Automated Splitting Gaussian Mixture Nonlinear Measurement Update

Kirsten Tuggle[1] and Renato Zanetti[2]
The University of Texas at Austin, Austin, Texas 78712

I. Introduction

Orbit determination, ground tracking, as well many other aerospace and non-aerospace applications, require very accurate estimates of system states paired with accurate characterizations of the uncertainties associated with those estimates. Furthermore, this is often in the presence of nonlinear dynamics and nonlinear measurement models. Nonlinear Bayesian estimation techniques are often applied, and due to constrained computational abilities, approximations to the true Bayesian solution are often used. In doing so, a choice is made among the competing notions of computational efficiency and estimation performance. For example, the Extended Kalman Filter (EKF) is relatively computationally simple but with the possibility of degraded performance in comparison to the more computationally taxing Particle Filter (PF) [1]. Gaussian mixture models (GMMs) offer a promising balance between these notions, garnering increased interest since the early 1970s [2]. In fact, particle representations used within PFs may be considered to be a certain class of Gaussian mixtures restricted to infinitesimal covariances [3]. This view of the GMM as a generalized version of the PF introduces a key attraction for its use - that is, retaining the ability of a PF to represent a wide class of probability density functions (pdfs) via a finite parameterization while also offering potential for reduced dimension within its parameterization due to flexibility in sizing component covariances. Additionally, algorithms utilizing GMMs can be facilitated by intuition based on Gaussian distributions and invocation of their associated properties. These strengths make GMMs an

---

[1] PhD Student, Department of Aerospace Engineering and Engineering Mechanics.
[2] Assistant Professor, Department of Aerospace Engineering and Engineering Mechanics

enticing option for any application relying on approximation of non-Gaussian pdfs. Thus, GMMs have been used across many technical disciplines including target tracking [1] and collision prediction [4, 5] within the aerospace community as well as clustering and Expectation-Maximization (EM) within probabilistic learning [6].

The use of GMMs corresponds to representing probability density functions not by their first two moments, or a collection of sigma points or particles, but rather by a collection of Gaussian components. The aggregate pdf, given by Eq. (1), is simply a weighted sum of the individual Gaussian pdfs.

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{i=1}^{N} w_i N(\boldsymbol{x}; \bar{\boldsymbol{x}}_{\boldsymbol{i}}, P_i) \tag{1}$$

Here, the number of components is denoted by N, individual weights are denoted by w_i, and individual Gaussian distributions parameterized by means $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ and covariances P_i are denoted by $N(\boldsymbol{x}; \bar{\boldsymbol{x}}_{\boldsymbol{i}}, P_i)$. A few comments will be provided for intuition concerning GMMs. Since the integral of any pdf over its support must equal 1, the integral of each Gaussian pdf in Eq. (1) must also be 1. The weight terms then, which must necessarily sum to 1 by the same property, represent the individual probabilities contributed by each component in this summation relative to that contributed by the remaining components. Considering the set of mixture components to be collection of probability entities which together represent an overall pdf, the weight of a component may be thought of as the normalized probability of that entity. The means place the entities within the state-space, and the covariances confine the entities about their means. The number of entities/components can be increased by the action of splitting a single component into a set of multiple components intended to approximate it, or decreased by merging a set of components into a single one intended to be representative of the set. Adaptation via these two actions seeks to manage the priorities of restricting component covariances for sufficient linearized estimator accuracy and limiting the number of components for computational feasibility. Effectiveness in managing these priorities relies on intelligent execution of splitting and merging.

Due to the accumulation of nonlinear effects by dynamics and the accruement of measurements, splitting is the more common action. Additionally, many splitting techniques have analogous merging counterparts. For these reasons, only the action of splitting is addressed in this work. Note

that splitting refers to choosing a direction along which a set of Gaussian components are placed in a manner to reapproximate the original component as a mixture itself. This has been referred to as a 1-D method [3], to which nearly all splitting schemes subscribe. Since prior uncertainty is only reduced along the splitting direction, the scheme must often be recursively applied to address other significant regions of the state space. Ref. [3] notes that 1-D methods are vulnerable to the curse of dimensionality when strong nonlinearities exist only within a subspace of the support and offers a method utilizing linear matrix inequalities to remedy this as well as consolidate splitting and merging actions. Since nonlinear models from orbital motion and tracking sensors are relatively tame, 1-D methods have nonetheless been useful.

The problem of designing a splitting routine requires answering two main questions: For the current update, does the potential gain in performance by splitting warrant its additional computation? If so, in which direction should the split take place to result in the best possible estimator performance? DeMars has done much work in answering these questions but with a philosophy more tailored to the prediction or propagation phase [7]. This is a logical theme in much of the existing literature [8],[4],[5], since orbit problems often involve periods of long propagation with sparse observations. Also, nearly all references [7],[9],[4] take the direction for splitting to be an eigenvector of the prior covariance matrix, typically the one associated with the maximum eigenvalue. Since this is essentially the dominant direction of prior uncertainty, these methods do address the priority of constraining sizes of component covariances. A guiding principle of this work is the additional consideration of the actual nonlinearities at play in the observation model. It is recognized that the quality of reapproximation of the prior is not in and of itself the chief interest, which is rather the quality of the filtered reapproximation with respect to the unknown true Bayesian posterior solution. Any method offered under this principle must explicitly depend on some characterization of the nonlinear measurement model, just as the Bayes' solution does. An algorithm is submitted in [10] that is tailored for use with a Sigma Point Filter with direct use of products from the Unscented Transform, thus speficiying applicability to that case. The current work utilizes an EKF but does not restrict to that case alone. The method in [11] employs statistical linearization, which is very nice for capturing the effect of linear truncation on uncertainty but can also require significant com-

putation. Since the splitting scheme may be applied recursively until the splitting criteria deems no further splits are necessary, a major desire for this work was computational efficiency. The Split and Merge Unscented Gaussian Mixture Filter in [9] executed a nonlinearity check to refine the number of components deemed applicable for splitting but still eventually took the direction for splitting to be the principal eigenaxis of the prior covariance matrix. Full filters were developed in Ref. [9] as well as in Ref. [3]. This work examines only the measurement update portion of the estimation filter with extension to a full filter left to future work. Somewhat conversely to Ref. [9], the procedure in [12] selected components to be split based on a function of uncertainty while the actual splitting was applied in a direction based on nonlinearity. A treatment of the measurement update more similar to the current work is that of [13], though there are significant differences in the derived objective function and solutions provided. A feature in [9] is a means for numerical 5 (and thus circumventing analytical) computation of the second order partial derivatives (Hessian). Finally, the current work is an advanced effort in the line of [14] with key extensions including execution of automated splitting and further development of Kullback-Leibler divergence (KLD) criteria. Simulated examples are provided for demonstration of algorithm performance and illustration of benefits associated with the choice of the chosen scheme.

II. Problem Formulation

In sequential state estimation, we may view any single measurement update as a parameter estimation problem at that instant with all previous filter knowledge provided by the prior state probability density function. The observation y accommodates a nonlinear transformation of the state $\mathbf{x}$.

$$\begin{gathered} y = h(\boldsymbol{x}) + \epsilon \in \mathbb{R} \\ \epsilon \sim N(0, \sigma_R^2) \\ \sigma_R^2 \geq 0 \\ \mathbb{E}\{\epsilon \boldsymbol{x}^T\} = 0_{1\times n} \end{gathered} \tag{2}$$

Note that the current work assumes scalar measurements and an error term resulting from a zero-mean Gaussian distribution with variance σ_R^2, denoted by $N(0, \sigma_R^2)$. The notation N here is

overloaded in that it is used to indicate a Gaussian pdf for a random scalar (or vector) taking its size from the mean argument in N. A priori knowledge of the state is:

$$\begin{gathered} \boldsymbol{x} \sim N(\bar{\boldsymbol{x}}, P) \\ \boldsymbol{x} \in \mathbb{R}^n \\ P \succeq 0 \end{gathered} \tag{3}$$

The expression $P \succeq 0$ indicates that symmetric covariance matrix P is positive-semidefinite. The stated problem may represent treatment of the full filter in the case of a normally distributed prior or treatment of a single Gaussian mixture component (which is initially normally distributed by formulation). The update will be performed by an EKF, the most ubiquitous nonlinear filtering scheme. As previously mentioned, determination of a reapproximation of the prior reduces to prescriptions for how and when splitting into a Gaussian mixture should occur. Before these ideas are developed, however, the concept of a measurement update for GMMs will be discussed.

A. GMM Measurement Update

Consider the task of updating a Gaussian mixture prior, as in Eq. (1), for the system formulated in Eq. (2). It can be shown [15] that the MMSE (minimum mean square error) posterior is given by:

$$p_{\boldsymbol{x}|y}(\boldsymbol{x}) = \sum_{i=1}^{N} w_i^+ N(\boldsymbol{x}; \hat{\boldsymbol{x}}_{\boldsymbol{i}}, P_i^+) \tag{4}$$

where

$$\begin{gathered} \hat{\boldsymbol{x}}_{\boldsymbol{i}} = \bar{\boldsymbol{x}}_{\boldsymbol{i}} + K_i(y - \bar{y}_i) \\ P_i^+ = P_i - K_i P_i K_i^T \\ K_i = P_{\boldsymbol{x}y,i} P_{yy,i}^{-1} \\ w_i^+ = \frac{\beta_i w_i}{\sum_{j=1}^{N} \beta_j w_j} \\ \beta_i = N(y; \bar{y}_i, P_{yy}) \end{gathered} \tag{5}$$

Examining Eq. (4), the updated pdf is simply a linear combination of the updated pdfs for each component, where updates were performed using standard LMMSE equations shown in Eq. (5). Due to nonlinearity in the observation model, terms $\bar{y}_i$, $P_{\boldsymbol{x}y,i}$, and $P_{yy,i}$ cannot in general be

exactly computed using finite computational power and must be approximated in some manner (i.e. Taylor series expansion for an EKF or stochastic linearization for a UKF, Unscented Kalman Filter). Note that coefficients, the updated weights, for combining individual updates in Eq. (4) rely on information from all components via the β_j terms in the summation for w_i^+. This property subtly implies that the overall filtered solution does not result from isolated treatment of mixture components, and it preserves the notion of the weights as (normalized) magnitudes of probability contributions for each component.

Since examples and discussion in the current work utilize an EKF, a short review of the EKF tool will be provided.

B. Summary of Traditional (First-order) EKF

In the first-order EKF framework [16], the observation function is represented by its linearized form, resulting from truncating the Taylor series expansion about the prior mean after first-order terms.

$$\begin{aligned} y &= h(\bar{\boldsymbol{x}}) + \frac{\partial h}{\partial \boldsymbol{x}}\Big|_{\bar{\boldsymbol{x}}}[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \mathcal{O}(\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2^2) + \epsilon \\ &\approx h(\bar{\boldsymbol{x}}) + \frac{\partial h}{\partial \boldsymbol{x}}\Big|_{\bar{\boldsymbol{x}}}[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \epsilon \\ &= h(\bar{\boldsymbol{x}}) + H[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \epsilon \\ &\triangleq y_{EKF} \end{aligned} \tag{6}$$

Terms H and y_{EKF} are defined to be the observation function Jacobian evaluated at the prior mean and the linearized observation model respectively. A Kalman filter may be applied for an optimal solution to the linearized problem.

$$\begin{aligned} \hat{\boldsymbol{x}} &= \bar{\boldsymbol{x}} + K_{EKF}(y - \bar{y}_{EKF}) \\ P^+ &= P - K_{EKF} W_{EKF} K_{EKF}^T \\ K_{EKF} &= P H^T W_{EKF}^{-1} \end{aligned} \tag{7}$$

where

$$\begin{aligned}
\bar{y}_{EKF} &= \mathbb{E}\{y_{EKF}\} = \mathbb{E}\{h(\bar{\boldsymbol{x}}) + H[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \epsilon\} = h(\bar{\boldsymbol{x}}) \\
W_{EKF} &= \mathbb{E}\{[y_{EKF} - \bar{y}_{EKF}][y_{EKF} - \bar{y}_{EKF}]^T]\} \\
&= \mathbb{E}\{H[\boldsymbol{x} - \bar{\boldsymbol{x}}][\boldsymbol{x} - \bar{\boldsymbol{x}}]^T H^T\} + \mathbb{E}\{H[\boldsymbol{x} - \bar{\boldsymbol{x}}]\epsilon^T\} + \mathbb{E}\{\epsilon[\boldsymbol{x} - \bar{\boldsymbol{x}}]^T H^T\} + \mathbb{E}\{\epsilon^2\} \\
&= HPH^T + \sigma_R^2
\end{aligned} \tag{8}$$

The estimated measurement using a priori state information and a 1st-order Taylor series expansion of the observation model is denoted by $\bar{y}_{EKF}$. The associated innovation, innovation covariance, and Kalman gain are denoted by $y - \bar{y}_{EKF}$, W_{EKF}, and K_{EKF} respectively.

C. Summary of Gaussian Second-order EKF

For the Gaussian second-order EKF [17], the overall update is designated to take the same linear form.

$$\begin{aligned}
\hat{\boldsymbol{x}} &= \bar{\boldsymbol{x}} + K_{SEKF}(y - \bar{y}_{SEKF}) \\
P^+ &= P - K_{SEKF} W_{SEKF} K_{SEKF}^T \\
K_{SEKF} &= PH^T W_{SEKF}^{-1}
\end{aligned} \tag{9}$$

However, the observation model is effectively replaced by its second-order Taylor series expansion rather than the linearized version, so estimated moments of the innovation process incorporate some higher-order terms - specifically, those that are quadratic in the state.

$$\begin{aligned}
y_{SEKF} &\triangleq h(\bar{\boldsymbol{x}}) + \frac{\partial h}{\partial \boldsymbol{x}}\Big|_{\bar{\boldsymbol{x}}}[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \frac{1}{2}[\boldsymbol{x} - \bar{\boldsymbol{x}}]^T \frac{\partial^2 h}{\partial \boldsymbol{x}\ \partial \boldsymbol{x}^T}\Big|_{\bar{\boldsymbol{x}}}[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \epsilon \\
&= h(\bar{\boldsymbol{x}}) + H[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \frac{1}{2}[\boldsymbol{x} - \bar{\boldsymbol{x}}]^T D[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \epsilon
\end{aligned} \tag{10}$$

Term D is defined to be the Hessian of the observation function evaluated at the prior mean. It can be shown that in the presence of Gaussian prior distributions, innovation process terms reduce to the following [18]:

$$\begin{aligned}
\bar{y}_{SEKF} &= \mathbb{E}\{y_{SEKF}\} = \mathbb{E}\{h(\bar{\boldsymbol{x}}) + H[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \frac{1}{2}[\boldsymbol{x} - \bar{\boldsymbol{x}}]^T D[\boldsymbol{x} - \bar{\boldsymbol{x}}] + \epsilon\} = h(\bar{\boldsymbol{x}}) + \frac{1}{2}tr(DP) \\
W_{SEKF} &= \mathbb{E}\{[y_{SEKF} - \bar{y}_{SEKF}][y_{SEKF} - \bar{y}_{SEKF}]^T]\} \\
&= HPH^T + \sigma_R^2 + \frac{1}{2}tr(DPDP)
\end{aligned} \tag{11}$$

D. Briefing on Information-theoretic Terms

Information theory provides additional mathematical structure for posing and solving statistical estimation problems. Terms central to the developed algorithm will be defined and motivated here for reference in later sections. For a continuous random variable, $\boldsymbol{X}$, its associated differential entropy (DE) is:

$$\begin{aligned} \mathcal{H}(\boldsymbol{X}) &= -\int_S p_{\boldsymbol{X}}(\boldsymbol{x})\ ln\ p_{\boldsymbol{X}}(\boldsymbol{x})\ \boldsymbol{dx} \\ &= \mathbb{E}\{-ln\ p_{\boldsymbol{X}}(\cdot)\} \end{aligned} \tag{12}$$

where $p_{\boldsymbol{X}}$ is the pdf for $\boldsymbol{X}$ (with subscripts furthermore dropped) and S is the support for $\boldsymbol{X}$. Note that examples of random variables can be conjured such that the integral in Eq. (12) does not converge but are not relevant to the current work. Also, the use of the natural logarithm defines units of information to be nats rather than say bits from a base-2 logarithm. In the case that $\boldsymbol{X}$ is a Gaussian random variable:

$$\begin{aligned} \mathcal{H}(N(\ \cdot\ ; \boldsymbol{m}, P)) &= -\int_S p_{\boldsymbol{X}}(\boldsymbol{x})\ ln\ p_{\boldsymbol{X}}(\boldsymbol{x})\ \boldsymbol{dx} \\ &= \frac{1}{2} ln|2\pi e P| \end{aligned} \tag{13}$$

DE is often described as the average surprisal or unpredictability of a random variable and is associated with uncertainty. This link with uncertainty is made very clear for Gaussians in Eq. (13), where the DE in this case is simply a function of the volume of the covariance ellipsoid.

In comparing information content between random variables, or equivalently between their probability density functions, a popular term is relative entropy / Kullback-Leibler divergence (KLD).

$$\begin{aligned} D_{KL}(p||q) &= \int_S p(\boldsymbol{x})\ ln\ \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\ \boldsymbol{dx} \\ &= \mathbb{E}_p \left\{ ln\ \frac{p}{q} \right\} \end{aligned} \tag{14}$$

For purposes of this work, Eq. (14) represents the amount of information lost when pdf q is used to approximate the "true" probability density function p [19]. It can also be cast as the amount of information gained when revising belief concerning the distribution of the probability of $\boldsymbol{X}$ from q to p. The KLD is often invoked as a distance between pdfs although it is not a true metric as evidenced by its failure to satisfy symmetry with respect to its arguments and the triangle inequality.

It does however satisfy non-negativity (which may be proved via Gibbs' inequality), reaches 0 if and only if the two distributions are identical, and is a measure. Note that for convergence of the integral, p and q must be defined over the same support and furthermore, as a matter of convenience, the value of $0\ ln\frac{0}{0}$ is evaluated to be 0 in Eq. (14).

Finally, a few comments will be made to connect these terms to a Bayesian framework and demonstrate their relevance within a familiar setting. Disregarding for now the known Bayesian solution, consider the problem of updating prior belief concerning a random phenomenon, $\boldsymbol{X}$, using available measurements. Prior belief is represented by pdf p_{prior}, and observational belief is represented by likelihood function p_L. Intuitively, some balance must be made between fitting beliefs defined by these two terms. It was observed in [20] that a solution purely tailored to fitting prior belief may be posed as one that minimizes the information gain from p_{prior} to p using:

$$p_{post} = \arg\min_{p} \{D_{KL}(p||p_{prior})\} = \arg\min_{p} \left\{\int_S p(\boldsymbol{x})\ ln\ \frac{p(\boldsymbol{x})}{p_{prior}(\boldsymbol{x})}\ \boldsymbol{dx}\right\} \tag{15}$$

A solution only seeking to fit belief from measurements (denoted in general here as vector $\boldsymbol{z}$) may be posed as one that maximizes the expected log-likelihood using:

$$p_{post} = \arg\min_{p} \left\{-\int_S p(\boldsymbol{x})\ ln\ p_L(\boldsymbol{z}|\boldsymbol{x})\ \boldsymbol{dx}\right\} \tag{16}$$

A compromise between these goals is the following optimization problem:

$$p_{post} = \arg\min_{p} \left\{\int_S p(\boldsymbol{x})\ ln\ \frac{p(\boldsymbol{x})}{p_{prior}(\boldsymbol{x})}\ \boldsymbol{dx} - \int_S p(\boldsymbol{x})\ ln\ p_L(\boldsymbol{z}|\boldsymbol{x})\ \boldsymbol{dx}\right\} \tag{17}$$

It can be shown [20] that a unique solution exists (when restricting p to satisfy defined properties of a pdf), and that solution is in fact Bayes' Theorem:

$$p_{post}(\boldsymbol{x}|\boldsymbol{z}) = \frac{p_L(\boldsymbol{z}|\boldsymbol{x})\ p_{prior}(\boldsymbol{x})}{\int_S p_L(\boldsymbol{z}|\boldsymbol{\xi})\ p_{prior}(\boldsymbol{\xi})\boldsymbol{d\xi}} \tag{18}$$

Furthermore, the author notes that in form (discounting the fact that the likelihood function is not a true pdf), the optimization problem in Eq. (17) may be rewritten as:

$$p_{post} = \arg\min_{p} \{D_{KL}(p||p_{prior}) + D_{KL}(p||p_L) + \mathcal{H}(p)\} \tag{19}$$

with the nice formal interpretation of Bayes' Theorem as minimizing an equally-weighted expression of information divergences from prior and observational beliefs along with some function of posterior uncertainty.

III. Development of the Splitting Scheme

In developing a philosophy for splitting, the effects both of prior uncertainty and nonlinearity on the EKF solution must be somehow quantified and ideally minimized, or at the very least significantly mitigated.

A. Splitting Direction

The governing idea for capturing nonlinearity in the current work relies on behavior of the observation function Jacobian. In the case of a linear observation model, the Jacobian is constant over the entire state space. In other words, the rate of change of the Jacobian with respect to the state, meaning the Hessian of the observation function, is identically zero when h is linear in $\boldsymbol{x}$. Large changes in the Jacobian then are associated with the significance of higher-order terms in the observation function. The directional derivative evaluated at the prior mean, or the rate of change of the Jacobian at $\bar{\boldsymbol{x}}$ in direction $\hat{\boldsymbol{u}}$, is a reasonable quantification of nonlinearity.

$$\nabla_{\hat{\boldsymbol{u}}} H(\boldsymbol{x})\Big|_{\bar{\boldsymbol{x}}} = \lim_{\alpha \to 0} \frac{H(\bar{\boldsymbol{x}} + \alpha \hat{\boldsymbol{u}}) - H(\bar{\boldsymbol{x}})}{\alpha} \tag{20}$$

The Taylor series for the Jacobian provided in Eq. (21) can be substituted into Eq. (20) to provide Eq. (22).

$$H(\boldsymbol{x}) = H(\bar{\boldsymbol{x}}) + \frac{\partial H}{\partial x}\Big|_{\bar{\boldsymbol{x}}} [\boldsymbol{x} - \bar{\boldsymbol{x}}] + ... \tag{21}$$

$$\begin{aligned} \nabla_{\hat{\boldsymbol{u}}} H(\boldsymbol{x})\Big|_{\bar{\boldsymbol{x}}} &= \lim_{\alpha \to 0} \frac{H(\bar{\boldsymbol{x}}) + \alpha \frac{\partial H}{\partial \boldsymbol{x}}\Big|_{\bar{\boldsymbol{x}}} \hat{\boldsymbol{u}} + \frac{1}{2}\alpha^2 \hat{\boldsymbol{u}}^T \frac{\partial^2 H}{\partial^2 \boldsymbol{x}}\Big|_{\bar{\boldsymbol{x}}} \hat{\boldsymbol{u}} + ... - H(\bar{\boldsymbol{x}})}{\alpha} \\ &= \frac{\partial H}{\partial \boldsymbol{x}}\Big|_{\bar{\boldsymbol{x}}} \hat{\boldsymbol{u}} \\ &= D\hat{\boldsymbol{u}} \end{aligned} \tag{22}$$

In order to quantify the magnitude of prior uncertainty for a specific direction, one may examine the standard deviation of the state pdf when the prior is conditioned on all directions orthogonal to that considered. It can be shown, as in the App., that this is the quantity given in Eq. (23).

$$\sigma_{\hat{u}} = \frac{1}{\sqrt{\hat{\boldsymbol{u}}^T P^{-1} \hat{\boldsymbol{u}}}} \tag{23}$$

With both contributors to degradation of the EKF-solution quantified, the following cost function is chosen for the current work.

$$\begin{aligned} J(\hat{\boldsymbol{u}}) &= \left\| \nabla_{\hat{\boldsymbol{u}}} H(\boldsymbol{x}) \Big|_{\bar{\boldsymbol{x}}} \right\|^2 \sigma_{\hat{\boldsymbol{u}}}^2 \\ &= \frac{\hat{\boldsymbol{u}}^T D^T D \hat{\boldsymbol{u}}}{\hat{\boldsymbol{u}}^T P^{-1} \hat{\boldsymbol{u}}} \end{aligned} \tag{24}$$

As mentioned, a major benefit of this choice is the incorporation of both prior uncertainty and nonlinearity of h in the choice of splitting direction. The form of $J(\hat{\boldsymbol{u}})$ may be thought of a means for selecting directions with relevant prior uncertainty, in that a direction in which the filter is significantly uncertain but along which the observation function is nearly linear will be given less priority than perhaps a direction with less uncertainty but much more nonlinearity. $J(\hat{\boldsymbol{u}})$ similarly may be thought of as capturing relevant nonlinearity.

Note that nonlinearity is evaluated only locally at the prior mean. Another scheme integrates nonlinearity along each direction but does not result in an exact, closed-form solution for the cost function like that which is provided here [11]. Furthermore, only eigendirections of the prior are considered, a common trait in other methods [7]. Another advantage of the chosen cost function is that it is not restricted to a certain estimator by formulation, as opposed to other schemes which for example require terms used in an unscented Kalman filter [10].

The maximizer of the cost function in Eq. (24) is easily attained once it is recognized as a Rayleigh quotient expression. The solution follows easily from a change of variables and the Rayleigh-Ritz inequality.

$$\begin{aligned} P^{-1} &= P^{-T/2} P^{-1/2} \\ \boldsymbol{v} &= P^{-1/2} \hat{\boldsymbol{u}} \end{aligned} \tag{25}$$

$$J(\boldsymbol{v}) = \frac{\boldsymbol{v}^T P^{T/2} D^T D P^{1/2} \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} \tag{26}$$

The Rayleigh-Ritz inequality states:

$$\lambda_{min}(P^{T/2} D^T D P^{1/2}) \boldsymbol{v}^T \boldsymbol{v} \leq \boldsymbol{v}^T P^{T/2} D^T D P^{1/2} \boldsymbol{v} \leq \lambda_{max}(P^{T/2} D^T D P^{1/2}) \boldsymbol{v}^T \boldsymbol{v} \tag{27}$$

where $\lambda_{min}(P^{T/2} D^T D P^{1/2})$ refers to the minimum eigenvalue of matrix $P^{T/2} D^T D P^{1/2}$ and similarly for the λ_{min} term. Substituting Eq. (27) into Eq. (26) gives the following result.

$$J(\boldsymbol{v}) \leq \frac{\lambda_{max}(P^{T/2} D^T D P^{1/2}) \boldsymbol{v}^T \boldsymbol{v}}{\boldsymbol{v}^T \boldsymbol{v}} = \lambda_{max}(P^{T/2} D^T D P^{1/2}) \tag{28}$$

Clearly then, the cost function is maximized when $\boldsymbol{v}$ is taken to be the eigenvector associated with the maximum eigenvalue of matrix $P^{1/2}D^TDP^{1/2}$ (including the property that P is symmetric), so the choice of a splitting direction is equivalent to an eigenvalue/eigenvector problem.

B. Splitting Criteria

A natural choice for a splitting criterion given the previous discussion utilizes the maximum of the cost function associated with the potential splitting direction.

$$\lambda_{max}(P^{T/2}D^TDP^{1/2}) \geq \tau \tag{29}$$

This would result in very few additional calculations in the event that splitting is deemed necessary. The term τ represents a threshold value.

The remaining options rely on comparison of the linearized EKF solution with a higher-order, chosen here to be the second-order EKF, solution. Knowing that incorporation of the quadratic terms in the SEKF appears within the innovation terms, one option for a criterion involves the difference in innovation covariances.

$$tr(DPDP) \geq \gamma HPH^T \tag{30}$$

In this scenario, if the additional term in the SEKF is deemed to be large enough relative to the term from the EKF (via some ratio γ), then splitting is chosen to occur. .

Another option that compares the EKF and SEKF solutions is the Kullback-Leibler divergence between Gaussian posteriors approximated by both filters. τ is used again to represent some threshold.

$$D_{KL}\left(N(\hat{\boldsymbol{x}}_{SEKF}, P^+_{SEKF}) \mid\mid N(\hat{\boldsymbol{x}}_{EKF}, P^+_{EKF}\right) \geq \tau \tag{31}$$

where

$$D_{KL}\left(N(\hat{\boldsymbol{x}}_{SEKF}, P^+_{SEKF}) \mid\mid N(\hat{\boldsymbol{x}}_{EKF}, P^+_{EKF}\right) = \frac{1}{2}(ln\frac{|P^+_{EKF}|}{|P^+_{SEKF}|} - n + [\hat{\boldsymbol{x}}_{EKF} - \hat{\boldsymbol{x}}_{SEKF}]^T P^{+^{-1}}_{EKF}[\hat{\boldsymbol{x}}_{EKF} - \hat{\boldsymbol{x}}_{SEKF}] + tr(P^+_{SEKF}P^{+^{-1}}_{EKF})) \tag{32}$$

A drawback for this method is that it does essentially require carrying the SEKF along with the EKF. However, from Eq. (11), few additional terms are necessary, and this burden is much less than if

the additional estimator were, for example, a sigma-point filter. The main advantage of this method is that it provides a clearer representation of the loss of information associated with a linearized solution. Starting from a Gaussian prior, the EKF framework results in a Gaussian posterior, although the true conditional state pdf will not be Gaussian for general nonlinear measurement functions. Eq. (32) is therefore a fitting quantity to represent the role that nonlinearities in h plays in making the update to a Gaussian component differ from a Gaussian posterior. A KLD that is above the threshold for splitting, signals significant failure in approximating the posterior solution as Gaussian. When this occurs, the prior is reapproximated as Gaussian mixture model, components then of the reapproximated prior will be more local and more suitable for an EKF solution and will contribute to a Gaussian mixture approximation of the posterior.

C. Automated Splitting Using the Kullback-Leibler Divergence

As mentioned, the splitting scheme provided can be implemented recursively as an automated splitting algorithm. The KLD was chosen as the splitting criterion for this purpose. The KLD lends itself to evaluating the ability of a more easily attained pdf in approximating a more intractable pdf and has for this reason been adopted by many machine learning and variational inference applications [21]. Furthermore, the informational interpretation of the KLD better relates this criterion to the purposes of probability and statistical inference rather than other, even normed, functional comparisons such as the distance in L_2. This interpretation also translates well to merging of Gaussian mixture components.

In the automated scheme, the KLD criterion for each component is weighted by the squared EKF-updated weight for that component. The check then for each component, i, is:

$$(w^+_{i,\,EKF})^2 D_{KL}\left(N(\hat{\boldsymbol{x}}_{i,\,SEKF}, P^+_{i,\,SEKF}) \;||\; N(\hat{\boldsymbol{x}}_{i,\,EKF}, P^+_{i,\,EKF}\right) \geq \tau \tag{33}$$

The squared weights ensure effort is not expended towards splitting irrelevant components by penalizing the KLD check for components that contribute little to the overall filtered mixture. Recall that the weight of a component in the EKF-update mixture captures the relative contribution of that component to the overall estimate (mean). Disregarding component weights can result in expending much effort in splitting a component within a region of the state-space that has

very little effect on the final solution. Use of the squared weights emphasizes this significance of component contribution relative to the KLD term and was seen to result in better termination of the splitting routine. Refinement of components then only occurs for portions of the mixture that both carry significant information and are at risk for significant degradation of performance induced by nonlinearities.

An important element in the design of an automated algorithm is the often overlooked designation of a threshold for its criterion. It is desirable for this threshold to be motivated by the logic of the problem, appear intuitive to the user, and avoid significant added computations. Care must be given in this case to resist explicit or implicit treatment of the KLD as a norm or metric since it is neither. An initial thought might be to compare the ratio of the computed criterion in Eq. (32) with an upper bound (or a percentage of an upper bound) for all possible values generated by Gaussian distributions. Examining Eq. (32), however, it can be seen that this value can be arbitrarily large with no upper bound.

Consider now two Gaussian random vectors characterized by the following first two moments:

Normal Distribution 1:	$\boldsymbol{\mu}$,	Σ
Normal Distribution 2:	$\boldsymbol{\mu} + c\Sigma^{\frac{1}{2}}\boldsymbol{v}$,	$\frac{1}{k}\Sigma$

where $\boldsymbol{v}$ is some unit vector direction and scalars $c > 0,\ k > 1$. Distribution 1 represents the Gaussian second-order solution while Distribution 2 represents some departure from it parameterized by scalars c and k. Values c and k may be selected by the user to represent some worst-allowable solution. A computed criterion in Eq. (32) that meets or exceeds the KLD associated with this scenario indicates a need for splitting. This worst-allowable KLD is given in Eq. (34).

$$D_{KL}(p_1||p_2) = \frac{1}{2}\left(ln\frac{|P_2|}{|P_1|} - n + [\boldsymbol{m_2} - \boldsymbol{m_1}]^T P_2^{-1}[\boldsymbol{m_2} - \boldsymbol{m_1}] + tr(P_1 P_2^{-1})\right) \tag{34}$$

$$= \frac{1}{2}(n(k - lnk - 1) + c^2 k) \tag{35}$$

Only single component is considered here, so the squared weight term is 1. A computed criterion greater than the KLD associated with the worst-allowable scenario indicates performance is poor enough to warrant splitting. Note that this takes place between quantities of the same units (nats),

and the check at each step is:

$$D_{KL}\left(N(\hat{\boldsymbol{x}}_{_{SEKF}}, P^{+}_{_{SEKF}}) \;||\; N(\hat{\boldsymbol{x}}_{_{EKF}}, P^{+}_{_{EKF}})\right) \geq \frac{1}{2}(n(k - ln(k) - 1) + c^2 k) \tag{36}$$

A nice property of this threshold is its invariability to the specific solutions at each step so that it can be precomputed and applied each time.

For more discussion on the user-defined values, a selection of k corresponds to a particular decrease in uncertainty from the SEKF solution to the EKF solution where the EKF indicates more confidence in its estimate than what is appropriate. Note that the additional positive term in the innovation covariance for the SEKF ensures that the produced estimation error covariance matrix is larger than or equal to that of the EKF. In the univariate case, a choice of c corresponds to displacement of the mean for Distribution 2 by a number of c-σ's from $\boldsymbol{\mu}$ in Distribution 1. Similarly for the multivariate case, this corresponds to displacement of $\boldsymbol{\mu}$ to the c-σ error ellipsoid in the direction of $\boldsymbol{v}$. This can be demonstrated using the fact (from the App.) that the magnitude of any vector to the c-σ error ellipsoid for Distribution 1 in direction $\boldsymbol{v}$ is $\frac{1}{\sqrt{\boldsymbol{v}^T\Sigma^{-1}\boldsymbol{v}}}$, so displacement of the mean to this ellipsoid in direction $\boldsymbol{v}$ gives $\boldsymbol{\mu} + \frac{c}{\sqrt{\boldsymbol{v}^T\Sigma^{-1}\boldsymbol{v}}}\boldsymbol{v}$ for the mean of Distribution 2. Carrying out calculations in Eq. (34) yields the same result.

IV. Simulated Examples

Three examples using range measurements , corresponding to the magnitude of state $\boldsymbol{x}$ with $h(\boldsymbol{x}) = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$, will be presented in order to exhibit beneficial properties of the automated splitting routine for filtering purposes. All examples utilize the splitting libraries in [7] and are of the form described in the Problem Formulation section for estimation of 2-dimensional position. This simplicity facilitates visualization, not only for plotted figures but also for terms within the algorithm. For example, the direction that maximizes only the nonlinearity term in Eq. (24) is always the direction orthogonal to the prior mean (with only one of such directions possible in 2-D). This property can be shown simply through eigenanalysis of the Hessian.

$$D(\bar{\boldsymbol{x}}) = \frac{1}{\sqrt{\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}}}\mathbb{I}_{2\times 2} - \frac{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T}{(\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}})^{\frac{3}{2}}} \tag{37}$$

$$D(\bar{\boldsymbol{x}})\bar{\boldsymbol{x}} = \frac{\bar{\boldsymbol{x}}}{\sqrt{\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}}} - \frac{\bar{\boldsymbol{x}}}{\sqrt{\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}}} = 0 = 0\bar{\boldsymbol{x}} \tag{38}$$

The Hessian is singular then with an eigenvalue of 0 and associated eigenvector in the direction of the prior mean. Now let $\boldsymbol{v}$ be a direction that is perpendicular to the prior mean, meaning $\bar{\boldsymbol{x}}^T\boldsymbol{v} = 0$.

$$D(\bar{\boldsymbol{x}})\boldsymbol{v} = \frac{\boldsymbol{v}}{\sqrt{\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}}} - 0 = \frac{1}{\sqrt{\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}}}\boldsymbol{v} \tag{39}$$

The maximum eigenvalue then is associated with this direction. Note that since the Hessian is symmetric, this analysis for matrix $D^T(\bar{\boldsymbol{x}})D(\bar{\boldsymbol{x}}) = D^2(\bar{\boldsymbol{x}})$ yields the same direction but with a squared eigenvalue. Now directions associated with dominant nonlinearity and uncertainty may be easily identified for the subsequent examples.

A. Example I.

The first example validates use of the automated splitting routine for a classic GMM example and begins to illustrate benefits for the approach derived. Fig. 1 depicts the true prior distribution, likelihood information for the range measurement, and the Monte Carlo-computed true Bayesian solution. Note that captions will list figures from left to right. The state components are initially correlated as indicated by the slight tilt of initial covariance ellipses with respect to the coordinate axes. The true solution is an example of what is called the "banana shape". The curvature of this shape can never be captured by Gaussian-generated equiprobability contours, which are always ellipses. Note that the EKF solution does not explicitly assume Gaussianity, but that is the usual representation considering only two moments are provided. Additionally from an information-theoretic point of view, the Gaussian distribution is the pdf that maximizes DE when the first two moments of the distribution are specified.

As expected in Fig. 2, the EKF solution fails to capture the basic shape of the true solution. For the automated GMM-generated solution, first note that the algorithm reached the exit condition and provided a solution; that is, splitting in the direction derived here resulted in a sufficient decrease in KLD for the splitting criterion to signal a stopping point for all components. Second, the GMM solution is closer to the desired banana shape, which speaks to the choice of splitting direction. This “closeness” in matching the true posterior pdf is evidenced qualitatively in Fig. 2 and can be confirmed by comparing KLDs between the true and approximated posteriors. Selecting KLD as the measure, the use of an EKF as an approximating solution to the truth resulted in a loss of

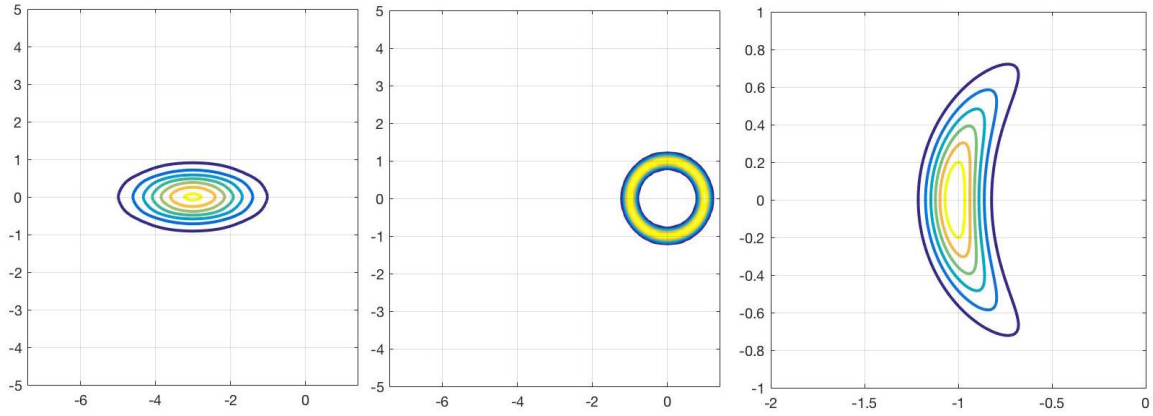

Fig. 1 True Prior, Measurement Distribution, and Posterior for Example I.

information of 1.0461 nats compared to 0.3813 nats with use of the GMM solution. In this example, components were split into sets of 5 at a time using the splitting library within [7], and the final number of components used in this case was 5 (meaning only one split was required).

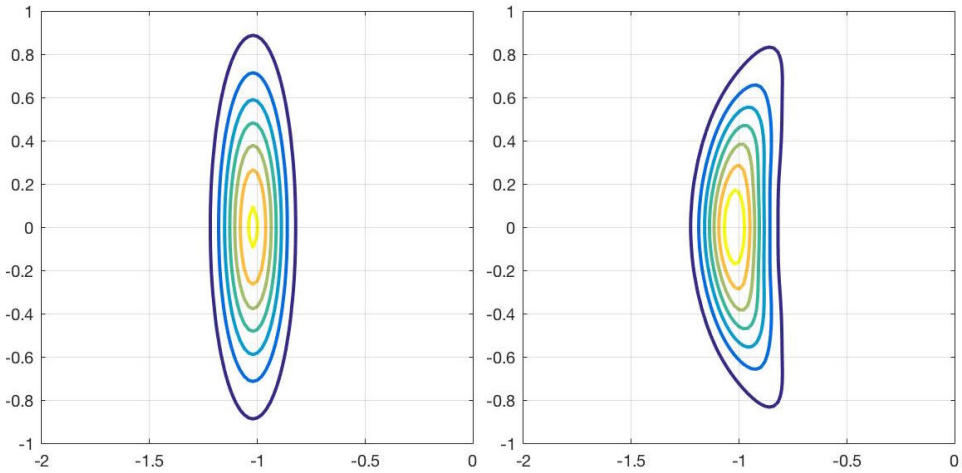

Fig. 2 EKF and GMM Posteriors for Example I.

To evaluate effectiveness of the choice of splitting direction with respect to a common solution, the simulation was replicated but with the direction always taken to be that of dominant prior uncertainty (principal eigenaxis of prior covariance).

The resulting solution in Fig. 3 is also feasible but requires a few additional components (9 rather than 5) for the same threshold. This is due to the fact that the standard method selects the initial splitting direction to be approximately the horizontal axis. However, that direction actually is associated with the least nonlinearity as noted in previous discussion. Maximum nonlinearity

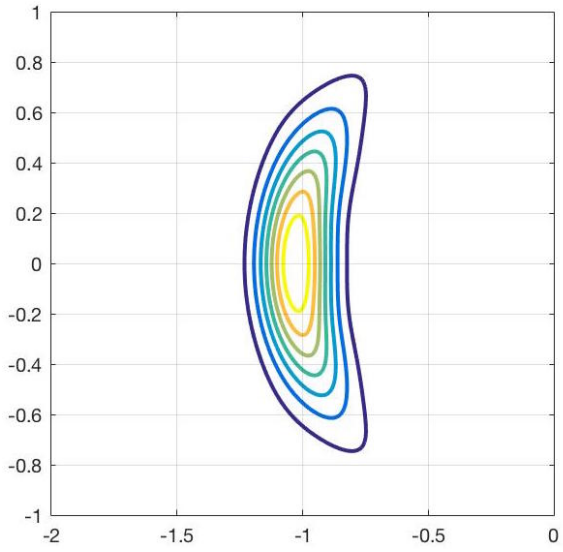

Fig. 3 GMM Posterior for Example I. Using a Standard Method

is associated with the vertical axis. The routine derived here neglects the axis in which there is large uncertainty but no designated nonlinearity in favor of a direction that has low but present uncertainty and maximum nonlinearity. In fact, the developed routine does first select the vertical axis for splitting while the standard method opts for the horizontal axis and requires another split.

B. Example II.

The second example displays a scenario in which the advantage of the provided routine is more pronounced. This setup follows that of the previous example but now with a greater prior uncertainty and more distinctly no dominant direction of prior uncertainty, as depicted in Figure 4.

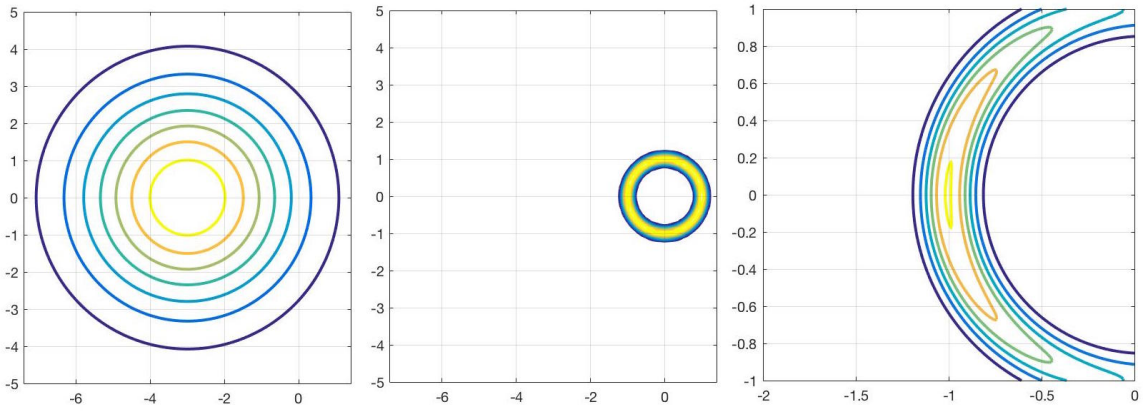

Fig. 4 True Prior, Measurement Distribution, and Posterior for Example II.

The resulting solution for the proposed algorithm is shown in Fig. 5, and Fig. 6 provides the solution using the mentioned standard direction. The proposed algorithm exhibits significantly better performance with fewer components (17 compared to the 33) for the same threshold. With no clear direction of prior uncertainty, the derived routine is sensibly guided by nonlinearity while the standard routine not only fails to adapt with nonlinearity but also begins with an entirely ambiguous idea concerning prior uncertainty.

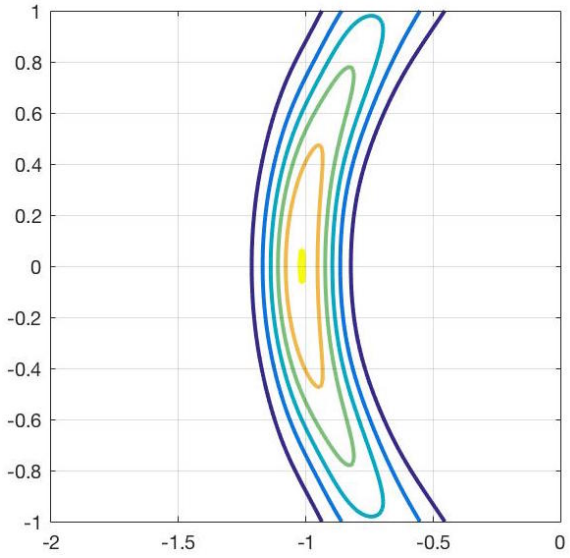

Fig. 5 GMM Posterior for Example II.

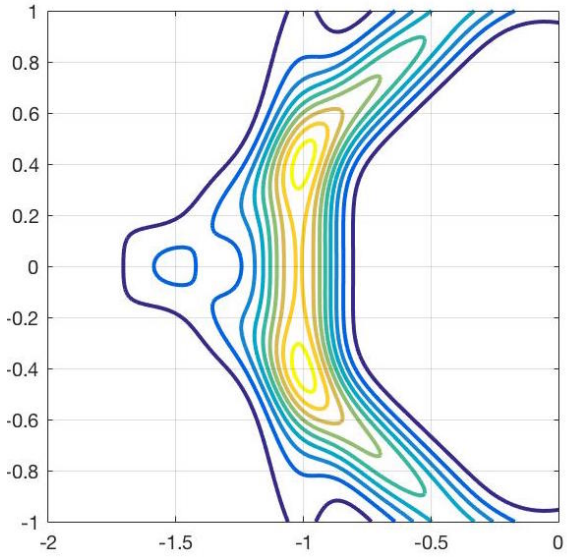

Fig. 6 GMM Posterior for Example II. Using a Standard Method

C. Example III.

The final simulation is included to further motivate the use of Gaussian mixtures in a more unique example and match the result from [14] using the automated version of the scheme. Now a much larger range measurement is used with the prior in Fig. 7 to produce the true solution shown there with virtually 2 blobs or bananas that are significantly separated.

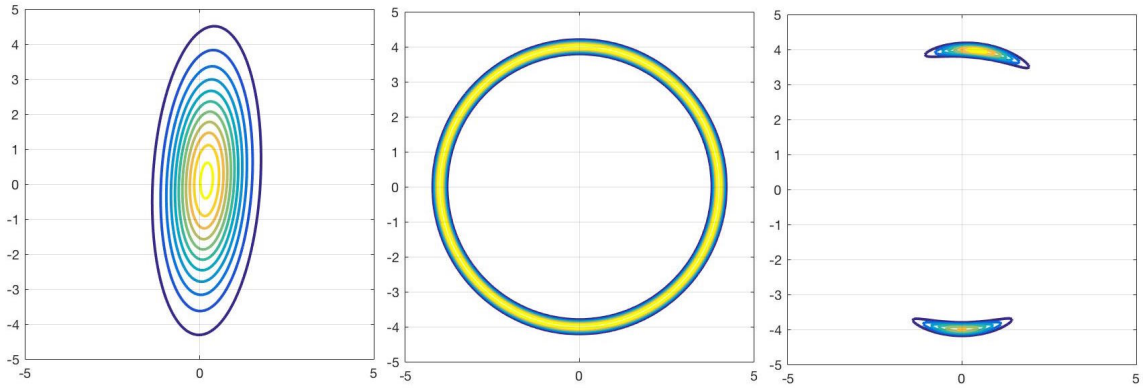

Fig. 7 True Prior, Measurement Distribution, and Posterior for Example III.

The EKF solution shown in Fig. 8 fails to reasonably approximate the truth even beyond displaying inadequate curvature. Although the EKF solution here attempts to respond to the much larger range measurement, it can never exhibit the bimodal nature of the true posterior pdf. Inherent in discussion concerning contours of the pdf is the implication of the estimate that would be provided. For example, it is easy to imagine that the expected value of the EKF solution in Fig. 8 would decidedly differ from that of the true pdf. The automated GMM solution, however, appears to much more reasonably approximate the truth.

The GMM solution using the standard direction is shown in Fig. 9. Again, the performance is visually degraded using slightly more components (25 compared to the 21 used in the solution provided by the proposed scheme).

V. Conclusions

Gaussian mixture models offer a promising path towards a generalized particle filter with mitigated dimensional concerns. Improved online management of the number of components is a crucial step towards that goal. The current work offers an adaptive algorithm for handling measurement

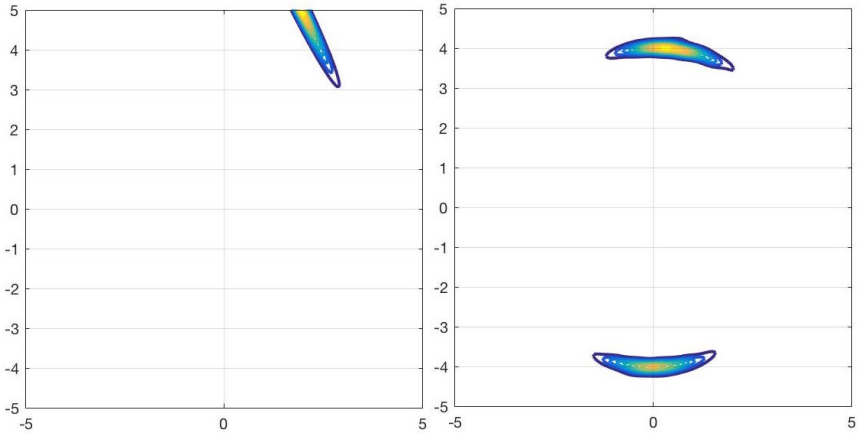

Fig. 8 EKF and GMM Posteriors for Example III.

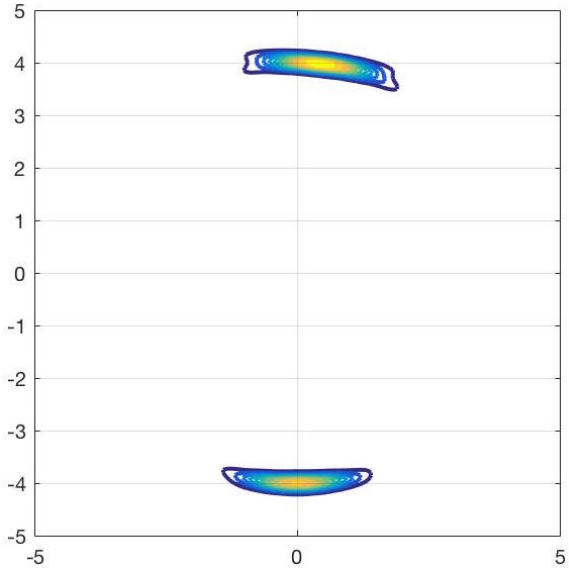

Fig. 9 GMM Posterior for Example III. Using a Standard Method

updates for splitting Gaussian mixture filters with the important consideration of the effects of both nonlinearities and uncertainty in doing so. The adaptable schemes employs the difference between first and second order approximations to detect nonlinearities, it also uses the Kullback-Leibler divergence to detect when the difference between the two approximations is sufficiently pronounced to merit a Gaussian component split. Examples are provided to demonstrate cases of observed effectiveness. The simulated examples show the improved accuracy of the proposed methodology over existing schemes in representing the posterior probability density function with a reduced number of Gaussian components.

Appendix: Directional Variance

The directional standard deviation given in Eq. (23) will be derived. First, suppose direction $\hat{\boldsymbol{u}}$ is an eigendirection of matrix P, that is, it is a principal axis of the covariance ellipsoid defined by the locus of points satisfying the following:

$$\boldsymbol{x}^T P^{-1} \boldsymbol{x} = 1 \tag{40}$$

Notice that the quantity in Eq. (23) corresponds to the length between the origin and the surface of the ellipsoid in the direction of $\hat{\boldsymbol{u}}$. This may be verified by taking a point called $\boldsymbol{x}^*$ that lies on the ellipsoid in direction $\hat{u}$ and solving for its magnitude using Eq. (40).

$$\|\boldsymbol{x}^*\|^2 \hat{\boldsymbol{u}}^T P^{-1} \hat{\boldsymbol{u}} = 1 \implies \|\boldsymbol{x}^*\|^2 = \frac{1}{\hat{\boldsymbol{u}}^T P^{-1} \hat{\boldsymbol{u}}} = \sigma_{\hat{\boldsymbol{u}}}^2 \tag{41}$$

It must be shown that conditional covariance (conditioned on all remaining principal directions) results in Eq. (40). Since $\hat{\boldsymbol{u}}$ is a principal direction, which we can call direction 1 without loss of generality, the conditional covariance is given by the well-known formula below for a Gaussian distribution.

$$P_{1|2} = P_{11}^{-1} - P_{12} P_{22}^{-1} P_{21} \tag{42}$$

where the prior covariance is partitioned as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \tag{43}$$

Using block matrix inversion, the squared-value given in Eq. (23) may be expressed as follows.

$$\begin{aligned} \frac{1}{\hat{\boldsymbol{u}}^T P^{-1} \hat{\boldsymbol{u}}} &= \frac{1}{\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} (P_{11}^{-1} - P_{12} P_{22}^{-1} P_{21})^{-1} & * \\ * & * \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}} \\ &= \frac{1}{(P_{11}^{-1} - P_{12} P_{22}^{-1} P_{21})^{-1}} = P_{11}^{-1} - P_{12} P_{22}^{-1} P_{21} = P_{1|2} \end{aligned} \tag{44}$$

Now, it only remains to be shown that the relationship holds when $\hat{\boldsymbol{u}}$ is not a principal direction. For this, an orthogonal change of basis is performed such that the first axis of the transformed basis

is aligned with $\hat{\boldsymbol{u}}$.

$$\tilde{\boldsymbol{x}} = V\boldsymbol{x} \text{ such that } V\hat{\boldsymbol{u}} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ and } VV^T = I \tag{45}$$

It will be shown that the conditional variance for direction $\hat{\boldsymbol{u}}$, which can be calculated using the traditional formula in the transformed frame, will match that provided by Eq. (23) in the original frame. Moments for the transformed state are:

$$\begin{aligned} \mathbb{E}\{\tilde{\boldsymbol{x}}\} &= V\bar{\boldsymbol{x}} \\ \mathbb{E}\{[\tilde{\boldsymbol{x}} - \mathbb{E}\{\tilde{\boldsymbol{x}}\}][\tilde{\boldsymbol{x}} - \mathbb{E}\{\tilde{\boldsymbol{x}}\}]^T\} &= VPV^T \end{aligned} \tag{46}$$

The result in Eq. (44) holds in the transformed frame as well, giving:

$$\begin{aligned} \tilde{P}_{1|2} &= \frac{1}{\begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \tilde{P}^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}} \\ &= \frac{1}{(V\hat{\boldsymbol{u}})^T (VPV^T)^{-1} V\hat{\boldsymbol{u}}} \\ &= \frac{1}{\hat{\boldsymbol{u}} V^T V P^{-1} V^T V \hat{\boldsymbol{u}}} = \frac{1}{\hat{\boldsymbol{u}}^T P^{-1} \hat{\boldsymbol{u}}} \end{aligned} \tag{47}$$

The result holds then for all directions in the state space.

References

[1] Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T., "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on signal processing*, Vol. 50, No. 2, 2002, pp. 174–188.

[2] Alspach, D. and Sorenson, H., "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE transactions on automatic control*, Vol. 17, No. 4, 1972, pp. 439–448.

[3] Psiaki, M. L., Schoenberg, J. R., and Miller, I. T., "Gaussian Sum Reapproximation for Use in a Nonlinear Filter," *Journal of Guidance, Control, and Dynamics*, Vol. 38, No. 2, 2015, pp. 292–303.

[4] Vittaldev, V. and Russell, R. P., "Collision probability for space objects using Gaussian mixture models," *Proceedings of the 23rd AAS/AIAA Space Flight Mechanics Meeting*, Vol. 148, Advances in the Astronautical Sciences, Univelt San Diego, CA, 2013.

[5] Vittaldev, V. and Russell, R. P., "Space Object Collision Probability Using Multidirectional Gaussian Mixture Models," *Journal of Guidance, Control, and Dynamics*, 2016, pp. 2163–2169.

[6] Plataniotis, K. N. and Hatzinakos, D., "Gaussian mixtures and their applications to signal processing," *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*, 2000.

[7] DeMars, K. J., Bishop, R. H., and Jah, M. K., "Entropy-based approach for uncertainty propagation of nonlinear dynamical systems," *Journal of Guidance, Control, and Dynamics*, Vol. 36, No. 4, 2013, pp. 1047–1057.

[8] Vittaldev, V., Linares, R., and Russell, R. P., "Spacecraft Uncertainty Propagation Using Gaussian Mixture Models and Polynomial Chaos Expansions," *Journal of Guidance, Control, and Dynamics*, Vol. 39, No. 9, 2016, pp. 2163–2169.

[9] Faubel, F., McDonough, J., and Klakow, D., "The split and merge unscented Gaussian mixture filter," *IEEE Signal Processing Letters*, Vol. 16, No. 9, 2009, pp. 786–789.

[10] Havlak, F. and Campbell, M., "Discrete and continuous, probabilistic anticipation for autonomous robots in urban environments," *IEEE Transactions on Robotics*, Vol. 30, No. 2, 2014, pp. 461–474.

[11] Huber, M. F., "Adaptive Gaussian mixture filter based on statistical linearization," *Information fusion (FUSION), 2011 proceedings of the 14th international conference on*, IEEE, 2011, pp. 1–8.

[12] Terejanu, G., "An adaptive split-merge scheme for uncertainty propagation using Gaussian mixture models," *49th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, 2011, p. 890.

[13] Raitoharju, M. and Ali-Loytty, S., "An adaptive derivative free method for Bayesian posterior approximation," *IEEE Signal Processing Letters*, Vol. 19, No. 2, 2012, pp. 87–90.

[14] Tuggle, K., Zanetti, R., and D'Souza, C., "A Splitting Gaussian Mixture Formulation for a Nonlinear Measurement Update," *Proceedings of the AAS/AIAA Space–Flight Mechanics Meeting*, San Antonio, TX, Feb 5–9, 2017 2017, AAS 17-430.

[15] DeMars, K. J., *Nonlinear orbit uncertainty prediction and rectification for space situational awareness*, Ph.D. thesis, 2010.

[16] Gelb, A., *Applied optimal estimation*, chap. 6.1, MIT press, 1974.

[17] Jazwinski, A. H., *Stochastic Processes and Filtering Theory*, Vol. 64, chap. 9.3, Academic Press, 1970.

[18] Zanetti, R., DeMars, K. J., and Bishop, R. H., "Underweighting nonlinear measurements," *Journal of guidance, control, and dynamics*, Vol. 33, No. 5, 2010, pp. 1670–1675.

[19] Burnham, K. P. and Anderson, D. R., *Model selection and multimodel inference: a practical information-theoretic approach*, Springer Science & Business Media, 2003.

[20] Bui-Thanh, T., "A Fresh Look at the Bayes' Theorem from Information Theory," Sept. 2016.

[21] Wainwright, M. J., Jordan, M. I., et al., "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, Vol. 1, No. 1–2, 2008, pp. 1–305.