A Gaussian mixture Monte Carlo filter for state estimation in very high dimensional nonlinear systems

Javad Rezaie, Jo Eidsvik

Abstract State and parameter estimation in high dimensional systems is one of the most problematic parts in both monitoring and control. In this paper we propose a robustified Gaussian mixture filter (RGMF). The proposed filter is a generalization of the ensemble Kalman filter, and is more flexible for skewed or multimodal distributions. We modify the proposed filter by Principal Component analysis to remove collinearity in the data and improve filter performance. By doing this we break up the Gaussian mixtures by separating their means according to the posterior distribution. Simulation results show that this modified version of the previously proposed RGMF increases filter performance while it uses limited number of samples as previous. Generally speaking the proposed filter works as fast as previous one but its estimation accuracy and performance is better.

1 Introduction

State estimation is an important problem in engineering and science. If we represent the system dynamics (differential or difference equations) in state space form, the measurements are transformed, noisy and an incomplete representation of the system state. Filtering methods extract the probability distribution of the state at every time point, given all measurements until that time. For dynamic systems it is natural to perform the estimation process as soon as new observations arrive. Thus, recursive Bayesian estimation algorithms are powerful for dealing with filtering problems. This consists of sequentially going forward in time according to a two-step routine: i) a forward propagation step using the system dynamics, and ii) an updating step when the new data gets available. Step i) is known as the *prediction problem*,

Javad Rezaie
Department of Mathematical Sciences, NTNU, Norway, e-mail: rezaie@math.ntnu.no

Jo Eidsvik
Department of Mathematical Sciences, NTNU, Norway, e-mail: joeid@math.ntnu.no

Ninth International Geostatistics Congress, Oslo, Norway, June 11. – 15., 2012

while step ii) is the *filtering problem*. Denote the state variable at time t by x_t, and let $X_t = (x_1,\ldots,x_t)$ be the collection of the state variables from time 1 to the current time t. Further, the observations at time t are denoted y_t, and $Y_t = (y_1,\ldots,y_t)$ is the collection of observations at this current time step. We assume continuous state and observation variables, i.e. $x_t \in \mathscr{R}^n$ and $y_t \in \mathscr{R}^m$, where the dimensions n and m tend to get large in most modern applications. The filtering task consists of sequential propagation and updating as we obtain new observations. At time t-1, consider that we have the updated (filtering) distribution of the state given all observations until that time, denoted by the density $\pi(x_{t-1}|y_1,...,y_{t-1}) = \pi(x_{t-1}|Y_{t-1})$. When the new observation y_t is available, we combine the system dynamics and the likelihood in Bayes rule for the updating:

$$\begin{aligned}\pi(x_t|Y_t) &= \frac{\pi(y_t|x_t)\pi(x_t|Y_{t-1})}{\pi(y_t|Y_{t-1})}\\ \pi(x_t|Y_t) &\propto \pi(y_t|x_t)\pi(x_t|Y_{t-1}) \propto \pi(y_t|x_t)\int \pi(x_t,x_{t-1}|Y_{t-1})dx_{t-1}\\ &\propto \pi(y_t|x_t)\int \pi(x_t|x_{t-1},Y_{t-1})\pi(x_{t-1}|Y_{t-1})dx_{t-1}\\ &\propto \pi(y_t|x_t)\int \pi(x_t|x_{t-1})\pi(x_{t-1}|Y_{t-1})dx_{t-1} \qquad (1)\end{aligned}$$

where the conditional independence assumption of the data y_t is used. This recursive Bayesian method gives the exact solution to the general filtering problem, but for practical applications we cannot implement it for large systems because we must calculate multi dimensional complicated integrals. Thus, some simplified conditions on the system dynamics and observations have to be considered, inducing some consistent approximations.

In this paper we present and extend an algorithm suggested in Rezaie and Eidsvik (2012), which approximates the prediction and filtering distributions by a shrinked Gaussian mixture. This is a flexible approach to the state estimation problem, going between the Ensemble Kalman filter and the Particle filter. The details of the algorithm are outlined in Rezaie and Eidsvik (2012). In this expanded abstract we extend the approach by using dimension reduction techniques for removing the collinearity in samples.

2 Robustified Gaussian mixture Monte Carlo filter

Assume that the system dynamics is generally nonlinear with additive Gaussian noise, $x_t = g(x_{t-1}) + n_t$, where $g(.)$ is a general nonlinear function and n_t is zero mean Gaussian process noise with covariance P, $n_t \sim N(n_t;0,P)$, thus we have:

$$\pi(x_t|x_{t-1}) = N(x_t;g_t(x_{t-1}),P), \qquad (2)$$

Also assume a linear likelihood model, with additive Gaussian noise:

$$\pi(y_t|x_t) = N(y_t; H_t x_t, R). \tag{3}$$

The matrix H_t is defined by the data acquisition of the problem, while R is the covariance matrix of the measurement noise.

By plugging eqn. (2) and (3) in eqn. (1), we have a general posterior distribution. The filtering goal is to sequentially approximate and/or generate samples from this posterior distribution as new observation arrives. Different algorithms have been proposed to approximate this posterior distribution with acceptable errors.

2.1 Ensemble Kalman filter

Ensemble Kalman filter (EnKF) approximates the predictive distribution with a Gaussian one, $\hat{\pi}(x_t|Y_{t-1}) = N(x_t; \overline{x}_t, \overline{P}_t)$, where $\overline{x}_t$ and $\overline{P}_t$ are the empirically estimated mean and covariance from forward propagated or predicted samples. These predicted samples are achieved by propagating the samples from previous filter distribution $x_{t-1}^1, x_{t-1}^2, \ldots, x_{t-1}^B \sim \pi(x_{t-1}|Y_{t-1})$, through forward model. Finally, when the predictive distribution is Gaussian, and also we have Gauss linear likelihood, then the posterior distribution is Gaussian (for more details see Evensen 2009):

$$\begin{aligned}
\hat{\pi}(x_t|Y_t) &\propto N(y_t; H_t x_t, R) N(x_t; \overline{x}_t, \overline{P}_t) \\
\hat{\pi}(x_t|Y_t) &= N(x_t; \overline{x}_t + \overline{P}_t H_t' \overline{Q}_t^{-1} (y_t - H_t \overline{x}_t), \overline{S}_t) \\
\overline{S}_t &= \overline{P}_t - \overline{P}_t H_t' \overline{Q}_t^{-1} H_t \overline{P}_t, \overline{Q}_t = H_t \overline{P}_t H_t' + R.
\end{aligned} \tag{4}$$

2.2 Gaussian mixture filter

Gaussian mixture filter (GMF) approximates the predictive distribution, $\pi(x_t|Y_{t-1})$ with a weighted mixture of Gaussian distributions with known parameters and by assuming Gauss linear likelihood, the posterior distribution is a new mixture of Gaussian (let v_b denote the weights for the different components in the Gaussian mixtures):

$$\begin{aligned}
\pi(x_t|Y_t) &\propto N(y_t; H_t x_t, R) \sum_{b=1}^{B} v_b N(x_t; g(b x_{t-1}^b, P) \\
\pi(x_t|Y_t) &= \sum_{b=1}^{B} w_b N(x_t; \hat{x}_t^b, S_t),
\end{aligned} \tag{5}$$

where $\hat{x}_t^b$ and S_t are the updated mean and covariance matrix, given component b. The components of the Gaussian mixture are obtained by the standard Kalman filter update:

$$\begin{aligned}\hat{x}_t^b &= g_t(x_{t-1}^b) + PH_t^{'}Q_t^{-1}(y_t - H_t g_t(x_{t-1}^b)),\\ S_t &= P - PH_t^{'}Q_t^{-1}H_tP, \quad Q_t = H_tPH_t^{'} + R.\end{aligned} \tag{6}$$

Where the weights w_b is proportional to the likelihood evaluated at samples, $w_b = \frac{N(y_t;H_t g(x_{t-1}^b),Q_t)}{\sum_{c=1}^B N(y_t;H_t g(x_{t-1}^c),Q_t)}$ (for more details see Rezaie and Eisvik, 2012).

2.3 Robustified Gaussian mixture filter

Another important filtering problem in high dimensional systems is sample degeneracy which means that all samples collapse to a few one and they cannot capture the statistical properties of the distributions. EnKF was proposed to deal with this problem in a consistent manner, but it fails if the predictive distribution is far a way from Gaussian. On the other hand, GMF can approximate general predictive and posterior distributions by selecting sufficient number of Gaussian mixtures. Unfortunately, GMF suffers from sample degeneracy which means one weight is close to 1 and the rests are almost zero. Rezaie and Eidsvik (2012) proposed an algorithm for handling both problems. The proposed method is a robustified version of GMF (RGMF) which combines EnKF and GMF, and it can approximate general posterior distributions without suffering from sample degeneracy in high dimensional systems. In RGMF, they define a new predicted sample set, $\{z_t^b, b = 1,\dots,B\}$ by linear combination of the predicted sample mean and predicted samples, $z_t^b = \alpha g(x_{t-1}^b) + (1-\alpha)\overline{x_t}$. By changing parameter α, the shrinked predicted sample, z_t^b moves on a line which connects $\overline{x_t}$ and $g(x_{t-1}^b)$, (Figure 1. explains the effect of this tuning parameter).

The tuning parameter, α , helps us to move between EnKF and GMF by shrinking the predicted samples toward the mean. By proper selection of α we can move predicted samples to high likelihood regions. Based on these newly defined predicted samples, the posterior distribution is a mixture of Gaussian with known parameters and weights (for more details see Rezaie and Eidsvik, 2012):

$$\begin{aligned}\tilde{\pi}(x_t|Y_t) &\propto N(y_t;H_t x_t,R)\tilde{\pi}(x_t|Y_{t-1}),\\ \tilde{\pi}(x_t|Y_t) &= \sum_{b=1}^{B} \tilde{w}_b N(x_t;\tilde{x}_t^b,\tilde{S}_t),\end{aligned} \tag{7}$$

where $\tilde{x}_t^b$ and $\tilde{S}_t$ are the updated mean and variance, given particle b, i.e.

$$\begin{aligned}\tilde{x}_t^b &= z_t^b + \tilde{P}_t H_t^{'}\tilde{Q}_t^{-1}(y_t - H_t z_t^b)\\ \tilde{S}_t &= \tilde{P}_t - \tilde{P}_t H_t^{'}\tilde{Q}_t^{-1}H_t\tilde{P}_t, \ \tilde{Q}_t = H_t\tilde{P}_tH_t^{'} + R.\end{aligned} \tag{8}$$

Where $\tilde{P}_t$ is prediction variance for each mixture component. Naturally, all matrices in this expression depend on the shrinkage parameter α. The weights are now given by

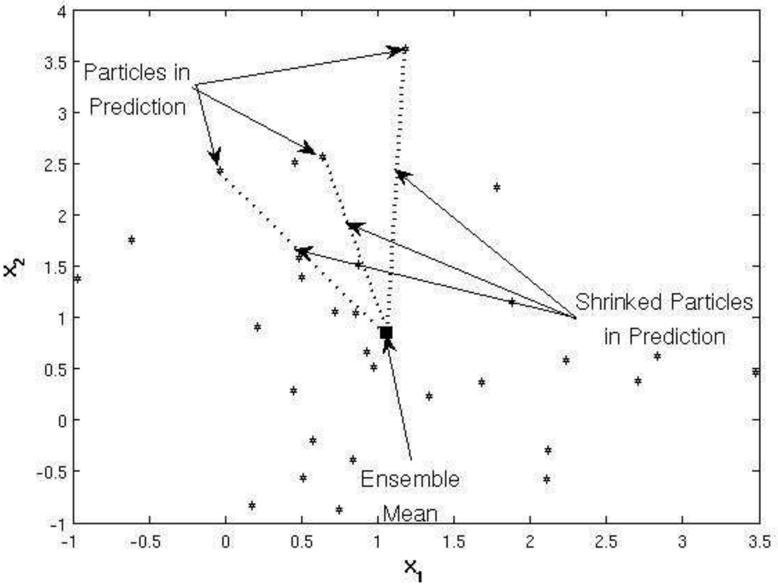

Fig. 1 A graphical description of shrinkage $z_t^b = \alpha g_t(x_{t-1}^b) + (1-\alpha)\bar{x}_t$, the shrunk samples move on the line (dot-line) which connects the ensemble mean (square) to the ensemble members (dot points).

$$\tilde{w}_b = \frac{N(y_t; H_t z_t^b, \tilde{Q}_t)}{\sum_{c=1}^{B} N(y_t; H_t z_t^c, \tilde{Q}_t)}. \tag{9}$$

Figure 2 illustrates the predictive densities of the GMF, EnKF and RGMF for a particular $0 < \alpha < 1$. The GMF gives a very wiggly predictive density plot, while the EnKF is a Gaussian density. Now, if data matches one of the spikes, the particle associated with this spike would get a very large weight w_b in the GMF. This could cause degeneracy. The RGMF is smoother, and closer to the Gaussian curve representing the EnKF. If data matches one of the spikes in the GMF representation, the associated increase in the weight for the RGMF, denoted $\tilde{w}_b$, would not get that much larger than the remaining weights. In order to escape from sample degeneracy, Rezaie and Eidsvik (2012) proposed an algorithm for selecting α in an adaptive manner.

2.4 Principal component analysis in conjunction with RGMF

One problem in the updating part of EnKF based filters is in using collinear/correlated data which causes model overfitting, and the estimated posterior covariance is under estimated as a result. Rezaie et al., (2012) used different statistical dimension

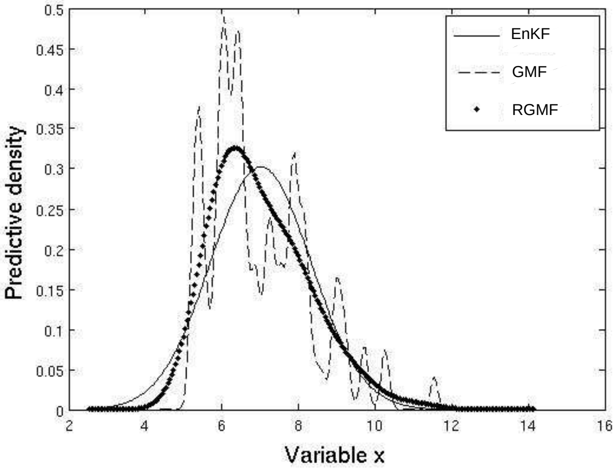

Fig. 2 The predictive distribution from EnKF (solid-line), GMF (dash-line) and RGMF (dot-line).

reduction techniques for finding and removing these collinear data. Based on their work, principal component analysis (PCA) seems promising. PCA is one of the most frequently used dimension reduction techniques. Its implementation is straight forward by using singular value decomposition (SVD). PCA focuses on finding the structure of data ensemble matrix. By finding the structure of data, we mean that PCA finds the directions which data has maximum variability (for more details see Hastie et al., 2009). Clearly speaking, if the data is in data space with its own structure according to basis of this space, the first principal component is defined as a vector which represents the first maximum variability direction of data, the second principal component is the direction of the second maximum variability of data etc. Besides, these PCs are chosen such that they are orthonormal. By transforming these data from data space to PC space (with PCs as the basis for this space) the structure of the data and its actual dimension are found and we can remove the less significant part of the data by removing the last PCs (Figure 3. represent the concept of PC direction and data variability).

A popular criterion for selecting the subspace dimension, is therefore to choose the number of components p so that the explained variance is larger than some tolerance level, β (e.g. 98%). According to Sætrom and Omre (2010) we can look at the update part of Kalman based filters as a linear regression of predicted state on observations and the regression coefficient matrix is the Kalman gain. Thus, the final step is regressing the shrinked samples on the selected PCs in order to find the Kalman gain.

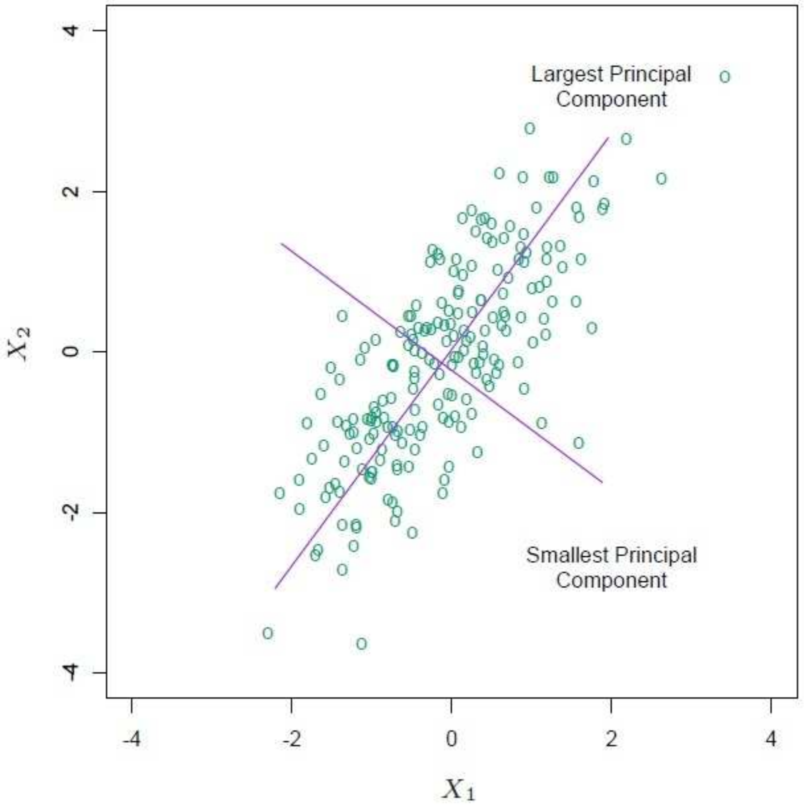

Fig. 3 The principal components for a two dimensional example, Hastie et al. 2009

3 Simulation results

In the example we compare the different algorithms in terms of mean square error (MSE), continuously ranked probability score (CRPS) and variance of the weights. Here, at any time t we have $MSE(t) = \sum_j (\hat{x}_{j,t} - x^{true}_{j,t})^2$, where $\hat{x}_{j,t}$ is the estimated mean of the filtering distribution and the sum is over all n state dimensions. An integrated MSE is achieved by summing out t. Further, the CRPS is defined by $CRPS(t) = \sum_j (\hat{F}(y_{j,t}) - I(y_{j,t} < y^{obs}_{j,t}))^2$. Here, $\hat{F}(.)$ is the empirical cumulative predictive distribution for data at time t, given all former data Y_{t-1}. Smaller values of CRPS means better predictive power. It shows that we often match the observed value, and that we have a narrow prediction band. The sum is over all m observation dimension, and an integrated CRPS is obtained by summing over all times t.

Rezaie and Eidsvik (2012) apply the robustified filter to seismic data assimilation, and Rezaie et al (2012) use the PCA dimension reduction for similar purposes. Here, we discuss a synthetic example for target tracking, which has many facets similar to petroleum reservoir monitoring.

3.1 Tracking targets with bimodal distributions

This example describes the position and velocity of planes or ships moving in two dimensions. If we imagine a monitoring system for planes or ships, their positions are measured by radar /sonar. The targets move in a dependent pattern, i.e. if one turn, others are likely to turn as well.

In this simulation we consider 100 sensor 100 target (system dimension is $n = 400$ and observation dimension is $m = 200$), also the number of ensembels is B=50. We let $x_t = [x_t\ \dot{x}_t\ y_t\ \dot{y}_t]^{'}$ be the state vector of one target. For one target, $(x_t\ y_t)$ is the (north,east) position, and similarly $(\dot{x}_t\ \dot{y}_t)$ is the (north,east) velocity. The absolute velocity is $v_t = \sqrt{\dot{x}_t^2 + \dot{y}_t^2}$, while the target is moving at bearing $\eta_t = \arctan(\frac{y_t}{x_t})$.

With constant velocities, a target moves in a straight line, and the dynamical model is linear. We consider a situation where a target manoeuvres (30 degrees) to the west whenever the velocity v_t becomes smaller than a threshold c. This model is nonlinear, and the dynamics can be phrased by $\pi(x_t|x_{t-1}) \sim N(x_t; g_t(x_{t-1}), P)$. Using a time-step dT , the one-target dynamics for large velocity is:

$$g_t(x_{t-1}) = \begin{bmatrix} 1 & dT & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & dT \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \dot{x}_{t-1} \\ y_{t-1} \\ \dot{y}_{t-1} \end{bmatrix} \tag{10}$$

while for small velocity:

$$g_t(x_{t-1}) = \begin{bmatrix} x_{t-1} + dTcos(\eta_t)v_{t-1} \\ cos(\eta_t)v_{t-1} \\ y_{t-1} + dTsin(\eta_t)v_{t-1} \\ sin(\eta_t)v_{t-1} \end{bmatrix} \tag{11}$$

$$\eta_t = \frac{\pi}{6} + \eta_{t-1}, \qquad \text{if} \quad v_{t-1} < c. \tag{12}$$

Thus, bearing η_t of one target at time t changes westward when the absolute velocity is small. This has effect on the north and east velocity, whereas the absolute velocity $v_t = v_{t-1}$ remains the same, on expectation. As a consequence, the predictions of the north and east positions will tend to be skewed or multimodal, when the distribution for velocity is near the critical velocity c.

The process noise covariance matrix is $P = \text{diag}([0.5^2, \quad 2^2, \quad 0.5^2, \quad 2^2])$ and initial conditions are drawn from $N(x_0; \mu_0, P_0)$ where $\mu_0 = [1000, \quad 75, \quad 1000, \quad 75]^{'}$ and $P_0 = 100P$. We introduce a fixed correlation of 0.9 between all targets, and the joint covariance is block diagonal in the multi-target situation.

We observe the north and east position at every time point, with Gaussian additive noise. Thus, the likelihood model for position data is linear and can be phrased by $\pi(y_t|x_t) \sim N(y_t; H_t x_t, R)$ where $R = \text{diag}([5^2 \quad 5^2])$ and:

$$H_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Simulation results show that using the PCA in the RGMF algorithm generally increase the filter performance. According to Figure 4., we see that the MSE of these filters are more or less in a same range (left plot in Figure 4.), but the CRPS value for RGMF-PCA filter is much lower than the others (upper right plot in Figure 4.). This means the performance of RGMF-PCA increased in CRPS sense without sacrificing the MSE. These two parameters show that the performance of RGMF-PCA for estimating the posterior distribution is better than the others. Besides, the variance of the weights, and equivalently the effective sample size, of the proposed filter (lower right plot in Figure 4.) is lower than the others, which shows the modified filter can better handle sample degeneracy.

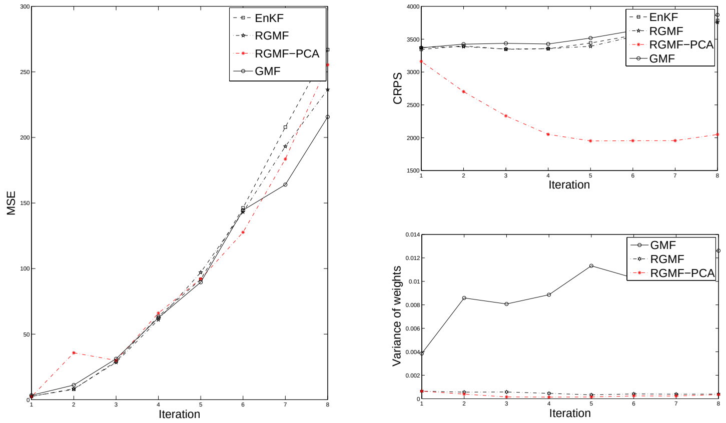

Fig. 4 Comparison of the filters in MSE (left), CRPS (upper right), and variance of the weights (lower right) senses.

4 References

Evensen G. (2009) Data assimilation, The Ensemble Kalman Filter, 2nd ed., Springer.

Hastie T., Tibshirani R., and Freidman J. (2009) The Elements of Statistical Learning; Data Mining, Inference, and Prediction. New York: Springer.

Rezaie J., Sætrom J., and Smørgrav E. (2012) Reducing the Dimensionality of Geophysical Data in Conjunction with Seismic History Matching, Copenhagen, SPE 153924.

Rezaie J., and Eidsvik J. (2012) Shrinked $(1-\alpha)$ ensemble Kalman filter and α Gaussian mixture filter, Computational Geosciences, doi:10.1007/s10596-012-9291-5.

Sætrom J., and Omre H. (2010) Ensemble Kalman filtering with shrinkage regression techniques, Journal of Computational Geosciences, Volume: 15 Issue:2, 271-292.