_i|\mathbb{G} \sim \mathbb{G} \tag{3}$$

$$y_i|\theta_i \sim f(.|\theta_i) \tag{4}$$

where y_i's are considered to be the samples from a mixture, and the mixing distribution $\mathbb{G}$ is following a DP prior.

2.2. Normal-Inverse Wishart priors

The use of conjugate priors is essential in our formulation. We aim to integrate out many of the parameters in the model and make the inference over the discrete variables and the hidden

[†] Emre Özkan, Saikat Saha and Fredrik Gustafsson are with the Department of Electrical Engineering, Linköping University, Linköping, Sweden,{emre, saha, fredrik}@isy.liu.se

[‡] Václav Šmídl is with the Institute of Information Theory and Automation, Prague, Czech Republic, smidl@utia.cas.cz

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

state. For this reason, we use a Normal-inverse-Wishart distribution as our base distribution $\mathbb{G}_0$. For multivariate Normal data z of dimension d, with unknown mean μ and covariance Σ, a Normal-inverse-Wishart distribution defines a conjugate prior. Let us denote it as $[\mu, \Sigma] \sim \mathrm{NiW}(\nu, V)$. Assuming a Normal-inverse-Wishart distribution as a prior defines a hierarchical Bayesian model given below:

$$z \sim \mathcal{N}(\mu, \Sigma), \quad \mu|\Sigma \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma}), \quad \Sigma \sim \mathrm{iW}(\nu - d, \Lambda) \quad (5)$$

where $\mathrm{iW}(.)$ denotes the Inverse Wishart distribution. The parameters ν and V represent the sufficient statistics and can be updated recursively. The relevant quantities are defined as,

$$\hat{\mu} = V_{11}^{-1} V_{1z}, \quad (6)$$

$$\hat{\Sigma} = V_{11}^{-1} \Sigma, \quad (7)$$

$$\Lambda = V_{zz} - V_{1z} V_{11}^{-1} V_{z1}, \quad (8)$$

$$V = \begin{pmatrix} V_{zz} & V_{1z} \\ V_{z1} & V_{11} \end{pmatrix}, \quad (9)$$

where d denotes the dimension of measurement vector z, then V_{zz} is defined as the upper-left $d \times d$ sub-block of $V \in \mathbb{R}^{(d+1)\times(d+1)}$

Via conjugacy, the posterior distribution is again a normal-inverse-Wishart distribution with updated statistics. The update equations of the statistics are as follows,

$$\bar{V} = V + \begin{pmatrix} z \\ 1 \end{pmatrix} \begin{pmatrix} z^T & 1 \end{pmatrix} = \begin{pmatrix} V_{zz} & V_{z1} \\ V_{1z} & V_{11} \end{pmatrix} \quad \text{(10a)}$$

$$\bar{\nu} = \nu + 1 \quad \text{(10b)}$$

Furthermore, the predictive distribution for z becomes a t-distribution for a NiW prior.

$$p(z|\nu, V) = t_{\nu-d+1}\left(\mu, \frac{(1 + V_{11})}{(\nu - d + 1)V_{11}} \Lambda\right) \quad (11)$$

where μ and Λ are computed according to (6) and (8). $t_v(\mu, \zeta)$ is the multivariate student-t distribution with v degrees of freedom, located at μ with scale parameter ζ.

3. PROBLEM FORMULATION

Consider the following nonlinear discrete time state space model describing the dynamics of the hidden state x_t and its relation with the observation y_t

$$x_t = f_t(x_{t-1}) + v_t \quad (12)$$

$$y_t = h_t(x_t) + w_t \quad (13)$$

Here t denotes the time index. $f(.)$ and $h(.)$ are possibly nonlinear functions of the state vector x_t. v_t is Gaussian process noise with known mean and covariance. w_t is the measurement noise having an unknown noise distribution $\mathbb{G}_w$. We will use a DPM model for the unknown noise distribution such that,

$$\mathbb{G}_w \sim DP(\mathbb{G}_0, \alpha) \quad (14)$$

$$\theta_t | \mathbb{G}_w \sim \mathbb{G}_w \quad (15)$$

$$w_t | \theta_t \sim \mathcal{N}(\mu_t, \Sigma_t) \quad (16)$$

Where $\theta_t = (\mu_t, \Sigma_t)$ is the mean and the covariance of the cluster.

We aim to establish an algorithm which is capable of estimating online the hidden state x_t as well as the unknown noise distribution including its variation in time. This indeed is a very difficult problem as it is hard to find a flexible algorithm which can adopt to changes in the noise statistics while the ambiguity in the unobserved state is inherent. The main difficulty of the problem arises from the fact that the estimation of the hidden state also depends on the unknown noise statistics. Therefore the joint estimation of the unobserved state and the noise statistics is required.

4. METHODOLOGY

In order to obtain analytical substructures which are essential for marginalization of the joint density, we utilize the *cluster variables* or *labels*. We assume that the measurement noise term w_t in (13) is i.i.d. according to a mixture density and the cluster variables c_t is defined for each measurement y_t indicating the specific component of the unknown mixture from which the noise w_t is sampled from. Next, we define our *unknowns* as $\bar{\theta} = [\theta, c_{0:t}]$ (θ denotes the mean and the covariance of the mixture component). Then the target density to be estimated becomes,

$$p(\bar{\theta}, x_{0:t}|y_{0:t}) = p(\bar{\theta}|x_{0:t}, y_{0:t}) p(x_{0:t}|y_{0:t}) \quad (17)$$

$$= p(\theta|c_{0:t}, x_{0:t}, y_{0:t}).p(c_{0:t}, x_{0:t}|y_{0:t}) \quad (18)$$

Conditional on the measurements, their labels and the unobserved states, the sufficient statistics of θ can be computed analytically. Here we will make use of a conjugate family of distributions such that the posterior density of the parameters of the specific component of the mixture $p(\theta|c_{0:t}, x_{0:t}, y_{0:t})$ will follow normal-inverse Wishart distribution. Moreover, in the measurement likelihood computations we will fully make use of the underlying conjugacy by integrating out the cluster parameters θ's and use the student-t as the predictive distribution. The inference is done by approximating the nonlinear state $x_{0:t}$ and the cluster variables $c_{0:t}$ by particles. The joint density we want to approximate with particles, which appears as the second factor in (18), admits the following recursion.

$$p(\xi_{0:t}|y_{0:t}) \triangleq p(c_{0:t}, x_{0:t}|y_{0:t}) \quad (19)$$

$$= p(\xi_{0:t-1}|y_{0:t-1}) \frac{p(y_t|y_{0:t-1}, \xi_{0:t}) p(\xi_t|\xi_{0:t-1})}{p(y_t|y_{0:t-1})} \quad (20)$$

The prior distribution of the cluster variables and the state can be factorized as follows,

$$p(c_{0:t}, x_{0:t}) = \prod_{k=1}^{t} p(c_k|c_{0:k-1}) \prod_{k=1}^{t} p(x_k|x_{k-1}). \quad (21)$$

If the approximation at time $t-1$ is available as the particles $\{\xi_{t-1}^{(i)}\}_{i=1}^{N}$, $\{w_{t-1}^{(i)}\}_{i=1}^{N}$ and the weights can be updated as,

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p(y_t|y_{0:t-1}, x_{0:t}^{(i)}, c_{0:t}^{(i)}) p(x_t^{(i)}|x_{t-1}^{(i)}) p(c_t^{(i)}|c_{0:t-1}^{(i)})}{q(c_t^{(i)}|c_{0:t-1}^{(i)}, y_t) q(x_t^{(i)}|x_{t-1}^{(i)}, y_t)} \tag{22}$$

Each particle, will hold the sufficient statistics of its clusters, and the measurement likelihood will be computed using the sufficient statistics. Algorithm iterations are given as a pseudo-code at the end of the section. The $p(c_t^{(i)}|c_{0:t-1}^{(i)})$ will be in accordance with the famous Chinese Restaurant Process induced by the DP.

$$p(c_{n+1} = c_j|c_1, c_2, ..., c_n) = \frac{n_{c_j}}{\sum_i n_{c_i} + \alpha} \tag{23}$$

$$p(c_{n+1} = c_{new}|c_1, c_2, ..., c_n) = \frac{\alpha}{\sum_i n_{c_i} + \alpha} \tag{24}$$

where c_j is the label of one of the existing clusters that appeared in the set $\theta_1, \theta_2, .., \theta_n$ and c_{new} is to be the label for a new cluster, n_{c_i} is the number of measurements previously assigned to cluster c_i. In the measurement likelihood computation, the unknown parameters of the noise can be integrated out such that,

$$p(y_t|y_{0:t-1}, x_{0:t}^{(i)}, c_{0:t}^{(i)}) = \int p(y_t|\theta_t) p(\theta_t|y_{0:t-1}, x_{0:t}^{(i)}, c_{0:t}^{(i)}) d\theta_t. \tag{25}$$

The posterior distribution of $p(\theta_t|y_{0:t-1}, x_{0:t}^{(i)}, c_{0:t}^{(i)})$ will follow and NiW distribution whose sufficient statistics can be computed through the given conditions. Then the integral above will be the student t distribution as in (11)

It is possible to explore the full support of $p(c_t|c_{0:t-1})$ as it takes a finite number of discrete values for each particle [7]. Then the particles having the N-best weights can be kept through the next step. Resampling still can be done if necessary. In the simulations, 100- best particles are kept at each time step.

4.1. Time varying noise distribution

4.1.1. Exponential forgetting

In order to adapt the changes in the noise distribution, it is possible to utilize the principle of forgetting in updating the noise statistics. The simplest form is known as exponential forgetting, where the update equations (10a)-(10b) are replaced by

$$\bar{V} = \lambda V + \begin{pmatrix} z \\ 1 \end{pmatrix} \begin{pmatrix} z^T & 1 \end{pmatrix}, \quad \bar{\nu} = \lambda \nu + 1. \tag{26}$$

where the forgetting factor $0 \leq \lambda \leq 1$ is a scalar real number. The use of this operation correspond to application of exponential window with effective length $h = \frac{1}{1-\lambda}$. The statistics relies on roughly the measurements within last h frames/time instances. That allows the algorithm to adapt the changes in the noise statistics in time.

```
• Iterations:
• For $t = 1, 2, \ldots$ do
    – For each particle $i = 1, .., N$ do
        * sample $x_t^{(i)} \sim q(x_t^{(i)}|y_t, x_{t-1}^{(i)})$
        * sample $c_t^{(i)} \sim q(c_t^{(i)}|y_t, c_{0:t-1}^{(i)}, x_t^{(i)})$
    – For $i = 1, .., N$, update the weights
      $w_t^{(i)} =$
      $w_{t-1}^{(i)} \frac{p(y_t|y_{0:t-1}, x_{0:t}^{(i)}, c_{0:t}^{(i)}) p(x_t^{(i)}|x_{t-1}^{(i)}) p(c_t^{(i)}|c_{0:t-1}^{(i)})}{q(c_t^{(i)}|c_{0:t-1}^{(i)}, y_t) q(x_t^{(i)}|x_{t-1}^{(i)}, y_t)}$
    – Update statistics of the noise, using the pseudo measurements in $z_t^{(i)} = y_t - h_t(x_t^{(i)})$ with the equations (10a)-(10b).
    – Normalize weights, $\omega_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{i=1}^{N} \tilde{w}_t^{(i)}}$.
    – Compute $N_{\text{eff}} = \frac{1}{\sum_{i=1}^{N} (\omega_t^{(i)})^2}$.
        * If $N_{\text{eff}} \leq \eta$, Resample the particles, and set $\omega_t^{(i)} = 1/N$.
```

4.1.2. r-order Markov model

In order to track the changes in the number of components of the unknown noise distribution, we condition the prior probabilities of the cluster variables in a sliding window fashion. In that case the probabilities in (23) and (24) are computed by considering the values of the cluster variables for the last r time steps, $c_{t-r:t}$ instead of the whole past values $c_{0:t}$ i.e., $p(c_t|c_{0:t-1}) \approx p(c_t|c_{t-r:t-1})$. The clusters which have not been updated for the last r time steps are deleted. A number of methods to extend DPM for time varying densities are proposed in [8] which can be applied here.

5. SIMULATIONS

We use the following benchmark scalar nonlinear time series model for our illustrations:

$$x_t = \frac{x_{t-1}}{2} + \frac{25 x_{t-1}}{1 + x_{t-1}^2} + 8\cos(1.2t) + v_t, \tag{27}$$

$$y_t = \frac{x_t^2}{20} + w_t, \quad t = 1, 2, \ldots \tag{28}$$

where $v_t \sim N(0, 1)$ and $w_t \sim \sum_{i=1}^{K_t} \pi_{i,t} \mathcal{N}(\mu_{i,t} \Sigma_{i,t})$. Here, the measurement noise distribution is changed every 500 time steps. Figure 1 and Figure 2 illustrate the variation of the true noise density and the estimated noise density in time. Figure 3 shows the *slices* of the same result for a single time step explicitly showing estimated and the true noise densities. The relevant marginalizations in the formulation results an efficient algorithm such that only 100 particles with highest weights are kept at each time step throughout the simulation. NiW distribution is used as the base distribution $\mathbb{G}_0$ and the initial parameters are set to $[\nu_0, V_0] = [(5, \begin{pmatrix} 15 & 0 \\ 0 & 1 \end{pmatrix})]$. Other variables are, $\alpha = 2$, forgetting factor $\lambda = 0.98$, and the order of Markov model $r = 30$. The transition density $p(x_t|x_{t-1})$ is used as the importance density while sampling x_t.

6. CONCLUSION

A novel algorithm is proposed for noise density estimation in nonlinear models. DPM, which defines flexible priors to estimate the unknown number of components in the mixtures, are utilized for the unknown distribution of the noise. The inference is done via particle filters and the estimation of the unknowns are made online. Time varying extension of the method is also provided. The performance of the algorithm is illustrated on a numeric example.

7. REFERENCES

[1] S. Saha, E. Özkan, F. Gustafsson, and V. Smidl, “Marginalized particle filters for Bayesian estimation of Gaussian noise parameters,” in *Proceedings of International Conference on Information Fusion*, Edinburgh, Scotland, July 2010.

[2] L. Mihaylova and D. Angelova, “Noise parameters estimation with Gibbs sampling for localization of mobile nodes in wireless networks,” in *Proceedings of 13th International Conference on Information Fusion*. ISIF, 2010.

[3] F. Caron, M. Davy, A. Doucet, E. Duflos, and P. Vanheeghe, “Bayesian inference for linear dynamic models with Dirichlet process mixtures,” *Signal Processing, IEEE Transactions on*, vol. 56, no. 1, pp. 71 –84, jan. 2008.

[4] N. Viandier, J. Marais, A. Rabaoui, , and E. Duflos, “GNSS pseudorange error density tracking using Dirichlet process mixture,” in *Proceedings of 13th International Conference on Information Fusion*. ISIF, 2010.

[5] Y. W. Teh, “Dirichlet processes,” in *Encyclopedia of Machine Learning*. Springer, 2010.

[6] E. Özkan, I. Y. Özbek, and M. Demirekler, “Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time-varying Dirichlet process mixture models,” *Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 8, pp. 1518–1532, 2009.

[7] P. Fearnhead, “Particle filters for mixture models with an unknown number of components,” *Statistics and Computing*, vol. 14, pp. 11–21, 2004.

[8] F. Caron, M. Davy, and A. Doucet, “Generalized Polya urn for time-varying Dirichlet process mixtures,” in *International Conference on Uncertainty in Artificial Intelligence, Vancouver, Canada*, 2007.

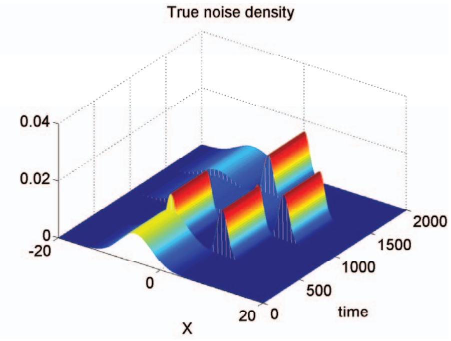


Fig. 1. Variation of true noise density in time.

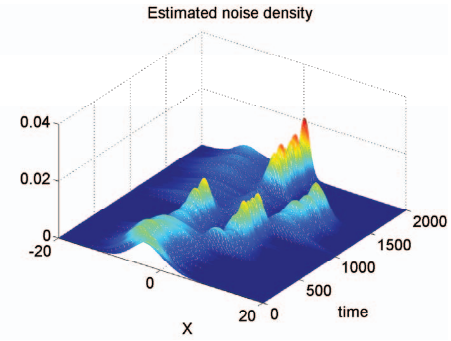
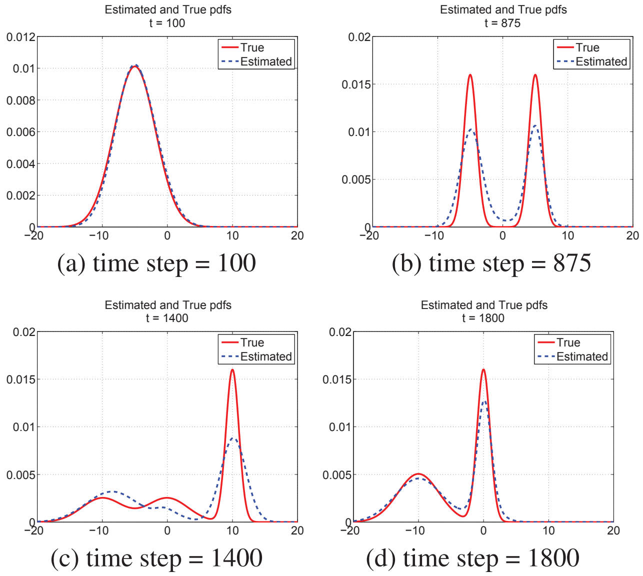


Fig. 2. The estimated noise density.



(a) time step = 100

(b) time step = 875

(c) time step = 1400

(d) time step = 1800

Fig. 3. True and estimated noise densities at different time steps.