Ensemble-localized Kernel Density Estimation with Applications to the Ensemble Gaussian Mixture Filter

Andrey A. Popov[1,2*], Enrico M. Zucchelli[2] and Renato Zanetti[1,2]

[1*]Oden Institute for Computational Engineering & Sciences, The University of Texas at Austin, 201 E. 24th Street, POB 4.102, Austin, 78712, Texas, USA.
[2]Dept of Aerospace Engineering & Engineering Mechanics, The University of Texas at Austin, 2617 Wichita Street, C0600, Austin, 78712, Texas, USA.

*Corresponding author(s). E-mail(s): andrey.a.popov@utexas.edu;
Contributing authors: enricomarco@utexas.edu; renato@utexas.edu;

Abstract

The ensemble Gaussian mixture filter (EnGMF) is a non-linear filter suited to data assimilation of highly non-Gaussian and non-linear models that has practical utility in the case of a small number of samples, and theoretical convergence to full Bayesian inference in the ensemble limit. We aim to increase the utility of the EnGMF by introducing an ensemble-local notion of covariance into the kernel density estimation (KDE) step for the prior distribution. We prove that in the Gaussian case, our new ensemble-localized KDE technique is exactly the same as more traditional KDE techniques. We also show an example of a non-Gaussian distribution that can fail to be approximated by canonical KDE methods, but can be approximated well by our new KDE technique. We showcase our new KDE technique on a simple bivariate problem, showing that it has nice qualitative and quantitative properties, and significantly improves the estimate of the prior and posterior distributions for all ensemble sizes tested. We additionally show the utility of the proposed methodology for sequential filtering for the Lorenz '63 equations, achieving a significant reduction in error, and less conservative behavior in the uncertainty estimate with respect to traditional techniques.

1

The sample covariance is a good measure of the global relationship between state variables, but fails to capture local relationships between variables for distributions whose curvature is highly non-Gaussian. Canonical kernel density estimation techniques [1, 2] often rely only on the global curvature of the data, and are thus ill-suited for advanced state estimation techniques relying thereon.

The first law of geography [3] states that "everything is related to everything else, but near things are more related than distant things." In the context of state estimation and data assimilation for geophysical systems [4, 5] this law begets heuristics known as covariance localization. Common approaches to covariance localization rely on spatial distances between state and observation variables [4]. The two predominant approaches are B-localization, where the prior covariance is scaled in a way such that far away state variables do not influence each other, and R-localization, where the observation covariances are scaled in a way such that state variables that are far away from an observation are not influenced by it. These approaches fundamentally rely on a spatial understanding of state variables or observations, and are not fit for general problems where the state variables might not have a complete spatial structure. A new more general approach to localization is therefore required.

The aim of this work is to build a new kernel density estimation (KDE) technique for particle filtering. While several computationally-intensive methods exist for density estimation in particle filters, such as clustering [6] and 'blob' filtering [7, 8], our goal is to construct a method that can both perform accurate kernel density estimation while simultaneously being fast enough to utilize in an online setting.

2

The previous conference paper [9] on which this work is based, applied a largely heuristic idea of finding a local measure of covariance through the use of distance-based weightings. In this work we formalize the notion of local covariance by posing the problem in terms of recovering the optimal Gaussian covariance from a localized conditional distribution.

A related parallel effort [10] developed an EnGMF where each kernel's covariance is dependent on the relative density of neighboring particles. All the k closest particles in the state-space equally contribute to the computation of the covariance of a kernel. With the aid of k-dimensional trees, the cost in the covariance generation is small. The resulting filter is used to track maneuvering spacecraft, whose transitional prior distributions are strongly non-Gaussian.

We propose a novel approach to localization that does not require an innate spatial structure in the state variables, but instead relies on some measure of inter-particle distance. Continuing the naming trend, we call our approach E-localization, standing for ensemble localization. In this work we apply this methodology to kernel density estimation, creating the E-localized KDE (ELKDE) and to the EnGMF to beget the E-localized EnGMF (ELEnGMF).

We show that the proposed approach is able to more accurately describe the prior distribution with fewer samples, and that the resulting description of the prior leads to a better posterior. Results consist of both qualitative and quantitative analysis on a simple test problem. We additionally show that our approach begets better performance in the sequential filtering setting.

This work is organized as follows: we first provide some background on kernel density estimation, on the filtering problem, and on the EnGMF in section 1. We next introduce the E-localization methodology in section 2. Penultimately, we showcase the ELKDE on a bivariate spiral example, and the ELEnGMF on the Lorenz '63 equations, in section 3. Finally, we have some concluding remarks in section 4.

3

1 Background

We now introduce kernel density estimation (KDE) and the ensemble Gaussian mixture filter (EnGMF).

1.1 Kernel Density Estimation

Suppose that we have an ensemble $\mathbf{X}_N = [x_1, \ldots, x_N]$, of N samples of dimension n from the random variable X, which has the probability distribution p_X.

When we do not know p_X directly, our goal is to approximate p_X using the ensemble $\mathbf{X}_N$ and all our other knowledge. We can build $\widehat{p}_X$, a kernel-based estimate of p_X,

$$\widehat{p}_X(x) = \frac{1}{N} \sum_{i=1}^{N} K_{\Sigma_i}(x - x_i), \tag{1}$$

where each Σ_i is a positive definite matrix, and the kernels K_{Σ_i} are functions that satisfy the following,

$$
\begin{aligned}
& K_{\Sigma_i}(x) \geq 0, \forall x, \\
& \int_{\mathbb{R}^n} K_{\Sigma_i}(x)\mathrm{d}x = 1, \\
& \mathbb{E}[K_{\Sigma_i}(x)] = 0, \\
& 0 \prec \mathrm{Cov}(K_{\Sigma_i}(x)) \prec \infty, \\
& K_{\Sigma_i}(x) = |\Sigma_i|^{-1/2} K_I\left(\Sigma_i^{-1/2} x\right),
\end{aligned}
\tag{2}
$$

where I is the identity matrix. The first two properties ensure that the kernels are probability densities. The third ensures that they are mean zero, with the fourth ensuring that they have finite non-zero covariance (in the SPD sense). The last property relates all the different parameterizations of the kernels to one another through the kernel with unit covariance.

Note that there are $n(n-1)N+1$ parameters of interest in the KDE estimate eq. (1), namely the choice of kernel K, and the set of covariances $\{\Sigma_i\}_{i=1}^{N}$.

In this work we focus on the Gaussian kernel,

$$K_{\Sigma_i}(x) = (2\pi)^{-n/2}|\Sigma_i|^{-1/2}e^{-\frac{1}{2}(x^T\Sigma_i^{-1}x)}, \tag{3}$$

as this makes the Kernel density estimate a Gaussian mixture model (GMM). GMMs are useful in state estimation as a GMM prior with a linear GMM observation leads to a GMM posterior [11].

Given knowledge of the covariance $\mathfrak{S}$ of the data, the canonical choice of the covariances in eq. (1), is

$$\Sigma_i = \beta_N^2\mathfrak{S}, \quad i = 1, \dots, N, \tag{4}$$

where β_N^2 is a scalar *bandwidth* factor. The optimal bandwidth factor both depends on the choice of kernel K and crucially on the underlying distribution of the data. As the underlying distribution is typically not known, the scaling factor that is optimal under Gaussian assumptions is typically chosen.

The following result about the choice of optimal bandwidth is due to [1]:

Lemma 1. *Without proof, if the distribution of interest, p_X is Gaussian with mean $\mathfrak{m}$ and covariance $\mathfrak{S}$, and the kernel is Gaussian eq. (3), then the bandwidth factor,*

$$\beta_N^2 = \left(\frac{4}{N(n+2)}\right)^{\frac{2}{n+4}}, \tag{5}$$

minimizes the mean integral squared error,

$$\mathbb{E}_{\mathbf{X}_N}\left[\int_{\mathbb{R}^n}(p_X(x) - \widehat{p}_X(x))^2\,\mathrm{d}x\right], \tag{6}$$

between the true distribution and its KDE given by eq. (1).

The approach used in lemma 1 we term the canonical KDE, or CKDE for short.

While the result in lemma 1 is satisfactory for Gaussian distributions, it does not provide us any intuition about how to choose the covariances in eq. (1) for non-Gaussian KDE.

One approach is to manually tune the bandwidth estimate eq. (5), as it is known to typically be too large in the case of highly non-Gaussian data [1, 12]. A constant scaling factor,

$$\beta_N^2 \leftarrow s_\beta \beta_N^2, \tag{7}$$

with $0 < s_\beta$, can be applied to account for this deficiency. Tuning s_β is difficult to perform correctly in a sequential filtering scenario, as a constant scaling factor may not be optimal for the wide range of possible distributions that can arise. Another alternative is to determine the scaling factor in an adaptive, online, fashion [13] using the observation likelihood. The adaptive scaling of the covariances can still suffer from overfitting if the algorithm is not carefully tuned to the problem at hand. A third alternative idea, called adaptive kernel density estimation (AKDE) [1], is to create a sequence of adaptive scaling factors,

$$\lambda_i = (\widehat{p}_X(x_i)/g)^{-\alpha},$$
$$\text{where} \quad \log g = \frac{1}{N} \sum_{i=1}^{N} \log \widehat{p}_X(x_i), \tag{8}$$

and α is a scaling factor typically taken to be n^{-1}, and every covariance is scaled by the factor

$$\beta_i^2 = \lambda_i^2 \beta_N^2. \tag{9}$$

Yet still, the AKDE idea suffers from the fact that the local curvature of the KDE estimate is determined by the global covariance of the whole distribution, and thus might be a poor approximation to the true non-Gaussian distribution.

1.2 Ensemble Gaussian Mixture Filter

We first provide a brief overview of the state estimation/data assimilation problem. Assume that we wish to estimate the true state x^t of some physical process, but only have access to noisy non-linear observations of it,

$$y = h(x^t) + \eta, \tag{10}$$

where η is unbiased additive Gaussian noise,

$$\eta \sim \mathcal{N}(0, R), \tag{11}$$

and R is the observation error covariance. Given a random variable X^- that describes our prior knowledge about the state, we use Bayesian inference to find a random variable,

$$X^+ = X^- \big| (Y = y), \tag{12}$$

that describes our posterior knowledge of the state, given the newfound observation eq. (10).

We now describe the inner workings of the EnGMF [13–16], in order to understand some of its deficiencies.

The EnGMF operates in three discrete steps: (i) estimation of the prior density with an ensemble Gaussian mixture, (ii) update of the Gaussian mixture from prior to posterior, and (iii) resampling of exchangeable samples from the posterior.

Given an ensemble of N independently and identically distributed (iid) samples from the prior distribution, $\mathbf{X}_N^- = \left[x_1^-, \ldots, x_N^- \right]$, using the KDE form eq. (1) with Gaussian kernel eq. (3) to construct an approximation of the prior distribution as a

Gaussian mixture model (GMM),

$$\widehat{p}_{X^-}(x) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{N}\left(x\,;\,x_j^-\,,\,\widehat{\Sigma}_j^-\right), \tag{13}$$

where each mean is centered at one of the ensemble members (particles), and the covariances $\widehat{\Sigma}_j^-$ are chosen in a way that attempts to make the distribution estimate more accurate.

The EnGMF formulas, used to construct the posterior GMM estimate, are,

$$\begin{aligned}
x_j^{\sim} &= x_j^- - \mathbf{G}_j\left(h(x_j^-) - \mathbf{y}\right), \\
\widehat{\Sigma}_j^+ &= \left(\mathbf{I} - \mathbf{G}_j\mathbf{H}_i^T\right)\widehat{\Sigma}_j^-, \\
\mathbf{G}_j &= \widehat{\Sigma}_j^-\mathbf{H}_j^T\left(\mathbf{H}_j\widehat{\Sigma}_j^-\mathbf{H}_j^T + \mathbf{R}_i\right)^{-1}, \\
v_j &= \frac{\mathcal{N}\left(\mathbf{y}\,;\,|\,h(x_j^-),\,\mathbf{H}_j\widehat{\Sigma}_j^-\mathbf{H}_j^T + \mathbf{R}\right)}{\sum_{j=1}^{N}\mathcal{N}\left(\mathbf{y}\,;\,h(x_j^-),\,\mathbf{H}_j\widehat{\Sigma}_{i,j}^-\mathbf{H}_j^T + \mathbf{R}\right)}, \\
\mathbf{H}_j &= \left.\frac{dh}{dx}\right|_{x=x_j^-},
\end{aligned} \tag{14}$$

where $x_j^{\sim}$ are the means of the posterior GMM, $\widehat{\Sigma}_j^+$ are the posterior covariances, $\mathbf{G}_j$ are the particle gain matrices, v_j are the weights of the posterior GMM modes, and $\mathbf{H}_j$ are the linearizations of the observation operator eq. (10) about the ensemble members (particles). When the observations are linear, this update formula is exact.

Remark 1. *Note that the gain matrices $\mathbf{G}_j$ in eq. (14) are actually statistical gains, and as such are biased estimators [17] of the exact gains. Techniques such as ensemble inflation can be applied to compensate for this bias, though in practice, for the EnGMF with resampling, this issue has not shown to have a significant impact.*

Remark 2. *The mean and covariance updates in eq. (14) behave in a similar fashion to that of the extended Kalman filter. While this ensures that the EnGMF converges [13] to 'exact Bayesian inference', in the case of a non-linear model and non-Gaussian distribution, a poor description of the prior would lead to an even worse description of the posterior. An alternative formulation, using the unscented transform has been utilized by [18, 19].*

Through the use of the mean and covariance updates in eq. (14), the estimate of the posterior GMM can be constructed,

$$\widehat{p}_{X^+}(x) = \sum_{j=1}^{N} v_j \, \mathcal{N}\left(x\,;\, x_{\tilde{j}}^{\sim}, \, \widehat{\Sigma}_j^+\right), \tag{15}$$

which is an approximation of the true posterior.

In the final step of the EnGMF, an ensemble of exchangeable samples from the posterior GMM eq. (15) is computed, resulting in $\mathbf{X}_N^+ = \left[x_1^+, \ldots, x_N^+\right]$. There are many ways by which a Gaussian mixture model can be sampled [15, 20]. In this work we focus on the naive approach. First sample N indices j from the distribution defined by the weights v_j from eq. (14). For each index, take its corresponding Gaussian component in eq. (15) and sample a random variable from it, by the canonical procedure. For details on implementation of the resampling procedure for Gaussian mixture models see [15].

Remark 3. *This resampling process does not generate iid samples from the approximate posterior distribution, as the posterior GMM weights, means, and covariances are all sample statistics, and are thus random variables derived from the prior ensemble. This means that samples from the posterior distribution are all conditioned on the prior ensemble, and the samples are therefore merely exchangeable. In practice, the error of treating these samples as iid does not seem to make a significant impact when the ensemble size is large enough.*

9

The choice of covariance in the prior GMM eq. (13) dictates the convergence of the EnGMF algorithm. In the canonical EnGMF [14] the prior covariances are chosen as to align with lemma 1, with the global covariance of the system estimated by the empirical covariance,

$$\widehat{\Sigma}^- = \frac{1}{N-1} \mathbf{X}_N^- \left(\mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) \mathbf{X}_N^{-,T}, \tag{16}$$

which is an unbiased estimator of the true covariance when the samples $\mathbf{X}_N$ are iid.

In state estimation and data assimilation, the prior state X^- is information sampled at a preceding step and propagated through a highly non-linear model, thus the distributions resulting from this propagation become highly non-Gaussian even if originally sampled from a Gaussian; consequently the empirical central second moment of the distribution eq. (16) carries less of the total information than for Gaussian filters.

Though the EnGMF contains the word "ensemble", it is really a particle filter, though the distinction is not discussed further in this work.

1.3 Motivation through a bimodal example

We now show why the CKDE estimate with the Silverman bandwidth lemma 1 is a poor choice for some non-Gaussian distributions.

We construct a bivariate (in terms of two state variables $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$) Gaussian mixture model consisting of two constituent terms such that the local covariance of each constituent term in the mixture is significantly different from the global covariance of the total mixture. Each constituent term has the same local covariance,

$$\mathfrak{S}_{\text{local}} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{17}$$

where ρ is the correlation coefficient. We separate the modes of the two Gaussian terms in the x_2 direction by a constant half-height υ, setting the means of the two constituent terms to,

$$\mathfrak{m}_1 = \begin{bmatrix} 0 & \upsilon \end{bmatrix}^T, \quad \mathfrak{m}_2 = \begin{bmatrix} 0 & -\upsilon \end{bmatrix}^T. \tag{18}$$

Explicitly, the Gaussian mixture distribution that we analyze is,

$$p_X(x) = \frac{1}{2} \left[\mathcal{N}(x; \mathfrak{m}_1, \mathfrak{S}_{\text{local}}) + \mathcal{N}(x; \mathfrak{m}_2, \mathfrak{S}_{\text{local}}) \right], \tag{19}$$

where each term is equally weighted.

By a simple application of known formulas [21] the global covariance of eq. (19) is given by,

$$\mathfrak{S}_{\text{global}} = \begin{bmatrix} 1 & \rho \\ \rho & 1 + \upsilon^2 \end{bmatrix}, \tag{20}$$

meaning that as we increase the separation half-height υ, the second state variable starts to dominate the global covariance.

We can formalize this intuition by looking at the dominant eigenvectors of the covariances. When the correlation coefficient ρ is positive, the dominant eigenvector of the local covariance eq. (17) is,

$$\begin{bmatrix} 1 & 1 \end{bmatrix}^T, \tag{21}$$

which is not dependent on either the correlation coefficient ρ or the separation half-height υ. The dominant eigenvector of the global covariance eq. (20), on the other hand, is,

$$\begin{bmatrix} \frac{-\upsilon^2 + \sqrt{\upsilon^4 + 4\rho^2}}{2\rho} & 1 \end{bmatrix}^T, \tag{22}$$

which tends towards an eigenvector of,

$$\begin{bmatrix} 0 & 1 \end{bmatrix}^T, \tag{23}$$

11

as the separation half-height υ tends towards infinity. The eigenvalue corresponding to this eigenvector is,

$$\frac{1}{2}\left(\sqrt{\nu^4 + 4\rho^2} + \nu^2 + 2\right), \tag{24}$$

which tends towards infinity as the separation half-height ν increases. This creates a large discrepancy between the global covariance eq. (20), and the local covariance eq. (17). Thus, the CKDE approximation induced by the bandwidth scaling of the global covariance eq. (4) fails to capture the local density for a sufficiently large ν. In other words, a distribution that is representable in terms of a Gaussian mixture with finite covariances would degenerate to not being representable using the CKDE method.

From this example, it is clear that a method of determining the local covariance structure of the distribution is needed.

2 Ensemble-localized Kernel density estimation

When the distribution of interest p_X is non-Gaussian, the global covariance can create a poor approximation of the local relationship of the data. Instead of looking for the optimal bandwidth like in lemma 1, we go back to the fundamentals of KDE estimates eq. (1), and attempt to find covariance matrices $\{\Sigma_i\}_{i=1}^{N}$ that approximate the distribution locally around each particle.

We focus on attempting to understand the local covariance structure of the distribution, p_X, around a sample, x_i, therefrom. Essentially, we attempt to answer the following question: if the local behavior of the distribution around the point x_i extended globally, what would the covariance of the resulting distribution be?

In order to attempt to answer this question, we introduce the synthetic random variable $\mathcal{L}$ with the same support as X and whose conditional distribution $\mathcal{L}|X = x$ is the description of the local information around the point x, and attempt to extract the local covariance around x through understanding the distribution of $X|\mathcal{L} = x_i$ which

is the distribution of X conditioned on the local information around the realization x_i. We term this idea 'E-localization', and its application to kernel density estimation, E-localized KDE or 'ELKDE' for short.

From Bayes' rule, we can see that,

$$p_{X|\mathcal{L}=x_i}(x) \propto p_{\mathcal{L}|X=x}(x_i)\, p_X(x), \tag{25}$$

thus, the knowledge contained in the distribution of $p_{\mathcal{L}|X=x}$ is sufficient to explain all the additional, local, information.

In this work we explore the following choice:

$$p_{\mathcal{L}|X=x}(\xi) = Ce^{-\frac{1}{2}((\xi-x)^T S_\xi^{-1}(\xi-x))}, \tag{26}$$

where C is a normalizing constant and the symmetric positive definite matrix S_ξ is an application-specific tunable parameter that is unique to every ξ around which we wish to localize. This means, in effect that a different choice of eq. (26) is made for every individual particle.

Our goal is to find a covariance matrix, $\widetilde{\Sigma}_i$, that represents the local covariance structure of p_X around the point x_i. In the case when the random variable X is Gaussian, the ELKDE method should exactly recover the CKDE method, providing equivalence and some limited sense of optimality by lemma 1.

Thus, we first look at the case when p_X is Gaussian with known mean and covariance. If the distribution is Gaussian, its local behavior around every single point has to be Gaussian as well. Thus, it should be possible to recover the global covariance from the information contained in $p_{X|\mathcal{L}=x_i}$, for any choice of the parameter S_ξ in eq. (26).

13

In the subsequent proof, we show that by only knowing information about $p_{X|\mathcal{L}}$ and the covariance S_ξ from eq. (26), it is possible to recover the covariance of the information contained within $p_X(x)$, with the knowledge of X is Gaussian.

Theorem 1 (Approximate Gaussian covariance with E-localization). *Take the random variable X to be Gaussian,*

$$p_X(x) = \mathcal{N}(x; \mathfrak{m}, \mathfrak{S}), \tag{27}$$

with p_X being the distribution, $\mathfrak{m}$ the mean, and $\mathfrak{S}$ the covariance. Take also the ensemble of N samples $\mathbf{X} = [x_1, \ldots, x_N]$ therefrom.

If the distribution of the conditional realizations $\{x_i\}_{i=1}^N$ are as in eq. (26),

$$p_{\mathcal{L}|X=x}(x_i) = \frac{C}{|S_i|^{\frac{1}{2}}} e^{-\frac{1}{2}((x_i - x)^T S_i^{-1} (x_i - x))}, \tag{28}$$

with matrices $\{S_i\}_{i=1}^N$, standing for $\{S_{x_i}\}_{i=1}^N$, then the ensemble localized density estimate,

$$\widehat{p}_X(x) = \frac{1}{N} \sum_{i=1}^N K_{\widehat{\Sigma}_i}(x - x_i), \tag{29}$$

with covariances,

$$
\begin{aligned}
\widehat{\Sigma}_i &= \beta_N^2 \widetilde{\Sigma}_i, \\
\widetilde{\Sigma}_i &= \text{Cov}(X|\mathcal{L}=x_i)[S_i - \text{Cov}(X|\mathcal{L}=x_i)]^{-1} S_i,
\end{aligned}
\tag{30}
$$

where $i = 1, \ldots N$, is exactly equivalent to the canonical kernel density estimate from lemma 1 for any choice of $\{S_i\}_{i=1}^N$.

Proof. It suffices to show that for all i, every local covariance is equal to the global Gaussian covariance, $\widetilde{\Sigma}_i = \mathfrak{S}$. From the Kalman filter equations, it is known that,

$$\text{Cov}(X|\mathcal{L}=x_i) = \mathfrak{S} - \mathfrak{S}(\mathfrak{S} + S_i)^{-1}\mathfrak{S},$$
$$= \mathfrak{S}\ (\mathfrak{S} + S_i)^{-1}(\mathfrak{S} + S_i) - \mathfrak{S}(\mathfrak{S} + S_i)^{-1}\mathfrak{S},$$
$$= \mathfrak{S}(\mathfrak{S} + S_i)^{-1}S_i,$$

therefore the covariances are,

$$\widetilde{\Sigma}_i = \mathfrak{S}(\mathfrak{S} + S_i)^{-1}S_i \left[S_i - \mathfrak{S}(\mathfrak{S} + S_i)^{-1}S_i \right]^{-1} S_i,$$
$$= \mathfrak{S}(\mathfrak{S} + S_i)^{-1}S_i \left[S_i^{-1} + S_i^{-1}\mathfrak{S}S_i^{-1} \right] S_i = \mathfrak{S},$$

independent of S_i, as required. $\qquad\qquad\square$

It is also of note that in the Gaussian case $\text{Cov}(X|\mathcal{L}=x_i)$ does not depend on the outcome x_i, but only on the covariance parameter S_i which is not generally true in the non-Gaussian case. In the general case, we make the notational definition,

$$\mathfrak{C}(x_i, S_i) := \text{Cov}(X|\mathcal{L}=x_i). \tag{31}$$

We now look at what happens when the distribution p_X is non-Gaussian. If the covariance S_i from eq. (26) tends towards infinity, then the local covariance from eq. (30) should tend towards the global covariance of the distribution of X.

Corollary 2 (Local covariance becomes global). *When the Gaussian distribution assumptions in theorem 1 are relaxed and X is taken from some arbitrary non-Gaussian distribution, the local covariance matrix $\widetilde{\Sigma}_i$ from eq. (30), approaches the*

15

global covariance as the synthetic covariance S_i from eq. (26) approaches infinity,

$$\lim_{|S_i| \to \infty} \widetilde{\Sigma}_i = \mathrm{Cov}(X), \tag{32}$$

where $\mathrm{Cov}(X)$ is the global covariance of the (now) non-Gaussian random variable X, and $|S_i|$ is some notion of generalized variance.

From this we can glean that the larger the synthetic covariance S_i gets, the less local the information becomes, and the covariance approaches the global estimate. It is therefore reasonable to ask if the converse is also true: as the covariance parameter S_i approaches zero, does the covariance in eq. (30) tend towards a local description of the covariance around the point x_i?

Remark 4 (Generalized variance). *It is of note that generalized variance in corollary 2 can take multiple forms. Common choices include,*

$$|S_i| = \det S_i, \tag{33}$$

which is the determinant of the matrix S_i, and,

$$|S_i| = \mathrm{tr}\, S_i, \tag{34}$$

which is its trace. The choice of generalized variance is ancillary to the main results of this work.

It is of note that the covariance eq. (30) in theorem 1 is only a good estimate in the case that the distribution of $X|\mathcal{L}$ is actually Gaussian. It is possible—either exactly, or numerically—that,

$$S_i \not\succ \mathfrak{C}(x_i, S_i), \tag{35}$$

16

which would make the covariance eq. (30) not positive definite, thus, in order to generalize theorem 1 to non-Gaussian distributions, an alternate formulation of the covariance must be chosen.

Utilizing the information we have gleaned from theorem 1, from corollary 2, and from the above discussion, we propose a definition of the local covariance matrix for generic, non-Gaussian distributions:

Definition 1 (Local covariance). *Given a random variable X with finite first two moments, and given a Gaussian random variable $\mathcal{L}|X\!=\!x_i$ that encapsulates our local uncertainty around x_i with covariance S_i, from eq. (26), the covariance that extends the local behavior of the distribution globally around the point x_i is given by,*

$$\widetilde{\Sigma}_i := \widetilde{\Sigma}(x_i, S_i) = \mathfrak{C}(x_i, S_i)[S_i - \mathfrak{C}(x_i, S_i)]^{-1} S_i,$$

$$where \quad \left\{ \begin{array}{c} S_i = \underset{S \in SP(n)}{\arg\min} |S|, \\[2mm] such\ that \quad \widetilde{\Sigma}(x_i, S_i) \in SP(n), \end{array} \right\}, \tag{36}$$

meaning that the local covariance $\widetilde{\Sigma}_i$ is the covariance that minimizes its argument S_i while remaining in the set of all n-dimensional symmetric positive definite matrices, $SP(n)$.

Definition 1 is one attempt at extending theorem 1 to non-Gaussian distribution through the same reasoning as in corollary 2. However, it is not the only possible choice of extension. The optimality of formulation proposed therein is of relevant interest but beyond the scope of this work.

Note that when the random variable X is Gaussian, the proposed sense of local covariance in definition 1 is exactly equivalent to the derivation in theorem 1.

17
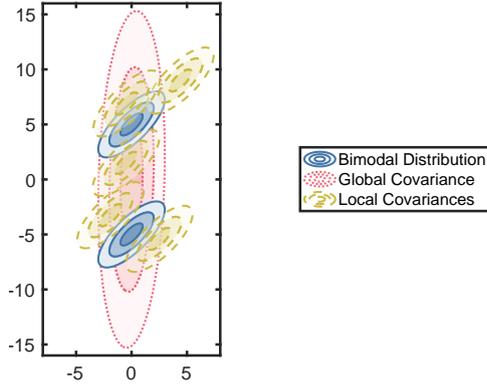
Fig. 1 A demonstration of the local covariance technique definition 1 on the bimodal example section 1.3. The blue ellipses with solid lines represent the 1, 2, and 3-σ bounds on the true bimodal distribution. The large red vertical ellipses with dotted outlines represent the global covariance. The yellow ellipses with dashed outlines represent the local covariances centered at hand-sampled points from the true distribution.

Remark 5 (Using $\mathfrak{C}(x_i, S_i)$). *In [9], on which this work is based, a scaling of the conditional covariance matrix*

$$\Sigma_i = s_{\mathcal{L},i}\mathfrak{C}(x_i, S_i), \tag{37}$$

where $s_{\mathcal{L},i}$ is a scaling factor determined for each covariance, was used for the covariance estimates. As there are no formal guarantees on when such a covariance is optimal, this idea is not further explored in this work.

2.1 Revisiting the bimodal example

We now revisit the bimodal example in section 1.3, and show that the proposed ELKDE method is robust to increase in the separation half-width parameter ν.

Take a sample x_i from the distribution defined by eq. (19), then the distribution of X given $\mathcal{L} = x_i$, would be defined by,

$$
\begin{aligned}
p_{X|\mathcal{L}=x_i}(x) = \ & u_1\mathcal{N}(\ell_1, \mathfrak{S}_{\text{local}}(\mathfrak{S}_{\text{local}} + S_i)^{-1}S_i) \\
& + u_2\mathcal{N}(\ell_2, \mathfrak{S}_{\text{local}}(\mathfrak{S}_{\text{local}} + S_i)^{-1}S_i),
\end{aligned}
\tag{38}
$$

18

where the means ℓ_1 and ℓ_2 are of no consequence.

As $\nu \to \infty$, the sample almost surely comes from the mode whose mean is closest. Without loss of generality, assume that this is the first mode, thus the weights collapse, $u_1 \to 1$ and $u_2 \to 0$, which can be seen from the EnGMF equations eq. (14). Therefore in this case the distribution collapses to a simple Gaussian, thus by theorem 1, the covariance around the point x_i by the ELKDE method is a scaling of $\mathfrak{S}_{\text{local}}$, and not of $\mathfrak{S}_{\text{global}}$ as by the CKDE method.

Moreover, in the case of a fixed $\nu = 5$, for any point in the distribution, the local covariance as computed by the optimization problem in definition 1, is always the local covariance. This can be visually seen in fig. 1 where the bimodal distribution, the global covariance and the local covariances at select points are plotted. As can be seen, the local covariances produced by definition 1, with determinant generalized variance as a metric, are identical to the covariances of the bimodal distribution, and do not resemble the global covariance.

We thus show that the ELKDE method is robust to the degenerate behavior shown in section 1.3.

2.2 Practical Implementation

There are several considerations to be made before a practical implementation of the covariance estimate in definition 1 can be attempted.

First we have to consider how to compute the covariance matrix $\mathfrak{C}(x_i, S_i)$, from eq. (31), given only the ensemble $\mathbf{X}$. We begin by defining local weights,

$$w_{i,j} \propto \mathcal{N}(x_i; x_j, S_i) \tag{39}$$

such that the column vector of weights $\mathbf{w}_i$ defines the importance of each ensemble member given the random variable $\mathcal{L} = x_i$. The covariance can then be estimated

from the ensemble and the importance weights by the unbiased estimator,

$$\mathfrak{C}(x_i, S_i) \approx \frac{1}{1 - \mathbf{w}_i^T \mathbf{w}_i} \left[\mathbf{X} \left(\boxbslash \mathbf{w}_i - \mathbf{w}_i \mathbf{w}_i^T \right) \mathbf{X}^T \right], \tag{40}$$

where $\boxbslash$ is the vector to diagonal matrix operator.

A practical numerical consideration is the problem of degenerate weights, meaning that the covariance estimate in eq. (40) can be numerically singular. A simple way to overcome this is by slightly nudging the weights towards uniformity,

$$w_{i,j} \leftarrow (1 - \alpha)w_{i,j} + \frac{\alpha}{N}, \tag{41}$$

where the factor α is chosen to be small. In this work we fix $\alpha = 10^{-4}$.

The local covariance in definition 1 requires the solution of a matrix-valued constrained optimization problem, which is practically useful. We can approximate the local covariance optimization eq. (36) by the simpler formula,

$$\widetilde{\Sigma}_i = \lim_{|S_i| \to 0} \Pi_{\mathrm{SP}(n)} \left(\mathfrak{C}(x_i, S_i)[S_i - \mathfrak{C}(x_i, S_i)]^{-1} S_i \right), \tag{42}$$

where $\Pi_{\mathrm{SP}(n)}$ is the projection of a matrix onto the set of n-dimensional symmetric positive definite (SPD) matrices.

The next consideration is the choice of covariance matrix S_i. If our ensemble $\mathbf{X}$ is finite, then it is impractical to choose S_i to be arbitrarily small, thus the aim is to find a formula for S_i that tends towards zero as the ensemble size grows, but still encapsulates numerically useful information about the distribution from the finite ensemble.

As our aim is to limit the impact of far away samples and only consider the impact of local samples, we choose,

$$S_i = r_i^2 I, \tag{43}$$

where the scalar 'radius', r_i is chosen such that the weights eq. (39) are meaningful in some sense. To that end we again take inspiration from Kernel density estimation literature [1] and from the ensemble transform particle filter [22] in determining the localization radius. For the ith particle, take the distance to the $\sqrt{N}$th (rounded) nearest neighbor according to some distance $d(i, k)$, and call it d_i. This idea is loosely based on the use of inter-particle distance in the ensemble transform particle filter (ETPF) [22], though in the ETPF the inter-particle distance is used to solve an optimal transport problem. We take the localization radius to be a scaling of this distance,

$$r_i = s_r d_i, \tag{44}$$

where s_r is the radius scaling factor (in this work taken to be $s_r = 1$). Thus the covariance S_i should converge to 0 as $N \to \infty$ in probability, as long as the distribution of interest has connected support. The above procedure is a modification of the one used in [10].

Remark 6. *Note that taking $r_i \to 0$ is equivalent to taking any measure of generalized variance remark 4 to zero, meaning that $|S_i| \to 0$ as in eq. (42).*

Remark 7 (Using S_i as the local covariance)**.** *An immediately obvious question to ask, is why not just use S_i for the covariances in eq. (1) instead, such as,*

$$\widehat{p}(x) = \sum_{i=1}^{N} \mathcal{N}(x; x_i, S_i). \tag{45}$$

While this estimate would converge to the exact distribution from which the samples originate given that the distribution has support over all real space, it would not converge to the optimal estimate in lemma 1. Additionally, the distribution in eq. (45) would not even account for the global scaling of the distribution, as the covariance would be diagonal.

The final consideration is finding a useful projection operator $\Pi_{\mathrm{SP}(n)}$ in definition 1. In order to see the two different ways in which this projection can be performed, it is useful to think about the logarithm of the covariance in eq. (30),

$$\log \widetilde{\Sigma}_i = \log \left[\mathrm{Cov}(X|\mathcal{L}=x_i)[S_i - \mathrm{Cov}(X|\mathcal{L}=x_i)]^{-1} S_i \right], \tag{46}$$

$$\begin{aligned} = \log \mathrm{Cov}(X|\mathcal{L}=x_i) - \log\left[S_i - \mathrm{Cov}(X|\mathcal{L}=x_i)\right] \\ + \log S_i, \end{aligned} \tag{47}$$

as the logarithm of the covariance merely has to be a symmetric matrix. There are two key elements, first ensuring that the final covariance estimate is SPD, which is equivalent to ensuring that eq. (46) is symmetric and real, and the second is ensuring that certain key constituent covariances in the calculation themselves are SPD, which is equivalent to ensuring that every matrix logarithm in eq. (47) is symmetric and real.

Let the operator on the arbitrary symmetric matrix A,

$$\mathcal{E}_\epsilon A, \tag{48}$$

define the operation that bounds the eigenvalues of A by ϵ from below. We can first think about ensuring that the matrix eq. (46) is symmetric and real, then one possible projection operator is,

$$\Pi_{\mathrm{SP}(n)} = \mathcal{E}_\epsilon, \tag{49}$$

which simply eliminates possible negative eigenvalues.

If we instead concentrate on ensuring that the logarithms in eq. (47) are symmetric and real, then one possible projection operator is,

$$\Pi_{\mathrm{SP}(n)}(\cdot) = \mathcal{E}_{\epsilon_1} \left[\mathrm{Cov}(X|\mathcal{L})[\mathcal{E}_{\epsilon_2}(S_i - \mathrm{Cov}(X|\mathcal{L}))]^{-1} S_i \right], \tag{50}$$

22

Algorithm 1 Local covariance estimation in the ELKDE

Input: Ensemble $\mathbf{X}_N$, projection function $\Pi_{\mathrm{SP}(n)}$
Output: E-localized covariances $\{\Sigma_i\}_{i=1}^N$

  1: **for** $i = 1$ to N **do**
  2:     % Find the distances between x_i and other samples
  3:     **for** $j = 1$ to N **do**
  4:        $d_{ij} \leftarrow \|x_i - x_j\|$
  5:     **end for**
  6:     $r_i \leftarrow \sqrt{N}$th sorted d_{ij}
  7:     $S_i \leftarrow r_i^2 I_n$
  8:     % Find the weights of the local information
  9:     **for** $j = 1$ to N **do**
10:        $a_{ij} \leftarrow \log \mathcal{N}(x_j; x_i, S_i)$
11:     **end for**
12:     % Numerically stable algorithm for computing the weights
13:     $\mathbf{w}_i \leftarrow \exp\left(\mathbf{a}_i - \mathrm{LogSumExp}(\mathbf{a}_i)\right)$
14:     $\mathrm{Cov}(X|\mathcal{L}) \leftarrow$ by eq. (40)
15:     % Compute the unprojected estimate of the covariance
16:     $\widetilde{\Sigma}_i \leftarrow \mathrm{Cov}(X|\mathcal{L})[S_i - \mathrm{Cov}(X|\mathcal{L})]^{-1} S_i$
17:     % Compute the projected and scaled covariance
18:     $\Sigma_i = \beta_N^2 \Pi_{\mathrm{SP}(n)}(\widetilde{\Sigma}_i)$
19: **end for**

which bounds the eigenvalues not only of the matrix, but of the constituent terms as well.

We provide an algorithmic overview of the E-localization procedure for finding the local covariance matrices in the ELKDE in algorithm 1.

3 Numerical Experiments

The goal of our numerical experiments is to first showcase the utility of the ELKDE approach on a simple-to-visualize data set, for which we use a simple bivariate spiral. Our second experiment, utilizing a more complicated chaotic dynamical system fulfils the second goal, to showcase the ELKDE methodology on the EnGMF, and thus showcase its utility for sequential filtering.
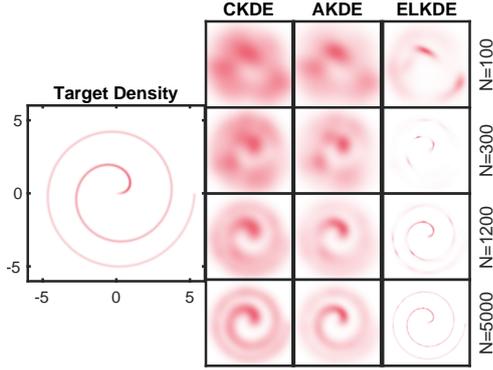
Fig. 2 A qualitative look at the accuracy of the different KDE methods explored in this work for the continuous Gaussian mixture model described in eq. (51). The true distribution is shown on the left, the canonical KDE method is labeled by 'CKDE', the adaptive KDE method is labeled as 'AKDE', and the E-localized KDE method is labeled as 'ELKDE'.
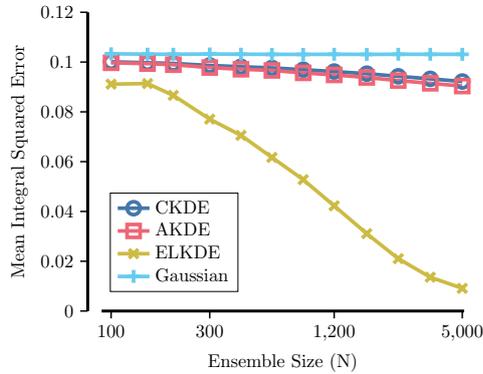
Fig. 3 A quantitative look at the accuracy of the different KDE methods explored in this work for the continuous Gaussian mixture model described in eq. (51) of ensemble size versus error in terms of MISE. The canonical KDE method is labeled by 'CKDE', the adaptive KDE method is labeled as 'AKDE', and the E-localized KDE method is labeled as 'ELKDE'. Additionally, the approximation by the empirical Gaussian distribution is provided as a baseline.

3.1 A bivariate spiral

Our first numerical experiment is meant to test the accuracy of the ELKDE in a highly non-Gaussian setting.

Consider the following distribution of continuous Gaussian mixture [23] form,

$$p_X(x) = \frac{1}{4\pi} \int_0^{4\pi} \mathcal{N}(x; \mathfrak{m}(z), \sigma^2 I_2)\, \mathrm{d}z, \tag{51}$$

24

where I_2 is the bivariate identity matrix, and the constituent mean,

$$\mathfrak{m}(z) = \frac{3}{2}\sqrt{z}\begin{bmatrix} \cos(z) \\ \sin(z) \end{bmatrix}, \tag{52}$$

represents a Fermat spiral [24], and the small covariance, defined by $\sigma^2 = 2^{-8}$, represents the local covariance of each one of the infinitesimal modes.

The mean of this distribution is given by the closed form expression,

$$\mathbb{E}[X] = \frac{3}{8\sqrt{2\pi}}\begin{bmatrix} -S\left(2\sqrt{2}\right) \\ C\left(2\sqrt{2}\right) - 2\sqrt{2} \end{bmatrix} \approx \begin{bmatrix} -0.06 \\ -0.35 \end{bmatrix}, \tag{53}$$

where S and C are the Fresnel functions [25], and the covariance $\mathrm{Cov}(X) = \mathfrak{S}$ is,

$$\begin{aligned} \mathfrak{S}_{1,1} &= \frac{9\pi}{4} - \frac{9S\left(2\sqrt{2}\right)^2}{128\pi} + \frac{1}{256} \\ \mathfrak{S}_{2,1} &= -\frac{9\left(\left(2\sqrt{2} - C\left(2\sqrt{2}\right)\right)S\left(2\sqrt{2}\right) + 8\pi\right)}{128\pi} \\ \mathfrak{S}_{2,2} &= \frac{9\left(4\sqrt{2}C\left(2\sqrt{2}\right) - C\left(2\sqrt{2}\right)^2 - 8 + 32\pi^2\right)}{128\pi} \\ &\quad + \frac{1}{256} \end{aligned} \tag{54}$$

which is approximately,

$$\mathfrak{S} \approx \begin{bmatrix} 7.07 & -0.58 \\ -0.58 & 6.95 \end{bmatrix}. \tag{55}$$

This distribution is particularly challenging as it requires an increasing number of of Gaussian mixture terms to approximate accurately, as it cannot be fully represented by a finite Gaussian mixture. Additionally, there is three orders of magnitude separating the local and global covariance terms, thus the CKDE and AKDE techniques should have particular difficulty with this problem.

We present results on various amount of samples N ranging from $N = 100$ to $N = 5000$, and make use of the MISE eq. (6) as a measure of error. We compute the MISE over 12 Monte Carlo runs for each ensemble size N. The samples from the spiral distribution eq. (51) are computed exactly, while the MISE computation uses a uniformly spaced approximation to the integral with 10000 evenly spaced points of z. We compare the CKDE, the AKDE, and the ELKDE methods.

For this experiment we make use of the second projection method from eq. (50), with parameters $\epsilon_1 = 10^{-4}$ and $\epsilon_2 = 10^{-2}$ as this method produces covariance estimates that are in line with performing the optimization procedure in definition 1.

A qualitative look at the accuracy of the methods can be seen in fig. 2. It can be seen that the CKDE method suffers from the fact that the global estimate of the covariance is not appropriate for approximating the target probability density. The AKDE method seems to be able to better resolve the form of the spiral, though not significantly better than the CKDE method. The ELKDE method appears to be able to fully resolve the correct spiral with $N = 1200$ and $N = 5000$ samples, with a partially resolved spiral visible for as little as $N = 300$ samples.

For the quantitative experiment, we additionally look at approximating the distribution by the empirical Gaussian as a baseline. As can be seen in fig. 3, the CKDE is a slight improvement over the Gaussian approximation, with the AKDE having a noticeable, but marginal impact on the error of the CKDE method, though both global methods barely do better than the Gaussian baseline. The ELKDE method, on the other had, shows significant improvement, especially in the slope of the improvement with respect to the number of samples. For $N = 5000$, the ELKDE method results in a MISE of one order of magnitude less than the other methods.
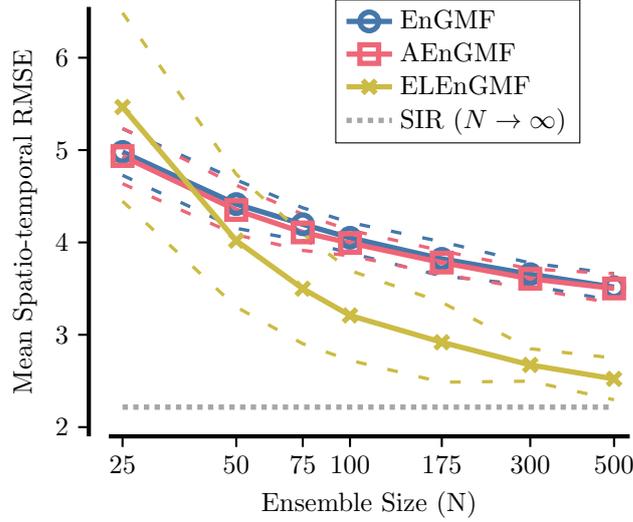
Fig. 4 A quantitative look at the performance of the EnGMF with CKDE prior estimation, the AEnGMF with AKDE prior estimation, and the ELEnGMF with ELKDE prior estimation for the Lorenz '63 equations. The x-axis represents the ensemble size N, and the y-axis represents mean spatio-temporal RMSE eq. (58) over 12 separate Monte Carlo runs. The solid lines represent the mean of the error while the dashed lines represent three standard deviations of the error over the Monte Carlo simulations. The gray dotted line represents the theoretical minimum RMSE computed using a sequential importance resampling (SIR) filter with $N = 25000$ ensemble members and optimal rejuvenation.
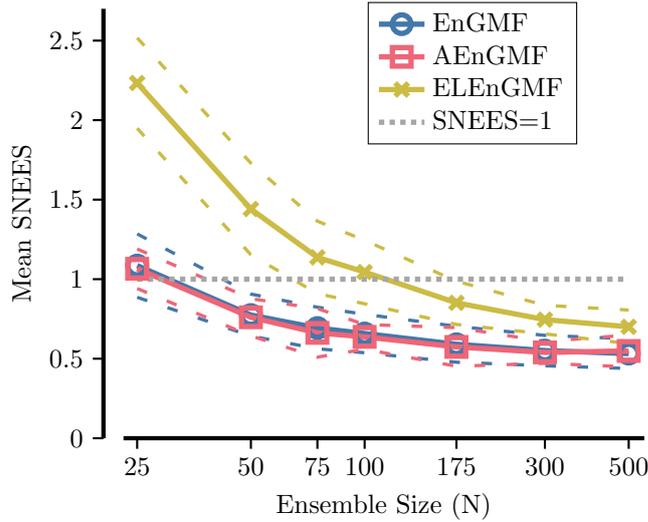
Fig. 5 A quantitative look at the performance of the EnGMF with CKDE prior estimation, the AEnGMF with AKDE prior estimation, and the ELEnGMF with ELKDE prior estimation for the Lorenz '63 equations. The x-axis represents the ensemble size N, and the y-axis represents the mean SNEES eq. (59) over 12 separate Monte Carlo runs. The solid lines represent the mean of the error while the dashed lines represent three standard deviations of the error over the Monte Carlo simulations. The gray dotted line represents the ideal SNEES of one.

27

3.2 Lorenz '63 sequential filtering example

The final round of experiments aims to put the ELEnGMF to practical use in a sequential filtering experiment. In sequential filtering our aim is to find an optimal estimate of the truth x^t at some time index i. This is performed by forecasting a state estimate to time index i to obtain a prior and performing inference eq. (12) to obtain a posterior estimate of the state x_i^+, then repeating the cycle over and over again *ad infinitum*.

To this end we use the three variable Lorenz '63 equations [26, 27],

$$
\begin{aligned}
\dot{x}_1 &= 10(x_2 - x_1), \\
\dot{x}_2 &= x_1(28 - x_3) - x_2, \\
\dot{x}_3 &= x_1 x_2 - \frac{8}{3}x_3,
\end{aligned}
\tag{56}
$$

which is a chaotic test problem that is one of the foundational problems for particle filters [5].

The following problem setup is taken from [9, 13]. We perform inference every $\Delta t = 0.5$ time, with 5500 sequential steps taken (the first 500 of which are discarded to account for spinup—discarding errors from initial over- or under-confidence of the algorithm). We take the scalar range observation,

$$
h(x) = \|x - c_2\|_2 \,,
\tag{57}
$$

where $c_2 = \begin{bmatrix} 6\sqrt{2} & 6\sqrt{2} & 27 \end{bmatrix}^T$ is the fixed point in the center of one of the wings of the Lorenz butterfly, with unbiased Gaussian observation error determined by the variance $R = 1$. All experiments are performed over 12 distinct Monte Carlo samples.

For the error metrics in this experiment, we look at the spatio-temporal root mean squared error (RMSE) and the scaled normalized estimation error squared (SNEES).

The spatio-temporal RMSE of the posterior means of the data with respect to the truth is,

$$\text{RMSE}(x^+) = \sqrt{\frac{1}{|x|} \sum_{i,j} \left(x_{i,j}^+ - x_{i,j}^t \right)^2} \tag{58}$$

where x^+ is the temporal collection of posterior means, x^t is the collection of true states of the system, i is the time index, j is the state index, and $|x|$ is the cardinality of the data. For the RMSE, the lower the better. The SNEES [16] of the posterior means with respect to the truth is defined as,

$$\text{SNEES}(x^+) = \frac{1}{n|x|} \sum_i (x_i^+ - x_i^t)^T \left(\mathfrak{S}_{i,\text{global}}^+ \right)^{-1} (x_i^+ - x_i^t), \tag{59}$$

where everything is the same as in eq. (58), except $\mathfrak{S}_{i,\text{global}}^+$, which represents the global covariance of the posterior uncertainty at time index i. A SNEES value of one is considered optimal, as the distribution of the error of the posterior with respect to the truth would have the same covariance as predicted by the filter. If the SNEES is less than one, then the filter is considered to be too conservative, while a filter with a SNEES greater than one is too confident in the estimate. In general, a conservative filter is preferable to one that is overconfident.

Remark 8. *The Lorenz '63 system has a Kaplan–Yorke dimension of 2.06 (computed using a method based on [28]) that is close to two, meaning that the covariance estimate in eq. (59) can frequently become close to singular numerically. We thus discard outlier values (those that are greater than 100 under the sum) in order to ensure that these numerical effects don't bias our results in a manner that would make their utility hard to interpret.*

For this experiment we make use of the first projection method from eq. (49), with parameter $\epsilon_1 = 10^{-4}$ as this method produces covariance estimates that are similar to those obtained by performing the optimization procedure in definition 1.

For the RMSE experiment in fig. 4, it can be seen that the ELEnGMF performs worse than the CKDE or the AKDE in the case of $N = 25$ ensemble members, but performs better in the case of $N \geq 100$ ensemble members, approaching the theoretical minimum RMSE (computed using a sequential importance resampling filter) at around $N = 500$ ensemble members, having a significant improvement in error over the global covariance filters.

For the SNEES experiment in fig. 5 it is evident that both the EnGMF and the AEnGMF behave almost identically, producing highly conservative estimates of the covariance of the system. This is not exactly surprising, as it can be shown [15] that the covariance of the prior distribution is a scaled by a factor of $1 + \beta_N^2$. The ELEnGMF is less conservative than the other two filters, as the individual estimates of the covariances allow the EnGMF to decrease the conservative nature of the prior estimate, translating to a less conservative estimate of the posterior. While the trend of the SNEES for all three methods is to grow lower and lower relative to the ensemble size, as all the methods are equivalent to a bootstrap particle filter in the ensemble limit, the SNEES should increase after a certain point, though the amount of ensemble members needed to accomplish this is unknown, and impractical to compute.

4 Conclusions

In this work we have introduced the E-localization methodology for kernel density estimation (ELKDE) and have applied it to the ensemble Gaussian mixture filter, creating the E-localized ensemble Gaussian mixture filter (ELEnGMF). We have shown that the ELKDE is theoretically equivalent to canonical KDE methods in the Gaussian case, and is empirically more accurate on a highly non-Gaussian spiral distribution. We have also shown that this methodology addresses one issue—the disparity between local and global notions of covariance—that arises in the application of KDE methods to distributions such as those arising from non-linear dynamics. Additionally, through

the application of the methodology to the Lorenz '63 equations, we have shown that the ELEnGMF has the potential to be a superior particle filter to the EnGMF in the sequential filtering setup.

While this work has tackled some issues inherent to the EnGMF, it has left many as open problems, and has even introduced some new issues. In particular the selection of localization radius r_i in eq. (43) is an issue of immediate interest. The methodology presented in this work for tackling this issue clearly has nice empirical behavior, but its justification is largely heuristic. Exploring this choice in future work is desirable. Another, more significant issue is the choice of projection method in eq. (49) and eq. (50), and whether such methods are even necessary, or if a more robust alternative is viable.

Future non-theoretic work would involve applying adaptive covariance parameterization techniques [13] to the ELEnGMF, such that the choice of bandwidth scaling factor eq. (7) and radius scaling factor eq. (44) could be performed adaptively in the sequential filtering regime. Another future direction would involve applying the ELEnGMF to a practical orbit determination problem [16].

Supplementary information. Not Applicable

Declarations

Competing interests

The authors have no competing interests to declare that are relevant to the content of this work.

References

[1] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Routledge, ??? (2018)

[2] Chen, Y.-C.: A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology **1**(1), 161–187 (2017)

[3] Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic geography **46**(sup1), 234–240 (1970)

[4] Asch, M., Bocquet, M., Nodet, M.: Data Assimilation: Methods, Algorithms, and Applications. SIAM, ??? (2016)

[5] Reich, S., Cotter, C.: Probabilistic Forecasting and Bayesian Data Assimilation. Cambridge University Press, ??? (2015)

[6] Yun, S., Zanetti, R.: Clustering methods for particle filters with gaussian mixture models. IEEE Transactions on Aerospace and Electronic Systems **58**(2), 1109–1118 (2021)

[7] Psiaki, M.L., Schoenberg, J.R., Miller, I.T.: Gaussian sum reapproximation for use in a nonlinear filter. Journal of Guidance, Control, and Dynamics **38**(2), 292–303 (2015)

[8] Psiaki, M.L.: Gaussian mixture nonlinear filtering with resampling for mixand narrowing. IEEE Transactions on Signal Processing **64**(21), 5499–5512 (2016)

[9] Popov, A.A., Zanetti, R.: Ensemble gaussian mixture filtering with particle-localized covariances. In: 2023 26th International Conference on Information Fusion (FUSION), pp. 1–7 (2023). https://doi.org/10.23919/FUSION52260.2023.10224227

[10] Zucchelli, E.M., Jones, B.A.: A Bayesian approach to low-thrust maneuvering spacecraft tracking. Journal of Guidance, Control, and Dynamics **47**(8), 1586–1601 (2024)

[11] Alspach, D., Sorenson, H.: Nonlinear Bayesian estimation using gaussian sum approximations. IEEE transactions on automatic control **17**(4), 439–448 (1972)

[12] Janssen, P., Marron, J.S., Veraverbeke, N., Sarle, W.: Scale measures for bandwidth selection. Journal of Nonparametric Statistics **5**(4), 359–380 (1995) https://doi.org/10.1080/10485259508832654 https://doi.org/10.1080/10485259508832654

[13] Popov, A.A., Zanetti, R.: An adaptive covariance parameterization technique for the ensemble Gaussian mixture filter. arXiv preprint arXiv:2212.10323 (2022)

[14] Anderson, J.L., Anderson, S.L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly weather review **127**(12), 2741–2758 (1999)

[15] Liu, B., Ait-El-Fquih, B., Hoteit, I.: Efficient kernel-based ensemble Gaussian mixture filtering. Monthly Weather Review **144**(2), 781–800 (2016)

[16] Yun, S., Zanetti, R., Jones, B.A.: Kernel-based ensemble Gaussian mixture filtering for orbit determination with sparse data. Advances in Space Research **69**(12), 4179–4197 (2022)

[17] Popov, A.A., Sandu, A.: An explicit probabilistic derivation of inflation in a scalar ensemble kalman filter for finite step, finite ensemble convergence. arXiv preprint arXiv:2003.13162 (2020)

[18] Reifler, B.L., Yun, S., Jones, B.A., Zanetti, R.: Multi-target ensemble gaussian

mixture tracking with sparse observations. In: Proceedings of the 2021 Advanced Maui Optical and Space Surveillance Technologies Conference (2021)

[19] Reifler, B.L., Popov, A.A., Jones, B.A., Zanetti, R.: Large-scale space object tracking in a proliferated LEO scenario. In: Proceedings of the 26th International Conference on Information Fusion (2023)

[20] Gilitschenski, I., Steinbring, J., Hanebeck, U.D., Simandl, M.: Deterministic Dirac mixture approximation of Gaussian mixtures. In: 17th International Conference on Information Fusion (FUSION), pp. 1–7 (2014). IEEE

[21] Hardy, M.: Calculation of the Covariance of Gaussian Mixtures. Mathematics Stack Exchange. https://math.stackexchange.com/q/195984 (version: 2019-09-15) (2019). https://math.stackexchange.com/q/195984

[22] Reich, S.: A nonparametric ensemble transform method for Bayesian inference. SIAM Journal on Scientific Computing **35**(4), 2013–2024 (2013)

[23] Zucchelli, E.M., Jones, B.A.: A gaussian integral filter with multivariate laplace process noise. Proceedings of the 26th International Conference on Information Fusion (2023)

[24] Lawrence, J.D.: A Catalog of Special Plane Curves. Courier Corporation, ??? (2013)

[25] Olver, F.W., Lozier, D.W., Boisvert, R.F., Clark, C.W., Dalhuis, A.: Digital library of mathematical functions. National Institute of Standards and Technology from http://dlmf. nist. gov/(release date 2011-07-01), Washington, DC (2010)

[26] Lorenz, E.N.: Deterministic nonperiodic flow. Journal of atmospheric sciences

20(2), 130–141 (1963)

[27] Strogatz, S.H.: Nonlinear Dynamics and Chaos with Student Solutions Manual: With Applications to Physics, Biology, Chemistry, and Engineering. CRC press, ??? (2018)

[28] Dieci, L., Jolly, M.S., Van Vleck, E.S.: Numerical techniques for approximating lyapunov exponents and their implementation. Journal of Computational and Nonlinear Dynamics **6**(1) (2011)