Interpolation of Non-Gaussian Probability Distributions for Ensemble Visualization

Brad Hollister and Alex Pang

Abstract—A typical assumption is that ensemble data at each spatial location follows a Gaussian distribution. We investigate the consequences of that assumption when distributions are non-Gaussian. A sufficiently acceptable interpolation scheme needs to be devised for the interpolation of non-Gaussian distributions. We present two methods to calculate interpolations between two arbitrary distributions and compare them against two baseline methods. The first method uses a Gaussian Mixture Model (GMM) to represent distributions. The second method is a non-parametric approach that interpolates between quantiles in the cumulative distribution functions. The baseline methods for comparison purposes are: (a) using a Gaussian distribution and interpolating the means and standard deviations, and (b) forming a new distribution based on the interpolation of individual realizations of the ensemble. We show that the two proposed non-Gaussian interpolation methods have the following behavior: the interpolated distributions do not decompose to more constituent Gaussian distributions than the highest modality of those being interpolated, and do not have variances less than the smallest variance from the grid points being interpolated. Finally, we compare these four interpolation methods when used in the analysis of scalar and vector fields of ensemble data sets, particularly in areas where the distribution is non-Gaussian.

Index Terms—Interpolation, non-Gaussian, visualization, flow, uncertainty, ensemble

1 Introduction

A fundamental operation used in most visualization algorithms is interpolation. Interpolation is used in workhorse visualization techniques such as marching cubes, direct volume rendering, and stream line generation, and many other popular algorithms. Performing interpolation is well defined when the data points and the interpolants are single valued, or crisp. However, this is not the case when the data points and the interpolants are multivalued, or consist of a distribution.

With increasing interest in representing uncertainty in modeling and simulation with techniques based on Monte Carlo methods, we are now faced with the challenge of analyzing and visualizing ensemble fields. Ensemble fields are made up of individual realizations, each a possible outcome, of the simulation. Assuming that the ensemble fields are defined over a grid, a popular approach is to treat all the values at a given grid point from different realizations as a multivalue or a distribution. Recent works in this area have primarily assumed that the multivalue follow a Gaussian distribution. Even more recent efforts have tried to remove this assumption. In this paper, we examine two alternative interpolation methods that support non-Gaussian distributions and compare them against two other baseline methods.

There are several reasons for considering a more general representation for multivalue aside from a Gaussian model. The assumption of a normal distribution neglects the possibility that the multivalue represents overlapping sub-populations of data, which by themselves can be considered Gaussian component distributions. These often arise in various situations such as sub-voxel material classification for volume rendering, and ambiguity in resolving fiber orientation during DT-MRI tractography. Often times, it is at these "mixing" regions where interesting things happen e.g. presence or absence of a boundary, crossing or divergence of a path, etc. The distributions at these regions exhibit multimodal profiles. Their consideration requires representation of these distributions as non-Gaussian.

In this paper, we adopt the terms crisp to mean single valued, whereas multivalued is taken to mean a collection of values [11]. The concept of multivalues is general enough to represent (i) the collection of values of a variable at a particular location as reported by different realizations in an ensemble, (ii) a probability distribution of the same set of values represented as a probability density function (PDF) that requires the area under the function to sum to one, (iii) other representations e.g. as a signal. Using the operator based approach for manipulating multivalues [11], linear interpolations can be defined as:

$$M' \;=\; (1-\alpha)M_1 + \alpha M_2 \tag{1}$$

where M', M_1 and M_2 are multivalues, $\alpha \in [0,1]$. Note that $(1-\alpha)$ is a simple subtraction between 2 crisp values. The multiplication of a crisp value and a multivalue simply scales each member of the multivalue and results in a multivalue. On the other hand, the $+$ operator between two multivalues can be defined according to the needs of the application. Using this framework, one can also define and entertain other variations of simple linear interpolations e.g.

$$f(M') \;=\; (1-\alpha)f(M_1) + \alpha f(M_2) \tag{2}$$

where $f(.)$ operates on multivalue M, and $+$ is appropriately defined.

The two interpolation methods examined in this paper define $f(.)$ as: (i) a gaussian mixture model to represent M, and (ii) different quantiles of the PDF representing M. We refer to interpolation using method (i) as *GMM PDF interpolation*, and method (ii) as *Quantile PDF interpolation*. These are described later in Section 4. The two baseline methods used to compare these interpolations are: (i) one that uses a Gaussian representation of M – interpolation is referred to as *Gaussian PDF interpolation*, and (ii) one that uses the raw multivalues – interpolation is referred to as *Ensemble PDF interpolation*.

There are three main considerations in formulating the interpolation methods. Firstly, if additional modes are introduced during interpolation, this would imply that new sub-populations are somehow introduced during the process. While such populations may exist, there is nothing in the data set to suggest this. So, we impose the condition that the interpolation method cannot create additional modalities between known distributions. Secondly, a suitable interpolation method should not produce distributions that have variance less than the smallest variance from the grid points being interpolated between. As a contradiction, suppose that the interpolated distributions did in fact have variances less than those at the grid points. This is undesirable since the interpolated distributions should be less certain than at the observed grid point distributions, and should therefore not have variances that are smaller than those observed at the grid points. Thirdly, the method must naturally produce a total probability of 1.0. While one approach is to normalize the sum of components treated separately, we present more than one possible method that adheres to our specifications and that does not require explicit normalization. Therefore, a good interpolation method should ensure that: (i) no additional modes are introduced during the interpolation, (ii) the variance should not be smaller during interpolation, and (iii) interpolated results are also probability distributions. These are described further in Section 4. Aside from interpolation, the more general problem of curve fitting or data analysis over ensemble fields can also benefit from this work.

After reviewing relevant related works in Section 2, we summarize different interpolation methods that assume the data to have a Gaussian distribution in Section 3. This is followed by detailed descriptions of our two proposed interpolation methods in Section 4. In Section 5, we examine the behavior and relative performance of these two methods against two baseline methods, and use them to analyze an ensemble forecast of an ocean circulation model in Section 5.3.

2 Related Work

A nice overview of statistical techniques for spatial interpolation was presented by Myers [13]. The techniques range from simple linear models with no covariance, to those using spatial structure functions. The survey however does not include non-parametric distribution interpolation. The paper does claim that interpolation is a solution to an inherently ill-posed problem, namely that it is a problem of prediction with limited data. For that, multiple models with different purposes can be employed. A more detailed survey [9], but focusing on geostatistical applications, compare methods according to different criteria such as local vs global support, deterministic vs stochastic, univariate vs multivariate, linear vs nonlinear etc. Among the methods that consider stochastic data, they assume normal distribution.

Within the visualization community, there are also a number of recent publications that address stochastic interpolation. Scheuermann, et al. [28] present a form of Kriging interpolation of spatial data for Gaussian distributions using a parameter-based approach. This technique relies on computing a covariance matrix and that the underlying data be formed from a Gaussian process. Pfaffelmoser et. al [17] visualize isosurfaces via a raycasting scheme, and perform spatial interpolation assuming the data has a Gaussian distribution at each location. Likewise, Pöthkow et. al [21] discuss isocontour visualization of normally distributed data. They interpolated between grid points using the 0th and 1st moments without spatial correlation considerations. Their subsequent work [22] considered the effects of spatial correlation in visualizing isosurfaces using probabilistic marching cubes. An alternative method of looking at global correlation structures in a hierarchical fashion was presented in [18].

When data do not follow a Gaussian distribution, a more general uncertainty model is needed. Liu et. al [10] propose a Gaussian mixture to represent the distribution of voxel values in air temperature data. They perform volume rendering on the data set and interpolate between pairs of a fixed number of Gaussians components along cast rays. In their study, they found that four Gaussian kernels are sufficient for a variety of data sets that they examined. In addition, they support stationary and anisotropic correlations in the process, For non-parametric representations of non-Gaussian distributions, operations on the distributions require different handling. Love, et al. [11] discuss two forms of a non-parametric interpolation method via convolution addition of probability distributions as well as bin-wise addition. Pohl, et al., [20] first transform the (discrete) distribution to Euclidean space via a set of Log Odds operations, where they can then be manipulated using conventional addition and multiplication. Results are then mapped back to probabilistic space via a reversible transform. Read [26] delineates a method to interpolate histograms via quantiles.

Uncertainty in vector fields is of great interest to at least two broad fields: meteorological community and fiber tracking community. Most of the work to date assumes Gaussian random fields. Otto, et al. present analysis of 2D [14] and 3D velocity fields [15] using particle advection, critical points, and segmentation of

field topology. Petz et al. [16] also analyze uncertain velocity fields modeled as Gaussian random fields with spatial correlation.

There is a growing body of work on probabilistic fiber tracking. Unlike velocity fields, the tracks here represent fiber connectivity from one region to another and are obtained by integrating the major eigenvector field. The main source of uncertainty can be attributed to inadequate resolution in the data acquisition stage of diffusion tensor MRI. However, there are numerous other sources as well [3]. While most of the earlier works on probabilistic fiber tracking delved on the inadequacy of the simple tensor representation to show alternative trajectories due to multiple fiber populations within a cell, more recent works are based on high angular resolution diffusion imaging (HARDI) data which makes it is possible to describe fiber orientations using more sophisticated formulations such as spherical harmonics and multi-tensor representations. In a recent paper, Jiao et al. [7] describe a local, icon-based presentation of an ensemble field of fiber orientation distribution functions (ODF). The results of our paper can be used towards spatial analysis of such ensemble fields, for example.

There is much interest in the meteorological community to provide better visualization of forecast data. Slingsby et. al [29], discuss how users interpret and use weather data, specifically hurricane data. Storm path information are examined from historical data. They draw attention to spatial and temporal clustering and its undervalued status among those currently employing such visualization software. Weather forecasts are usually based on an ensemble of predictions. For that, Potter et. al [23] describe a framework for viewing stochastic information from ensembles. This package allows for visualization of spaghetti plotting, etc. of weather data. Zhang, et al. [27] present Noodles, a software package for displaying uncertainty in stream lines and other weather data visualization for ensemble forecasting. Potter et. al [24] describe a software tool to visualize two-dimensional sets of distribution data. It displays a contour of field PDF values and allows for a normed difference between data PDFs and an ansatz selected by the user. More recently, Phadke et. al [19] present two novel visualization methods for ensembles. Primarily, they allow simultaneous viewing of multiple ensemble members. They also present a technique called "Screen Door Tinting" which applies value changes to field points that show differences between ensembles.

From the point of view of users, Martin et. al [12] point out the difficulty of users to identify hurricane directional movement and speed from current data visualization, or directly on vector fields. In a similar study, Broad et. al [4], further emphasize interpretation and usage of complex weather data. They show how a general interpretation of a Gaussian distribution of hurricane direction prediction can lead to inaccurate views on the probability within a "cone of uncertainty". Clearly, if multimodal velocity distribution is calculated with such a broad region of uncertainty using a Gaussian assumption, incorrect estimation of the probability of hurricane direction can occur, most specifically within the general population who can be greatly impacted by such interpretation. A non-Gaussian consideration for vector field visualization together with a redesigned visualization may rectify this issue to a degree. We hope that with the results presented in this paper, we will be able to extend such visualizations to consider non-Gaussian mixing regions.

3 Gaussian Interpolation

In this section, we briefly summarize alternative strategies of performing spatial interpolation for distributions that are assumed to be Gaussian. In this discussion, we consider linear interpolation between two univariate Gaussian distributions. The interpolation parameter α indicates both the parameterized spatial distance and the parametric interpolation distance between the two distributions.

First, it is possible to interpolate Gaussian parameters: the mean, standard deviation (and other moments) independently. The interpolants remain Gaussian and can be reconstructed based on interpolated parameters. This method is simple yet allows for smooth translation of mode and smoothly varying moments as can be seen in figure 1. *Gaussian PDF interpolation* in this paper refers to this variant of Gaussian interpolation.

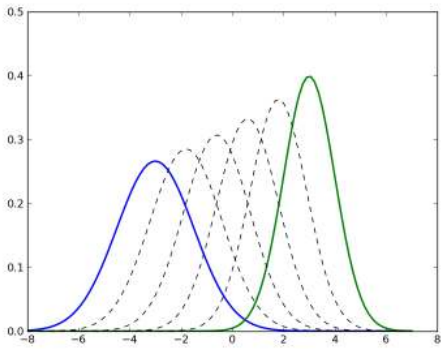
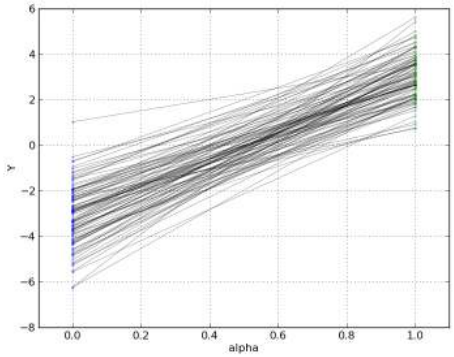


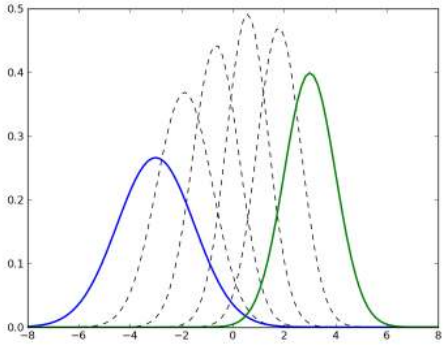
Fig. 1. Intermediate interpolants (black dashed curves) travel from the blue to the green Gaussian curve.

When the distribution is represented by samples rather than by Gaussian parameters, another approach is to interpolate the samples directly rather than fitting it with a Gaussian first. Here, samples drawn from each distribution are interpolated independently. For a random variable B (representing samples drawn from the blue curve), let the random sample $Y_1, Y_2, ..., Y_n$ be n independent and identically distributed (i.i.d.) variables. Similarly, a random sample from G (representing samples drawn from the green curve) are the n i.i.d. variables $Y_{1+n}, Y_{2+n}, ..., Y_{2n}$. The total number of all possible sample interpolants is the count of all possible pairings

between the members of the random samples, i.e. the cardinality of the Cartesian product: $|\{Y_1, Y_2, ..., Y_n\} \times \{Y_{1+n}, Y_{2+n}, ..., Y_{2n}\}|$, for any given $\alpha \in [0, 1]$. This method of PDF interpolation allows translation of mode but variance is potentially less than either the B or G distribution during interpolation. Figure 2 shows an instance of sample pairings between two PDFs and the resulting PDF interpolants. In this example, there are interpolants that have variance less than the distributions being interpolated.



(a) One set of sample pairs drawn independently from the distribution on the left (blue dots) and the distribution on the right (green dots).



(b) Intermediate interpolants (black dashed curves) show smaller variance than end points distributions.

Fig. 2. Sample interpolation for a given instance of distribution sample pairings. (a) Shows pairings and (b) depicts interpolants with dashed lines.

Thirdly, there is "probabilistic interpolation", also referred to as histogram interpolation. This method normalizes the range of the grid point distributions. For each "bin", frequencies are interpolated. With this approach, the PDF at one grid point morphs into the the PDF at the other grid point. In figure 3, the interpolant at $\alpha = 0.5$ is bimodal.

This third method might be suitable for some applications, such as volume rendering materials where a cell might contain multiple materials. That is, when one considers the situation where the populations are predominantly of different types on either side of a boundary, but is made up of both populations at the boundary region, then interpolations that increase the modality of the distributions might be desirable. On the other hand, when one considers the transport or transition of a population or mixture of populations e.g. volume of water at different temperatures, across some distance then we do not want to increase the modality of the interpolant distributions. In this paper, we consider the latter design criterion as we consider interpolation of non-Gaussian distributions.

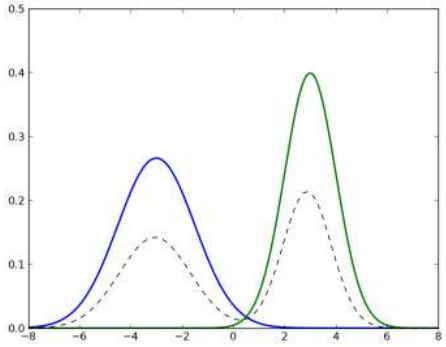


Fig. 3. An interpolant can become multimodal between unimodal distributions as shown by the dashed black interpolant at $\alpha = 0.5$.

4 Non-Gaussian Interpolation

We present two techniques for the linear interpolation of PDFs as represented by a GMM and a non-parametric quantile model. These techniques directly apply to the standard unit reference cell, where each grid point represents a distribution from an ensemble.

4.1 GMM PDF Interpolation

Our first approach is to linearly interpolate Gaussian parameters for a Gaussian Mixture Model (GMM) as outlined in figure 4. The final step may be optional depending on the application, as indicated by the dotted arrow and box. We describe fitting components and interpolating parameters in this section. Gathering samples is implementation specific and is influenced by the data source.

The *fit components* stage from figure 4 requires modeling the samples with Gaussian components. The GMM can be extracted using the Expectation-maximization (EM) algorithm ([1], [2], and [25]) in order to derive a mixture from the starting samples using m Gaussian components. The mixture is denoted as the random variable $V_{\mathbf{g}}$ located at grid point location $\mathbf{g}$, where $\mathbf{g} \subset \{\mathbf{p} | \mathbf{p} \subset \mathbb{R}^n, n \in \mathbb{N}, 0 < n \leq 3\}$. The GMM is determined by a linear combination of Gaussian basis functions Φ_i:

$$V_{\mathbf{g}} = \sum_{i=1}^{m} a_i \Phi_i \tag{3}$$

$$\sum_{i=1}^{m} a_i = 1 \tag{4}$$

$$\Phi_i = \mathcal{N}(\mu_i, \sigma_i^2) \tag{5}$$

In the next stage of the method, *interpolate parameters*, we first determine the how to pair each Gaussian component from different grid point distributions. For the separate grid points $\mathbf{g}_0$ and $\mathbf{g}_1$, whose Euclidean norm $||\mathbf{g}_0 - \mathbf{g}_1|| = 1$, we pair corresponding Φ_i from V_0 and V_1 (located at $\mathbf{g}_0$ and $\mathbf{g}_1$ respectively). The pairing heuristic for Gaussian components between each end point is based on a one-dimensional linear scale. For univariates, in order to minimize interpolation distance between the mean of paired Gaussian components, we allow substeps in which a possible re-pairing ranked by sorted Gaussian means takes place. In the multivariate case, we pair and sort based on the weight of each Gaussian.

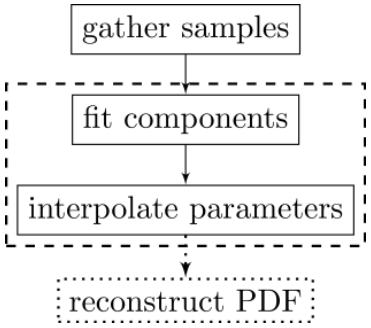


Fig. 4. Gaussian Mixture Model PDF interpolation method. Dashed outline signifies core method stages primarily discussed in paper. Dotted arrow and box signify optional stage.

We calculate α, and the interpolant Gaussian component parameters: $\bar{\mu}_i$, $\bar{\sigma}_i^2$ and their associated weights $\bar{a}_i$ using equations 6 through 9. Another index is used for each component to denote which $V_{\mathbf{g}}$ it is from. Therefore, we have $\mu_{0,i}$, $\sigma_{0,i}^2$ and $a_{0,i}$ from V_0. $\mu_{1,i}$, $\sigma_{1,i}^2$ and $a_{1,i}$ are from V_1.

$$\alpha = ||\mathbf{p} - \mathbf{g}_0|| \tag{6}$$

$$\bar{\mu}_i = (1-\alpha)\mu_{0,i} + \alpha\mu_{1,i} \tag{7}$$

$$\bar{\sigma}_i^2 = (1-\alpha)\sigma_{0,i}^2 + \alpha\sigma_{1,i}^2 \tag{8}$$

$$\bar{a}_i = (1-\alpha)a_{0,i} + \alpha a_{1,i} \tag{9}$$

Thus, our interpolant PDF is $\bar{V}_{\mathbf{p}}$ at location $\mathbf{p}$, defined on a line segment of unit length and with end points $\mathbf{g}_0$ and $\mathbf{g}_1$.

This interpolation method meets our design criteria. Interpolant PDFs will not have greater modality than end point distributions since we require a constant number of Gaussian components to be interpolated. Therefore no additional modes can be present in the interpolants. Linear interpolation of variances from components produce GMM interpolants whose component variances are bounded by those at the end points. Mean interpolation difference is minimized for univariates. Probability interpolation difference between components is minimized for multivariates. The interpolated weights will always sum to one. This is ensured, as long as the total of the weights at every α equal one, as we require. Because EM only returns weights that sum to one, and we only make one-to-one pairings with a fixed and the same number of Gaussian components at each end point, then any number of re-pairings will also have total weights equal to one.

4.2 Quantile PDF Interpolation

The quantile interpolation method overview is shown in figure 5.

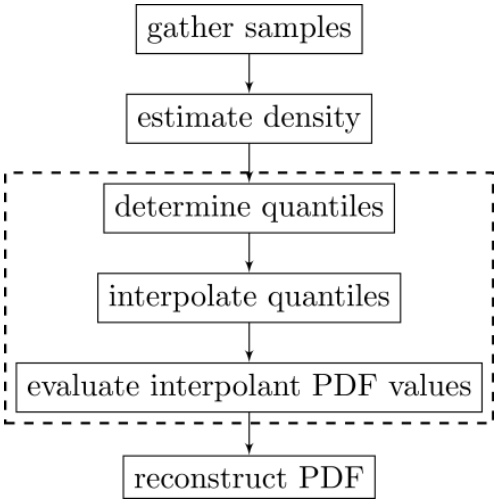


Fig. 5. Quantile PDF interpolation method. Dashed outline signifies core method stages discussed in the paper.

Stages *gather samples* and *estimate density* are implementation specific. We do not cover their implementation details here and the user may choose varying approaches depending on the data. For example, kernel density estimation (KDE) with different window setting techniques can be used for density estimation.

During the *determine quantiles* stage, we compute the random value from the cumulative distribution function (CDF) that will return the desired quantile.

The *interpolate quantiles* phase from figure 5 utilizes a linear interpolation between quantiles q_0 and q_1 of the cumulative density functions (CDF) of V_0 and V_1. This is expressed in equation 10 and uses α from equation 6.

$$\bar{q} = (1-\alpha)q_0 + \alpha q_1 \tag{10}$$

In the *evaluate interpolant PDF values* step, both grid point distributions' quantiles evaluate to the same cumulative density of the interpolant CDF over the sample space variable s:

$$\int_{-\infty}^{\bar{q}} \bar{V}_{\mathbf{p}}(s)ds = \int_{-\infty}^{q_0} V_0(s)ds = \int_{-\infty}^{q_1} V_1(s)ds \qquad (11)$$

Each interpolant probability value for the interpolant's qth quantile can be evaluated using the following expression (see [26] for a complete derivation):

$$\bar{V}_{\mathbf{p}}(\bar{q}) = \frac{V_0(q_0)V_1(q_1)}{(1-\alpha)V_1(q_1) + \alpha V_0(q_0)} \qquad (12)$$

While we can find a unique random value to obtain a desired quantile for univariates, this is not true for the bivariate (or multivariate) case. For the bivariate case, the *determine quantiles* stage requires that we sum over the two-dimensional sample space of the PDF estimate in order to collect (u, v) sample pairs that correspond to the same cumulative density. We do this only at the end points $\mathbf{g}_0$ and $\mathbf{g}_1$. Note that integration of density is performed over a discretized grid and compared within a specified tolerance of the quantile value.

The result of the *determine quantile* step is a set of points that have the same quantile. These points form a curve which we parameterize and refer to as a quantile curve. In the *interpolate quantiles* stage, we take corresponding points (u_0, v_0) and (u_1, v_1) on the curves from $\mathbf{g}_0$ and $\mathbf{g}_1$ respectively and find $(\bar{u}, \bar{v})$ along a line between (u_0, v_0) and (u_1, v_1) depending on α. The resulting interpolant is obtained using equation 13.

$$\bar{V}_{\mathbf{p}}(\bar{u}, \bar{v}) = \frac{V_0(u_0, v_0)V_1(u_1, v_1)}{(1-\alpha)V_1(u_1, v_1) + \alpha V_0(u_0, v_0)} \qquad (13)$$

For the final *reconstruct PDF* step, a reconstruction of the PDF curve or surface is performed using a suitable interpolation such as those available using [8]. For our study, we tessellate the input point set to n-dimensional simplices, and interpolate linearly on each simplex. Unlike the GMM method, PDF modes can only be estimated with a continuous curve or surface. In the case of infinitely many interpolant PDF data points, the surface reconstruction approaches a true PDF.

Interpolant PDFs will not have greater modality than end point distributions. Inflection points on the CDFs will only split and merge corresponding to the modality at the end points. Linear interpolation of the quantiles ensures this. In order for additional modes to form at interpolants, quantiles would have to interpolate to values outside of the range set by the end point PDF quantile values during the interpolation. Since this can not occur using linear interpolation, additional modes do not occur with this method.

Variance of the interpolants for Quantile PDF interpolation is never greater than either end point distributions. The interpolants have quantiles located "between" the end point PDF quantiles in the associated sample space defined by the end point distributions. If the interpolated quantiles were to take on values outside of their bounds set by the end point PDFs, then the variance constraint would be violated. However, linear interpolation does not allow that to happen. It can also be shown that Quantile PDF interpolation is similar to sample based interpolation discussed in section 3. The method interpolates paired samples based on *ordered* samples from both end point PDFs by cumulative density. In this way, no vertical cross-section of the interpolated samples has variance that is less than the least variance from either end point PDF in the interpolation.

5 Results

In the results below, we use four Gaussian components for GMM PDF interpolation as suggested by Liu et al. [10].

5.1 Ground "Truth" Comparison

We examine the behavior of our interpolation methods in figure 6 for a one-dimensional case between two non-Gaussian distributions. Six hundred samples are used to form a fixed-width kernel density estimate (FKDE [6]) at each end point. Our ground "truth" is derived from a linear interpolation of realizations. We then form a non-parametric distribution of each ensemble member interpolant using FKDE.

Figure 6 qualitatively shows that both quantile and GMM PDF interpolations are quite similar to our ground truth ensemble PDF interpolation. On the other hand, the simple Gaussian PDF interpolation shows marked difference from our ground truth. To obtain a more quantitative measure, we calculate the symmetric Kullback-Leibler (SKL) divergence which gives us a measure of dissimilarity between two distributions. Equation 14 is the SKL between probability distributions P and Q.

$$D_{\mathrm{SKL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i) + \sum_i \ln\left(\frac{Q(i)}{P(i)}\right) Q(i) \qquad (14)$$

SKL is computed from $\alpha = 0.0$ to $\alpha = 1.0$ for each PDF interpolation method. For each method, we compute and average 100 such SKL comparisons to remove measurement noise due to sampling and EM fitting. Because the SKL results for Gaussian interpolants are an order of magnitude greater than both GMM and Quantile PDF interpolants, we show the Gaussian SKL measurements separately. In figure 7, we can easily see that Quantile interpolants (blue line) have the least SKL values, while both GMM (red line) and Gaussian (purple line) have larger entropies. The color scheme used for each PDF interpolation method in figures 6 and 7 are used for the remainder of this paper.

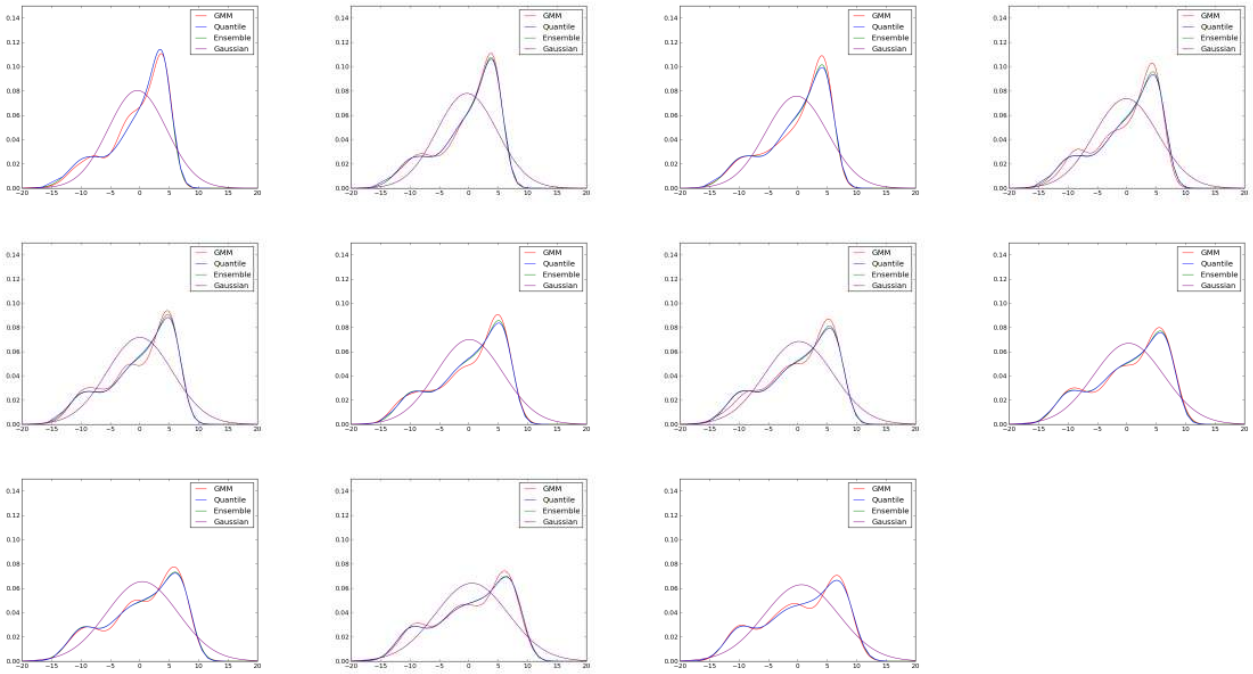
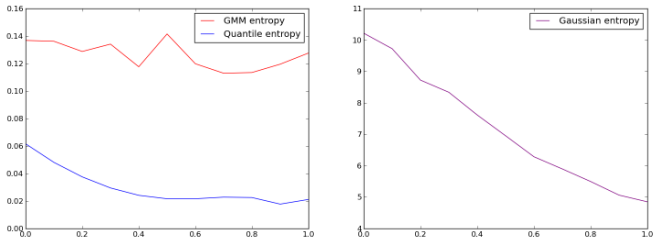


Fig. 6. Univariate PDF interpolant from α=0.0 to α=1.0: GMM (red), Quantile (blue), Ensemble (green) and Gaussian (purple).

Entropy at $\alpha = 0.0$ and $\alpha = 1.0$ are due entirely to the accuracy of the estimation and are not due to any of the interpolation methods. For intermediate α values, the SKL entropy is a combination of the entropy due to estimation errors and the entropy due to difference between the ensemble interpolant and the GMM, Quantile or Gaussian interpolant. Unfortunately, since density estimate and fitting of Gaussian components are needed to form the distributions at the end points, and we do not know how the estimation or fitting error varies as a function of α, we cannot distinguish between entropy due interpolation and those due to estimation or fitting.

Interestingly, as can be seen in figure 7 (b) for Gaussian interpolants, entropy at $\alpha = 1.0$ is less than any intermediate α. Quantile PDF interpolants are almost identical with ensemble interpolants and entropy is greatest at $\alpha = 0.0$ where estimation entropy is larger than for any interpolants. Quantile PDF interpolation effectively orders the samples by their cumulative probability. This corresponds closely with ensemble physical simulations per ensemble member.

Figure 8 shows a linear interpolation between two bivariate distributions. At the top of the figure, we have a bimodal distribution and at the bottom of the figure, we have a unimodal distribution. Some tears on the interpolant PDF can be observed in column (b) due to insufficient data samples.

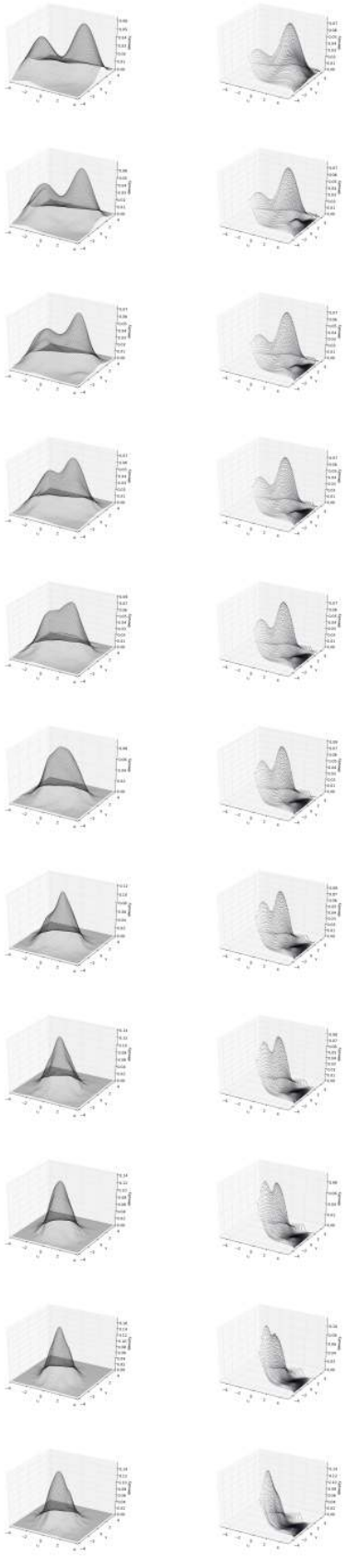


(a) GMM and Quantile SKL (b) Gaussian SKL

Fig. 7. Ten measurements of the SKL divergence for univariate interpolants from α=0.0 to α=1.0. Values are averaged from 100 independent comparisons. Entropy is shown on vertical axes and α on horizontal axes.

5.2 Synthetic Data

For covariant random variables, we describe interpolation in a synthetic velocity field where the velocity components are the bivariate random variables under consideration. In order to show the effect of considering a bivariate bimodal distribution when advecting in a velocity vector field, we construct a toy example consisting of a 3 x 3 grid where all grid points are defined as unimodal except the center grid point, which is defined by a bimodal distribution. Our mean parameter(s) for the velocity PDFs are the mean velocity vector $\mu_i = (u, v)^T$, where u and v are the velocity components aligned with the Cartesian x-y coordinate system. The left half and the top center of the grid is defined by a normal bivariate. Spherical covariance matrices are used, i.e. the covariance matrix designation is a multiple of the identity matrix.



(a) (b)

Fig. 8. One-dimensional PDF interpolation using (a) GMM and (b) Quantile from a bimodal bivariate (α = 0.0) at the top to a unimodal bivariate (α = 1.0) at the bottom.

$$\mathcal{N}_1(\mu_1, \Sigma_1), \mu_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{15}$$

The right side of the grid is defined by:

$$\mathcal{N}_2(\mu_2, \Sigma_2), \mu_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{16}$$

And, the center grid point is the Gaussian mixture of the following two bivariate normals where the first is weighted 0.6 and the second is weighted 0.4:

$$\mathcal{N}_3(\mu_3, \Sigma_3), \mu_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{17}$$

$$\mathcal{N}_4(\mu_4, \Sigma_4), \mu_4 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix} \tag{18}$$

We show the results of interpolating between a bimodal and a unimodal bivariate distribution in figure 8. The Quantile interpolants can be seen to have more pronounced modal separation. There are two discernible modes in all Quantile interpolants while the GMM interpolants are smoother and most lack multimodality. One noticeable artifact with the Quantile interpolants are "missing " lower quantiles. See section 6 for more details.

For visualizing uncertain vector fields, particularly where the distributions are non-Gaussian and more specifically multimodal, and therefore presenting multiple possible trajectories, we propose the use of *modal curves*. While spaghetti plots show bundles or clusters of (possibly intersecting) streamlines, we want modal curves to be parsimonious representations of the major trajectories of the flow, where major is taken to mean the top b most likely directions. That is, we allow modal curves to bifurcate, if along its path, the curve encounters a distribution that is significantly multimodal. To construct modal curves, we seed and advect massless particles much like conventional stream lines but using the interpolated PDF to make decisions. That is, we advect using the velocity corresponding to the highest peak of a bivariate (for 2D) distribution. Modal curves are allowed to bifurcate along PDF modes after a minimum number of advection steps. Advection is performed as usual, using the fourth-order Runge-Kutta method. Each branch is a separate traditional stream line in the sense that branches are seeded at the branch point and advected forward or backward in the velocity field using the same direction as the parent branch. In order to reduce clutter, we remove branches according to criteria outlined in algorithm 1. Figure 9 shows results using $b = 2$.

We prune branches that cross over one another with one exception. Modal curves do not prune themselves at crossings that occur between "root" curves. Up to two "root" modal flow curves may advect from the seed point in either forward or backward integration. Both

while *not at end of the branching modal flow curve list* **do**
advect current branch by taking vector from distribution that forms smallest angle between itself and previous velocity taken by current branch;
if *new advection position crosses branch that is older and it is not the root* **then**
mark current branch and all of its descendents for removal;
continue;
else
mark modal flow curve that was crossed by current modal flow curve and all of its descendents for removal
end
if *current modal flow curve's position prior to its own advection has encountered an interpolated multimodal distribution and its minimum number of advection steps have been reached for another bifurcation* **then**
create and advect new modal flow curve along remaining highest probable velocity and add new branch to list;
if *new advection position of new branch crosses another modal flow curve* **then**
remove new branch modal flow curve from list;
end
process modal flow curve branches marked for removal

Algorithm 1: Advection for modal flow curves

will be of the same age, i.e. have the same total advection steps at the end of an update cycle.

Pruning is performed to disallow ambiguation of primary flow paths and to keep computation to a minimum while allowing "feeler" breadth-search paths earlier in advection which can then be discontinued. Thus, we allow for the greatest divergence of advections along modes in PDF interpolants.

The GMM modal curves shown in figure 9 (top) contain only two branched forward advected curves, while for the Quantile modal curves in figure 9 (bottom), there are three branches, two root branches and a third child branch. Through monitoring intermediate advections, it was noted that all child branches encountered intersections and were subsequently pruned for the GMM advection. This can be explained by considering the entropy inherent in the GMM PDF interpolation method. GMM based modal curves tend to have more "noise" associated with their paths due to variations in Gaussian component parameter fitting (EM) at grid point PDFs. Thus, modal curves branching between maximal divergent branches (such as those shown in figure 9 (bottom)) often are completely pruned. In the toy example, the Quantile PDF interpolation method when applied, preserved one of the child branches and was not pruned because its path did not coincide with the rightmost root curve. Depiction of the most divergent flow paths are still observed in both methods, however.

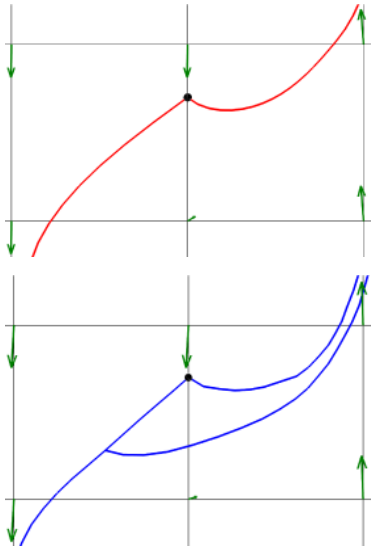


Fig. 9. Toy example modal curves for (top) GMM and (bottom) Quantile PDF interpolation. Black dot denotes seed point. Mean vector is shown at grid points.

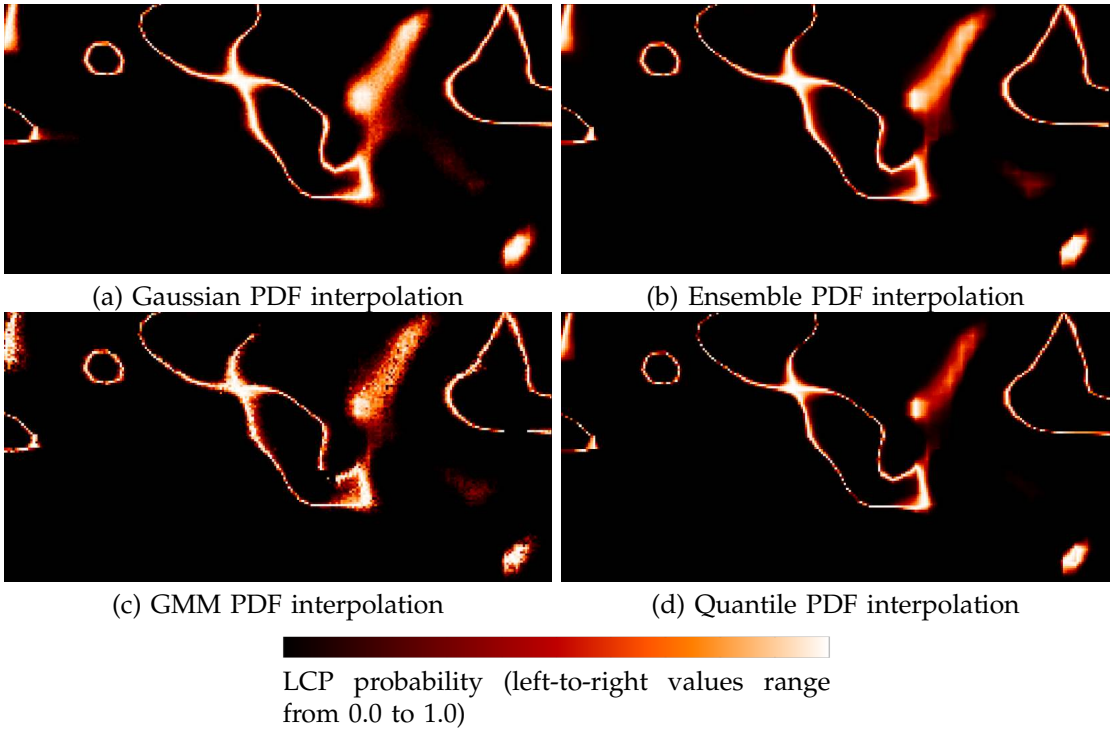
5.3 Simulation Data

Next, we provide verification of the interpolation methods and consideration of non-Gaussianity using simulation data. Our ensemble data-set covers a region of the Massachusetts Bay on the east coast of the United States of America [11] and is provided by Dr. Lermusiaux from MIT. The Massachusetts Bay volume in the study was divided into 53 x 90 grid with 16 depths. The depths at these 53 x 90 grid points vary significantly: depths as shallow as 90 meters and as deep as 196 meters were recorded. We use level zero, or the shallowest depth level in the ensemble and created visualizations using the temperature and velocity fields only.

The results of the GMM and Quantile PDF interpolation methods are shown for the level crossing probability (LCP) [21] at 35 degrees Fahrenheit (figure 10), using equation 20 in a mostly non-Gaussian region of the temperature field. Figure 11 shows the Shapiro-Wilk p-values for normality in the region where LCP is interpolated. Higher p-values of the Shapiro-Wilk test denote greater likelihood of a normal distribution. This region represents the lowest Gaussianity measured for the univariate temperature distributions at level zero of the ensemble data.

Quantile interpolated LCP matches closely with the Ensemble interpolated LCP. GMM interpolated LCP contains the most noise of all the interpolation methods and its probabilistic level set is also the most diffuse. The interpolated Gaussian assumption and the GMM interpolated LCP resemble each other more closely than do the Quantile and Ensemble interpolants.

We use equations 19 and 20 to calculate the LCP. Point $\mathbf{p}$ is a spatial location in the field, θ is the isovalue and $V_{\mathbf{p}}$ is a random variable at location $\mathbf{p}$. $V_{\mathbf{p}}$ is the interpolated temperature distribution at $\mathbf{p}$. Equation 20 is determined



(a) Gaussian PDF interpolation (b) Ensemble PDF interpolation

(c) GMM PDF interpolation (d) Quantile PDF interpolation

Fig. 10. LCP using (a) Gaussian, (b) Ensemble, (c) GMM and (d) Quantile PDF interpolation methods.

by considering whether the cumulative probability at the isovalue for the interpolated PDF is 0.5 at location $\mathbf{p}$. This formulation can be derived from [21].

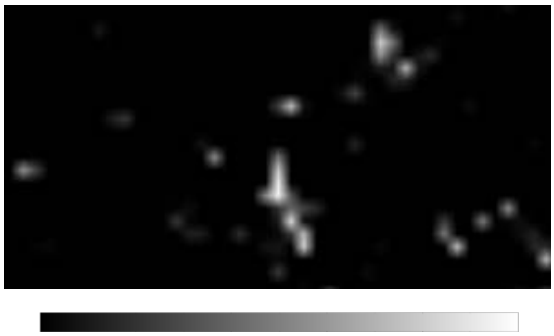


Fig. 11. Temperature field Gaussianity as measured with Shapiro-Wilk test for normality. Shapiro-Wilk test produce p-values that range from 0.0 to 1.0. Higher p-values (white) denote greater likelihood of a normal distribution.

$$F_{\mathbf{p}}(\theta) = \int_{-\infty}^{\theta} V_{\mathbf{p}}(s)ds \tag{19}$$

$$\mathrm{LCP}_{\mathbf{p}} = 1 - F_{\mathbf{p}}(\theta)^4 - (1 - F_{\mathbf{p}}(\theta))^4 \tag{20}$$

Next, we examine the modal curves using all four methods and compare against the spaghetti plots in figure 13. The Gaussian modal curves (purple) tend to follow the primary bundle of the spaghetti plots but do not branch because of the single mode. The ensemble modal curves (green) show similar behavior but with branching. Similarly, GMM (red) and Quantile (blue) modal curves bifurcate, but miss some of the stream line bundles of the spaghetti plots. The Quantile PDF interpolant modal curves have the closest paths in the rightmost part of the plot and GMM has a closer correspondence with the ensemble modal curves with its leftmost branches. There are two primary coherent bundles at the leftmost region of the spaghetti plots, where Quantile modal curves depict one bundle and GMM the other. Small variations in locality of the advections place both sets of modal curves closer to either stream line cluster and local modes dominate directional flow.

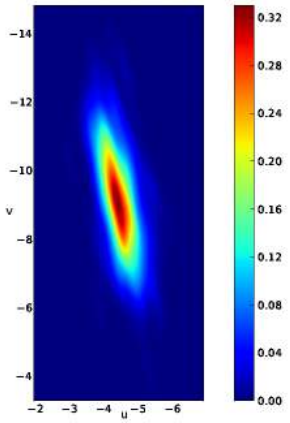


Fig. 12. Representative non-Gaussian grid point (p-value = 4.6×10^{-4})

Note that the bivariate velocity Gaussianity is very low in our dataset, where a typical example of a grid

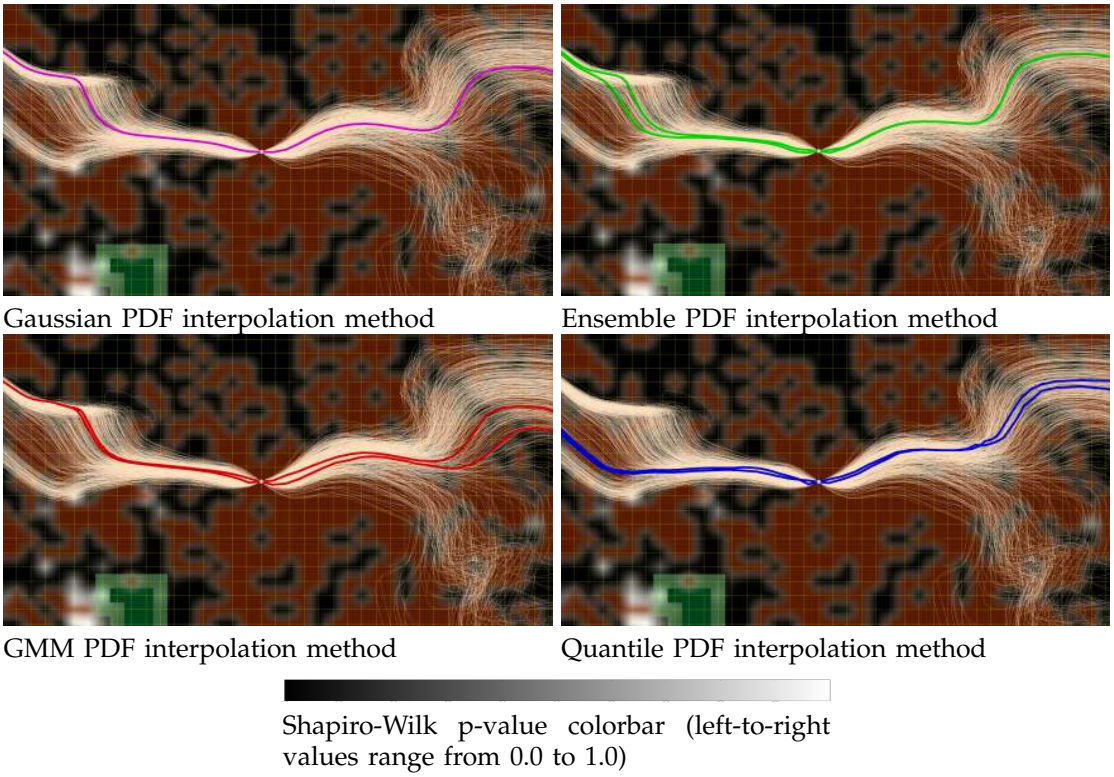


Fig. 13. Modal curves produced using (a) Gaussian, (b) Ensemble, (c) GMM and (d) Quantile PDF interpolation methods. White curves are spaghetti plots of stream lines. The greenish background represents land. The brownish-red background denotes bivariate multimodality greater than one. The black-gray-white background shows the p-values from the Shapiro-Wilk test (e), where higher p-values denote greater likelihood of a normal distribution. Most of the distributions in this region are multimodal non-Gaussian distributions.

point distribution having relatively low variance along the direction of the minor eigenvector of its covariance matrix as compared to the major eigenvector direction (see figure 12). Also note that non-Gaussianity alone is not sufficient for deciding whether modal curves should bifurcate or not. We also need a test for multimodality. We achieve this based on size and separation of peaks. If one considers multimodal marginal distributions individually, it is possible to generate samples that do not belong in the original bivariate distribution. Hence, it is important to consider the bivariate distribution itself rather than its marginals.

In figure 13, p-values are displayed for the Shapiro-Wilk test for Gaussianity along with modality from a Gaussian radial basis function (RBF) estimation. Each PDF has a set M of fitted Gaussian mean parameters. We calculate the greatest difference between any two Gaussian component means as a measure of multimodality. This is defined as follows: let $R = M \times M$, $r \in R$. Then, $D = \{d|d :=$ Euclidean distance between each member of r, $\forall r\}$.

For all two-dimensional ensemble velocity values at a grid point, there are values: $u_{min}, v_{min}, u_{max}$ and v_{max} that represent the minima and maxima of the velocity components. Let the velocity sample extent γ, be defined as in equation 21.

$$\gamma = \parallel (|u_{max} - u_{min}|, |v_{max} - v_{min}|)^{\mathrm{T}} \parallel \tag{21}$$

Multimodality of PDF at a grid point is considered to be $true$ or $false$ depending on the following condition in equation 22, where our weighting factor is 0.10. This is a heuristic that ensures adequate separation of Gaussian components in the mixture.

$$multimodal = \begin{cases} true & \text{if } \max D > 0.10\gamma \\ false & \text{if } \max D \leq 0.10\gamma \end{cases} \tag{22}$$

The modal curves use only local ensemble information (PDF modes) for advection. Thus, they do not always bifurcate along bundles of ensemble stream lines. Figure 14 shows good separation along ensemble stream line bundles but was only reproducible with GMM PDF interpolation (likely due to over-smoothing of multimodality from density estimation with bivariates).

We can also observe that modal curves do not always align themselves with regions of higher density of spaghetti plots. One of the contributing factors, if not the main contributing factor, is because we do not account for spatial covariance in our PDF interpolation. Stream lines in spaghetti plots are created from individual realizations where neighboring velocity information

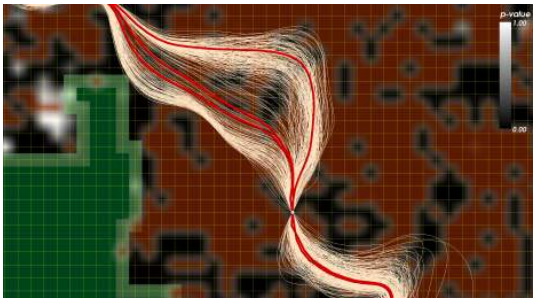


Fig. 14. GMM modal curve exhibiting bifurcation with ensemble spaghetti plots.

is available. The corresponding (i.e. pairing of) velocity information is lost in the PDF representation of the ensemble.

6 Discussion

Based on our limited investigation, Quantile PDF interpolation is the method of choice for the case of univariate interpolation of non-Gaussian distributions since it provides the best SKL score when compared to the ensemble PDF interpolants as baseline.

Both GMM and Quantile PDF methods rely on having a good density estimate either through EM or FKDE. However, Quantile PDF interpolation is particularly susceptible to the "curse of dimensionality" as one goes from univariate to multivariate interpolation. More data is needed to estimate the density. In our study, we use six hundred realizations for interpolating both univariate and bivariate joint distributions. Since PDF surface accuracy is proportional to the number of realizations, sample aliasing at lower frequencies may cause excess smoothing and can obscure modality. Aside from FKDE, there are other estimation methods such as adaptive kernel or projection pursuit density estimation [6] that can yield potentially better results with a limited number of samples for multivariates.

Limited samples also have adverse consequences during the integration stage for finding quantiles, where the sample space resolution needs to be increased in order to detect finer gradations of density per unit sample area. The complexity is proportional to n^d, where d is the dimension of the joint probability and n is the resolution of the sample space. Larger sample spacing can degrade high frequency probability surface detail. Such loss of detail may cause tearing in the reconstructed PDF because of incomplete quantile information during surface interpolation as can be seen in figure 8. This is not seen for univariates in our study but has been encountered for bivariates.

In contrast, because GMM will fit a given number of Gaussians to the data, GMM PDF interpolation is less susceptible to over-smoothing of the density estimate due to lack of data. Hence it can detect modality (up to the number of Gaussian components) better than Quantile PDF interpolation, but at the cost of accuracy associated with RBF. Another consideration is that the GMM at each grid point can be performed in a preprocessing step and its interpolation will outperform Quantile PDF interpolation in terms of fewer computations required per interpolant.

The interpolation methods presented in this paper do not account for spatial covariance with surrounding grid point distributions. With GMM, we dismiss PDF-wide summary parameters that simplify covariance measurements and as a consequence we do not currently have heuristics for paired Gaussian component covariance. In the quantile case, we are interpolating unique surface values of individual PDFs which do not relate as a whole to surrounding PDFs when considered in isolation.

From our example of a two-dimensional univariate PDF interpolation, we used LCP to visualize a probabilistic temperature field. Since LCP is determined based on the CDF, we can apply it directly to non-Gaussian fields.

7 Conclusion

This paper investigated two PDF interpolation methods for both univariate and bivariate non-Gaussian distributions, in one and two dimensional space, and compared them against two baseline methods. The fundamental problem with PDF interpolation is that there is no unique path or set of intermediate interpolations between PDFs (especially in the more general case of non-Gaussian distributions). Our methods assume no prior knowledge of the ensemble data, in order to be more broadly applicable.

The interpolation methods presented in this paper are designed to have certain properties: variance should be bounded by the variances at grid points, no additional modes are introduced during interpolation, and the interpolants are PDFs. Using LCP and modal flow curves, we compared the results of the 4 interpolation methods on random fields exhibiting non-Gaussian distributions and their effects on the visualizations.

The Quantile PDF interpolation appears to offer the best fitting interpolants relative to the ensemble. However, it suffers from the "curse of dimensionality". Improvements to this method can come in the form of alternative ways to estimate density e.g. projection based methods that can capture multimodality with smaller sample sets. Hybrid methods that take advantage of both GMM and Quantile interpolation is also another area to be explored. We currently do not include spatial covariance in PDF interpolation, and is another area of further investigation. Also, while we started out focusing on non-Gaussian distributions, the modality of the distribution is perhaps more significant particularly. In the results presented here, we used an ad-hoc method for testing the modality of a distribution. There are more formal multimodality tests that can be incorporated in the future [5].

Ensembles, when considered as a random field of (simulation) measurements, instead of merely disparate

parallel field data, offers promise for a much better insight into the nature of the ensemble when all members are visualized as their aggregate. Using interpolation on the grid point PDF directly provides a method for using the results of ensemble data in this more consolidated view. Additionally, if ensemble data can be stored as random field data exclusively, with better insight into the ensemble information, this approach may prove more viable than conventional methods (spaghetti plots for example) which are in large use today. Finally, the results presented in this paper is but the first step in analyzing and visualizing uncertainty in random fields.

REFERENCES

[1] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[2] Jeff Blimes. A gentle tutorial for the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 1998. http://crow.ee.washington.edu/people/bulyko/papers/em.pdf.

[3] R. Brecheisen. *Visualization of Uncertainty in Fiber Tracking Based on Diffusion Tensor Imaging*. PhD thesis, Technische Universiteit Eindhoven, 2012. Department of Biomedical Engineering.

[4] K. Broad, J. Leiserowitz, J. Weinkle, and M. Steketee. Misinterpretations of the cone of uncertainty in Florida during the 2004 hurricane season. *American Meterological Society*, pages 651–667, May 2007.

[5] N.I. Fischer, E. Mammen, and J.S. Marron. Testing for multimodality. *Computational Statistics & Data Analysis*, 18(5):499 – 512, 1994.

[6] Jeng-Neng Hwang, Shyh-Rong Lay, and Alan Lippman. Nonparametric multivariate density estimation: A comparative study. *IEEE Transactions on Signal Processing*, 42(10):2795–2810, 1994.

[7] F. Jiao, J.M. Phillips, Y. Gur, and C.R. Johnson. Uncertainty visualization in HARDI based on ensembles of ODFs. In *Proceedings of the 5th IEEE Pacific Visualization Symposium (PacificVis 2012)*, pages 193–200, February 2012.

[8] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[9] Jin Li and Andrew Heap. A review of spatial interpolation methods for environmental scientists. Technical Report GeoCat 68229, Geoscience Australia, 2008.

[10] Shusen Liu, Joshua A. Levine, Peer-Timo Bremer, and Valerio Pascucci. Gaussian mixture model based volume visualization. *IEEE Symposium on Large Data Analysis and Visualization*, pages 73–77, 2012.

[11] Alison Love, David L. Kao, and Alex Pang. Visualizing spatial multivalue data. *IEEE Computer Graphics and Applications*, 25(3):69–79, May/June 2005.

[12] J. Martin, E. Swan II, R. Moorehead, Z. Liu, and S. Cai. Results of a user study on 2D hurricane visualization. *Eurographics/IEEE Symposium on Visualization*, 27(3):991–998, 2008.

[13] Donald Myers. Spatial interpolation: An overview. *Geoderma*, pages 17–28, 1994.

[14] M. Otto, T. Germer, and H. Theisel. Uncertain 2D vector field topology. *Computer Graphics Forum (Proceedings of Euro-graphics 2010, Norrkpping, Sweden)*, 29(2):347–356, 2010.

[15] M. Otto, T. Germer, and H. Theisel. Uncertain topology of 3D vector fields. *Pacific Visualization Symposium (PacificVis)*, pages 67–74, 2011.

[16] Christopher Petz, Kai Pöthkow, and Hans-Christian Hege. Probabilistic local features in uncertain vector fields with spatial correlation. *Computer Graphics Forum*, 31(3):1045–1054, 2012.

[17] T. Pfaffelmoser, M. Reitinger, and R. Westermann. Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. *Eurographics/IEEE Symposium on Visualization (EuroVis 2011)*, 30(3):951–960, 2011.

[18] T. Pfaffelmoser and R. Westermann. Visualization of global correlation structures in uncertain 2D scalar fields. *Computer Graphics Forum*, 31:1025–1034, 2012. doi: 10.1111/j.1467-8659.2012.03095.x.

[19] M. Phadke, L. Pinto, O. Alabi, J. Harter, R. Taylor, X. Wu, H. Petersen, S. Bass, and C. Healy. Exploring ensemble visualization. *VDA*, pages 82940B–82940B–12, 2012.

[20] Kilian M. Pohl, John Fisher, Sylvain Bouix, Martha Shenton, Robert W. McCarley, W. Eric Grimson, Ron Kikinis, and William M. Wells. Using the logarithm of odds to define a vector space on probabilistic atlases. *Medical Image Analysis*, 11:465–477, 2007.

[21] K. Pöthkow and Hege H. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1393–1406, October 2011.

[22] Kai Pöthkow, Britta Weber, and Hans-Christian Hege. Probabilistic marching cubes. In *Proceedings of the 13th Eurographics / IEEE - VGTC conference on Visualization*, EuroVis'11, pages 931–940, Aire-la-Ville, Switzerland, Switzerland, 2011. Eurographics Association.

[23] K. Potter, A. Wilson, P. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. Johnson. Ensemble-Vis: A framework for the statistical visualization of ensemble data. In *IEEE Workshop on Knowledge Discovery from Climate Data: Prediction, Extremes.*, pages 233–240, 2009.

[24] Kristin Potter, Robert M. Kirby, Dongbin Xiu, and Chris R. Johnson. Interactive visualization of probability and cumulative density function. *International Journal for Uncertainty Quantification*, 2(4):397 – 412, 2012.

[25] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[26] A.L. Read. Linear interpolation of histograms. *Nuclear Instruments and Methods in Physics Research*, pages 357–360, 1999.

[27] Jibonananda Sanyal, Song Zhang, Jamie Dyer, Andrew Mercer, Philip Amburn, and Robert Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1421–1430, 2010.

[28] S. Schlegel, N. Korn, and G. Scheuermann. On the interpolation of data with normally distributed uncertainty for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2305 – 2314, December 2012.

[29] A. Slingsby, J. Strachan, P. Vidale, and J. Dykes. Discovery exhibition: Making hurricane track data accessible. http://www.discoveryexhibition.org/uploads/Entries/Slingsby_2010_HurricaneTrackData.pdf, 2010. Discovery Exhibition.