# An Adaptive Derivative Free Method for Bayesian Posterior Approximation

Matti Raitoharju*, Simo Ali-Löytty

***Abstract*—In the Gaussian mixture approach a Bayesian posterior probability distribution function is approximated using a weighted sum of Gaussians. This work presents a novel method for generating a Gaussian mixture by splitting the prior taking the direction of maximum nonlinearity into account. The proposed method is computationally feasible and does not require analytical differentiation. Tests show that the method approximates the posterior better with fewer Gaussian components than existing methods.**

## I. Introduction

In Bayes' theorem an $n$-dimensional state vector $x$ is estimated by updating its prior distribution using given measurements. The posterior distribution given measurement $y$ is

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{1}$$

where $p(x)$ is the prior probability density function (pdf) of the state, $p(y)$ is a normalizing constant, $p(y|x)$ is the measurement likelihood and $p(x|y)$ is the posterior pdf. In general the update cannot be done analytically.

In this paper the prior is assumed to be a Gaussian and the measurement $y$ to be a scalar that may be written in the form

$$y = h(x) + \varepsilon, \tag{2}$$

where $h(x)$ is the measurement function and $\varepsilon$ is the measurement error, assumed to be zero mean Gaussian independent of the prior.

If the measurement function is linear, i.e. $h(x)$ may be written as $Jx$, the posterior can be computed with the Kalman update [1]

$$\begin{array}{ll} z = h(x) & S = JPJ^T + R \\ C = PJ^T & K = CS^{-1} \\ x^+ = x + K(y - z) & P^+ = P - KSK^T \end{array}, \tag{3}$$

where $x$ and $x^+$ are the prior and posterior means, $P$ and $P^+$ are the prior and posterior covariances and $R$ is the variance of the measurement error. In this paper we assume that $P$ and $R$ are nonsingular.

If the measurement model is not linear the above update cannot be used directly. For nonlinear cases one of the simplest update methods is to compute the Jacobian of the measurement function in the prior mean and use it as $J$ in the Kalman update (3). This is used in the Extended Kalman Filter (EKF) [2, p. 278]. This requires analytical differentiation of $h(x)$, which can be difficult or impossible to perform, also the approximation may be poor if the Jacobian $J$ varies a lot in a small area around the prior mean.

The Unscented Kalman Filter (UKF) is an alternative to EKF that does not require analytical differentiation. The UKF update is based on the evaluation of the measurement function at the so called sigma points. The computation of the sigma points require the computation of a matrix square root of the covariance matrix

$$P = LL^T \tag{4}$$

using, for example, Cholesky decomposition. The extended symmetric sigma point set is

$$\begin{aligned} \chi_0 &= x \\ \chi_i &= x + \Delta_i,\ 1 \leq i \leq n \\ \chi_i &= x - \Delta_{i-n},\ n < i \leq 2n, \end{aligned} \tag{5}$$

where $\Delta_i = \sqrt{n+\xi}L_{:,i}$ ($L_{:,i}$ is the $i^{\text{th}}$ column of $L$) and $\xi$ is an algorithm parameter. The prior is updated by using the following approximations in the Kalman update (3)

$$\begin{aligned} z \approx& \sum \Omega_{i,m} h(\chi_i) \\ S \approx& R + \sum \Omega_{i,c} \left(h(\chi_i) - z\right)\left(h(\chi_i) - z\right)^T \\ C \approx& \sum \Omega_{i,c} \left(\chi_i - x\right)\left(h(\chi_i) - z\right)^T, \end{aligned} \tag{6}$$

where $\Omega_{0,m} = \frac{\xi}{n+\xi}$, $\Omega_{0,c} = \frac{\xi}{n+\xi} + (1 - \alpha_{\text{UKF}}^2 + \beta_{\text{UKF}})$, $\Omega_{i,c} = \Omega_{i,m} = \frac{1}{2n+2\xi}, (i > 0)$ and $\xi = \alpha_{\text{UKF}}^2(n + \kappa_{\text{UKF}}) - n$. The variables with subscript UKF are algorithm parameters [3], [4]. Although the UKF update evaluates the measurement in several points, the posterior distribution is approximated with a single Gaussian. In many cases a single Gaussian is not enough to give a good approximation of the posterior.

Gaussian mixture filters use a weighted sum of Gaussian components to approximate the pdfs [5],

$$p(\bullet) = \sum w_k p_{\text{N}}(\bullet|x_k, P_k), \tag{7}$$

where $w_k$ is the component weight and $p_{\text{N}}(\bullet|x_k, P_k)$ is a pdf of normal distribution with mean $x_k$ and covariance $P_k$. This allows better approximation of the posterior especially when the true posterior is multimodal. The update of any single component may be done using the EKF or UKF formula and the weight of a component is multiplied by the innovation likelihood

$$w_k^+ \propto w_k p_N(y|z_k, S_k), \tag{8}$$



Contact information: Tampere University of Technology. Department of Mathematics, PO Box 553, 33101 Tampere, Finland. e-mal: {matti.raitoharju, simo.ali-loytty}@tut.fi tel. +358456300763. EDICS: SAS-STAT, SAS-SYST. This work was supported by Nokia Inc.

and normalized so that the sum of weights is one.

Although this paper concentrates only on a single update the proposed method is intended to be part of a GMF that does the estimation of a time series. Usually GMFs do not have a constant number of components and a critical issue is to keep the number of components low while still estimating the distributions well.

In algorithms found in the literature the splitting of the prior component in case of nonlinearity into components depends only on the prior distribution. Examples of this kind of GMFs are Sigma Point GMF (SPGMF) [6], Box GMF (BGMF) [7] and Split and Merge Unscented GMF [8]. The first two algorithms require analytical differentiation of the measurement equation and the third uses a simple numerical method for testing if the measurement equation is nonlinear. In this work we propose a new method for prior component splitting that evaluates nonlinearity without the need for analytical differentiation and does the component splitting by taking into account both the prior distribution and the measurement function. In contrast to the methods found in the literature the proposed method does not add components in linear directions.

In the next section a measure of nonlinearity and a formula for its estimation is discussed. Then the splitting of prior according to nonlinearity is presented. In Section IV we show test results of performance of the proposed method compared to existing methods. The paper is concluded in Section V.

## II. Measuring Nonlinearity

A second order Taylor series expansion of a scalar function with a single vector parameter may be written as

$$h(x+\Delta) = h(x) + J\Delta + \frac{1}{2}\Delta^T H \Delta + \varepsilon(\Delta), \qquad (9)$$

where $H$ is the Hessian and $\varepsilon(\Delta)$ is the error caused by the higher order components of the measurement equation. If the quadratic term $\frac{1}{2}\Delta^T H\Delta$ and the higher order term $\varepsilon(\Delta)$ are zero then the Kalman update (3) may be used directly. If the quadratic and higher order terms are small the EKF and UKF approximations should work well.

In GMF it is necessary to evaluate whether the nonlinearity is large or small. In [6], it is proposed that a measurement should be considered highly nonlinear if

$$\operatorname{tr} PHPH > R. \qquad (10)$$

This criterion comes from the comparison of the EKF and the Second Order Gaussian filter [2, pp. 345-349], [9, p. 385]. The term $\operatorname{tr} PHPH$ is called nonlinearity in this paper.

Next we propose a numerical method for computing the term $PH$. In this derivation the higher order term $\varepsilon(\Delta)$ is assumed negligible and the matrix $L$ in (4) is computed using Cholesky decomposition of $P$. We define matrix $Q$ as

$$Q_{i,j} = \begin{cases} h(x+\Delta_i) + h(x-\Delta_i) - 2h(x) & , i = j \\ \frac{1}{2}[h(x+\Delta_i+\Delta_j) + h(x-\Delta_i-\Delta_j) - \\ \quad 2h(x) - Q_{i,i} - Q_{j,j}] & , i \neq j \end{cases} \qquad (11)$$

where $\Delta_i = \gamma L_{:,i}$. If $\gamma$ is chosen as $\gamma = \sqrt{n+\xi}$ then the computed values of the measurement equation in (11) may also be used in the UKF component update (6).

Using (9) with (11) we get

$$\begin{aligned} Q_{i,i} =& h(x) + J\Delta_i + \frac{\Delta_i^T H \Delta_i}{2} + h(x) - J\Delta_i \\ &+ \frac{(-\Delta_i)^T H(-\Delta_i)}{2} - 2h(x) = \Delta_i^T H \Delta_i \end{aligned} \qquad (12)$$

and

$$\begin{aligned} Q_{i,j} =& \frac{1}{2}[(\Delta_i+\Delta_j)^T H(\Delta_i+\Delta_j) - \Delta_i^T H\Delta_i - \Delta_j^T H\Delta_j] \\ =& \Delta_j^T H \Delta_i = \Delta_i^T H \Delta_j, H \text{ is symmetric.} \end{aligned} \qquad (13)$$

Thus matrix $Q$ may be written in matrix form

$$Q = \gamma L^T H \gamma L, \qquad (14)$$

which implies that matrix $PH$ may be computed by

$$PH = \frac{1}{\gamma^2}\gamma L L^T H \gamma L L^{-1} = \frac{1}{\gamma^2} L Q L^{-1}. \qquad (15)$$

The computation of the nonlinearity value (10) does not need the inverse of $L$, because

$$\operatorname{tr} PHPH = \operatorname{tr} \frac{1}{\gamma^2} L Q L^{-1} \frac{1}{\gamma^2} L Q L^{-1} = \frac{\sum_{i,j} Q_{i,j}^2}{\gamma^4}. \qquad (16)$$

## III. Splitting the Prior

In this section we propose a novel method for choosing the components of the Gaussian mixture formed from a Gaussian prior by finding the direction of the maximum nonlinearity. If the measurement is nonlinear according to criterion (10) within a Gaussian component, the component is split into a mixture of two Gaussians that preserves the mean and the covariance of the original component. If the nonlinearity is high in resulting components the split is done recursively for the nonlinear components. The recursive splitting helps to take higher order nonlinearities of the measurement equation into account.

The split vector $a$ is chosen from a set of vectors that have the same probability density

$$p(a) = \frac{1}{\sqrt{2\pi \det P}} e^{-\frac{1}{2}a^T P^{-1} a} = \text{ constant} \qquad (17)$$

that maximizes the absolute value of quadratic term in (9). This may be written as

$$\arg\max_a |a^T H a|, \text{ subject to } a^T P^{-1} a = \beta, \qquad (18)$$

where $\beta$ is a positive algorithm parameter. Using Lagrange multipliers we see that critical points of the optimization problem are vectors that satisfy the constraint and

$$2Ha = 2\lambda P^{-1} a \Leftrightarrow PHa = \lambda a. \qquad (19)$$

Thus, the critical points are eigenvectors of $PH$ that are scaled to satisfy $a^T P^{-1} a = \beta$. Using (19) with (18) we have

$$\arg\max_a |a^T H a| = \arg\max_a |a^T \lambda P^{-1} a| = \arg\max_a |\lambda \beta|, \qquad (20)$$

from which it is seen that an eigenvector $a$ corresponding to the eigenvalue having the largest absolute value is in the direction of maximum nonlinearity.

Equation (14) may be rewritten

$$\frac{1}{\gamma^2}Q = L^T H L. \tag{21}$$

Because matrix $L^T H L$ is real and symmetric its eigenvalues are real and eigenvectors may be chosen orthonormal. Now the eigenvalue decomposition of $L^T H L$ is

$$L^T H L V = V\Lambda, \tag{22}$$

where matrix $V$ is orthonormal having the eigenvectors as its columns; the diagonal elements of the diagonal matrix $\Lambda$ are the corresponding eigenvalues. Multiplying the above equation from the left by $L$ we have

$$LL^T H L V = LV\Lambda \Leftrightarrow (PH)LV = LV\Lambda, \tag{23}$$

from which we see that matrices $PH$ and $L^T H L = \frac{1}{\gamma^2}Q$ have the same eigenvalues and that an eigenvector of $\frac{1}{\gamma^2}Q$ multiplied from the left by $L$ is an eigenvector of $PH$. Now the split vector may be written as

$$a = \sqrt{\beta} L V e_i, \tag{24}$$

where $e_i$ is the $i^{\text{th}}$ column of the identity matrix and $i$ is the index of to the largest eigenvalue in magnitude. The parameters of a two component mixture that preserves the mean and covariance of the prior may be written

$$\begin{array}{ll} \tilde{x}_+ = x + a & \tilde{x}_- = x - a \\ \tilde{P} = P - aa^T & \tilde{w} = \frac{1}{2}w \end{array}, \tag{25}$$

where $w$ is the weight of the original component and the parameters marked with $\tilde{}$ are parameters of the new split components [8]. To ensure that the covariance matrix stays positive definite we have to ensure that $q^T \tilde{P} q > 0$ for any $q \neq 0$, that is,

$$\begin{aligned} q^T \tilde{P} q &= q^T (P - aa^T) q = q^T L L^T q - \beta q^T L V e_i e_i^T V^T L^T q \\ &= ||L^T q||^2 - \beta \cos^2\theta ||L^T q||^2 ||V e_i||^2 \geq ||L^T q||^2 (1 - \beta), \end{aligned} \tag{26}$$

where $\theta$ is the angle between $L^T q$ and $V e_i$. Thus $\beta$ must be chosen from the range $[0, 1[$.

Because of the trace properties it holds that $\operatorname{tr} PHPH = \sum \lambda^2$. The reduction of nonlinearity may be evaluated by looking at the change of the eigenvalues in the resulting component, assuming that the Hessian $H$ does not change. Because

$$\begin{aligned} \tilde{P} H L V &= \left(P - \sqrt{\beta} L V e_i (\sqrt{\beta} L V e_i)^T\right) H L V \\ &= \left(L L^T - \beta L V e_i e_i^T V^T L^T\right) H L V \\ &= \left(L - \beta L V e_i e_i^T V^T\right) L^T H L V \\ &= \left(L - \beta L V e_i e_i^T V^T\right) V \Lambda \\ &= L V \left(\Lambda - \lambda_i \beta e_i e_i^T\right). \end{aligned} \tag{27}$$

it follows that the new matrix $\tilde{P}H$ has the same eigenvectors as the original matrix $PH$ and only the $i^{\text{th}}$ eigenvalue is changed, from $\lambda_i$ to $(1 - \beta)\lambda_i$. Thus the nonlinearity is reduced by $(2\beta - \beta^2)\lambda_i^2$.

In Figure 1 the effect of parameter $\beta$ on the resulting components is presented. The original Gaussian represented by the dashed contour line is horizontally split into two new components. Using $\beta$ close to 1, the nonlinearity decreases fast in splits, but the resulting approximation may be bad. On other hand using a small value of $\beta$ reduces the nonlinearity more slowly. In our tests in Section IV we used $\beta = 0.5$ as a compromise that gave good results in our test scenario.

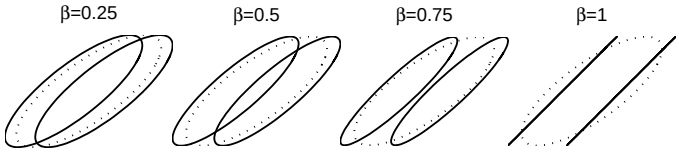


Fig. 1. Effect of $\beta$ on resulting components of splitting

If the measurement is not scalar the splitting could be done for each measurement component separately. Further if the measurements are independent then the update may be done separately for each measurement component.

## IV. Results

Evaluation of the performance is done by comparing the posterior distributions computed by several methods. The proposed method is called adaptive splitting (AS) where the prior is split until none of the mixture components is considered highly nonlinear according to criterion (10). The method is also tested in a variant where at most one split is allowed (AS2). Other methods in comparison are single component UKF, SPGMF [6] with parameters $\tau_{\text{SPGMF}} = 0.5$ and $\kappa_{\text{SPGMF}} = 4$ and BGMF [7], with $N = 1$ and $c_\Sigma = 1$. SPGMF and BGMF use analytic computation of nonlinearity to decide whether the prior shall be split. All tested methods use UKF update (6) with $\alpha_{\text{UKF}} = 10^{-3}$, $\kappa_{\text{UKF}} = 0$ and $\beta_{\text{UKF}} = 2$ [4]. The reference solution is computed using a dense grid where the probability density function is evaluated in each point using Bayes' update formula (1).

Our simulation scenario was a two dimensional positioning case. The measurement function used in simulations was a range measurement from the origin,

$$h(x) = ||x|| + \varepsilon, \tag{28}$$

where $\varepsilon$ is a zero mean Gaussian error term with variance $R$.

In the simulations the range measurement had mean chosen randomly from a uniform distribution in $[0, 10]$ and a unit variance. The prior mean was uniformly distributed with both dimensions in range $[0, 10]$ and covariance matrix had all 10 on diagonal and non diagonal elements were uniformly randomly chosen from the range $[-10, 10]$. The split distance parameter $\beta$ used in simulations was set to $0.5$ i.e. the eigenvalue of $PH$ in the split direction was halved in each split.

An example of a test is presented in Figure 2, where a single 2D range measurement is applied to a Gaussian prior. When comparing visually the true posterior and the different posterior approximations it is seen that the single Gaussian of UKF approximation is not enough in this case and that the proposed method (AS) produces estimates at least as good as the other methods.

In Table I are mean results from 10000 simulation runs. "Time" is the relative time of the method compared to UKF.

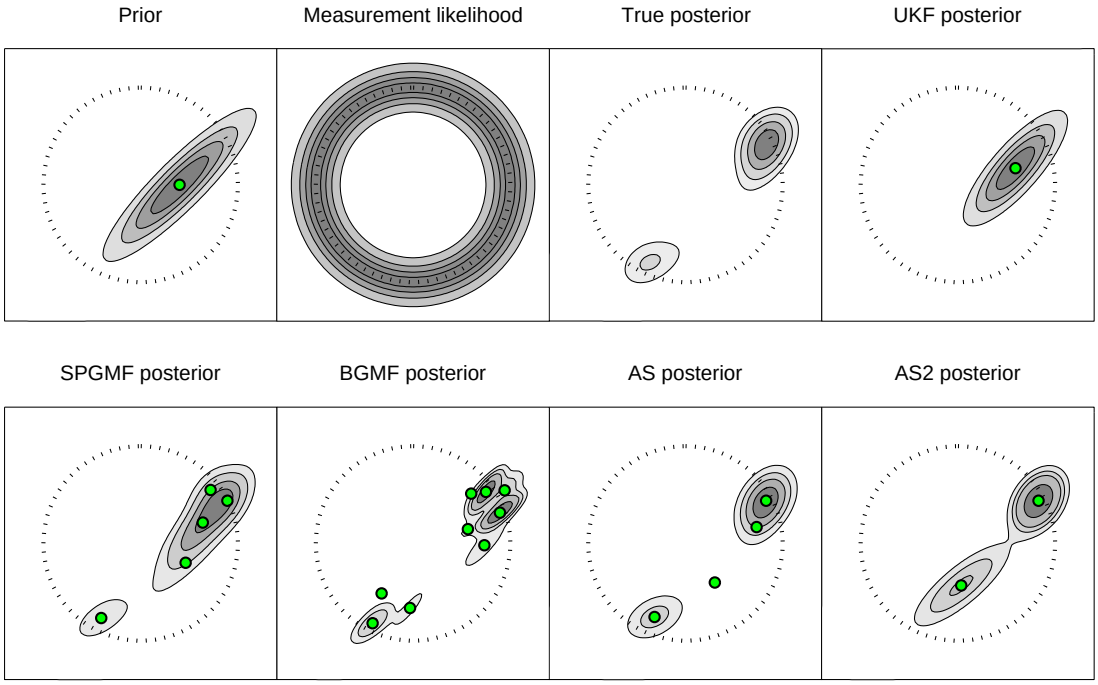


Fig. 2. Exemplary prior and posterior approximations in case of one range measurement. Pdfs are presented with contour maps and the component means are shown as dots.

TABLE I
COMPARISON OF DIFFERENT POSTERIOR APPROXIMATION METHODS

| Method | Time | K-L divergence | Components |
|---|---|---|---|
| UKF | 1.0 | 0.74 | 1 |
| SPGMF | 2.9 | 0.50 | 3.8 |
| BGMF | 12.5 | 0.47 | 6.6 |
| AS | 4.9 | 0.39 | 2.6 |
| AS2 | 2.4 | 0.47 | 1.7 |

"K-L divergence" (Kullback-Leibler divergence [10]) is defined as

$$D_{\mathrm{KL}}(p||q) = \int p(x) \log \frac{p(x)}{q(x)}, \tag{29}$$

where $p(x)$ is the reference pdf and $q(x)$ is the pdf of the Gaussian mixture approximation. "Components" is the number of components in the posterior approximations.

Results show that the proposed method produces a posterior that is clearly closer to the true posterior than UKF, SPGMF or BGMF, and that performs at least as well as the other methods even when the maximum number of components is limited to two. This is a clear indication that the measurement nonlinearity should be taken into account in splitting. The test for nonlinearity (10) gave same result in all 10000 cases for the analytical and numerical methods.

If the state would include more dimensions, for example, the 2D velocity, the number of components of SPGMF and BGMF would have increased from 5 to 9 and from 9 to 81 respectively, whereas AS would not have any more components. Although the SPGMF and BGMF could be programmed in a such way that they do not do splitting in linear dimensions, this would require manual work to customize the algorithms.

## V. CONCLUSION

In this paper it was shown that the nonlinearity of a measurement may be estimated numerically and that if the prior is split in the direction of the maximum nonlinearity the posterior approximation may be done accurately with a relatively small number of components. The proposed method produces better results with a smaller number of components than existing methods and may be used when the measurement equation is hard or even impossible to differentiate.

## REFERENCES

[1] Y. Ho and R. Lee, "A Bayesian approach to problems in stochastic estimation and control," *Automatic Control, IEEE Transactions on*, vol. 9, no. 4, pp. 333 – 339, oct 1964.

[2] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, ser. Mathematics in Science and Engineering. Academic Press, 1970, vol. 64.

[3] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *American Control Conference*, vol. 3, 1995, pp. 1628–1632.

[4] E. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, 2000, pp. 153 –158.

[5] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0005109871900975

[6] S. Ali-Löytty and N. Sirola, "Gaussian mixture filter and hybrid positioning," in *Proceedings of ION GNSS 2007, Fort Worth, Texas*, Fort Worth, September 2007, pp. 562–570. [Online]. Available: http://math.tut.fi/posgroup/ali-loytty_sirola_ion2007a.pdf

[7] S. Ali-Löytty, "Box Gaussian mixture filter," *IEEE Transactions on Automatic Control*, vol. 55, no. 9, pp. 2165–2169, September 2010.

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

[9] Y. Bar-Shalom, T. Kirubarajan, and X.-R. Li, *Estimation with Applications to Tracking and Navigation*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[10] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.