Comparative Visual Analysis of Vector Field Ensembles: Appendix

Mihaela Jarema Ismail Demir Johannes Kehrer Rüdiger Westermann

Technische Universität München

1 Introduction

A1. Modelling Directional Data

We model directional data, i.e., we estimate a pdf, using parametric mixture models. Another approach to approximate densities is non-parametric, where a pdf is estimated directly from the observations. The histogram is the most prevalent density estimator, while a more accurate method is the kernel density estimate [1]. A histogram is simple to construct, but suffers from major drawbacks: It involves an arbitrary choice for the starting point and group boundaries, which can potentially distort the presented information. Moreover, just like the smoother nonparametric density estimators, it requires the amount of smoothing of the estimate – the bin width – as input. Depending on the smoothing parameter, the shape of the density estimate can vary significantly. A very small amount can exhibit spurious fine structures, the number of "bumps" decreasing as the amount of smoothing increases, to the extent that the multimodality of a pdf can be obscured (cf. Fig. 1(a) and (b)). Furthermore, deriving additional quantitative information automatically, such as the modes of the pdf, is not trivial. Even if the modes can be assessed visually by the number of "bumps" or the local peaks of the density, the number of significant peaks, their locations and sizes still need to be determined. Because we use the modes not only to visualize the modality, but also to find the degree to which ensemble members agree in their behavior, we preferred a parametric model to a non-parametric one; parametric mixture models provide not only the approximated pdfs, but also the modes and their characteristics (cf. Fig. 1(c)).

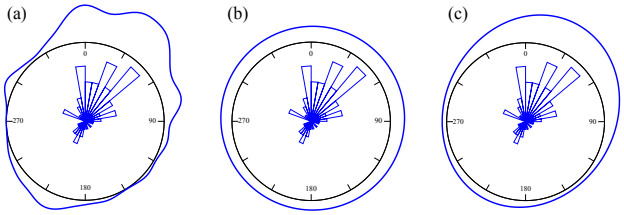

Figure 1: Density approximations (and rose diagrams - a type of angular histograms) for a bimodal set of directions [2]: (a)-(b) kernel density estimates with different amounts of smoothing, where (a) undersmoothing leads to the appearance of several minor modes, while (b) oversmoothing obscures relevant structures; (c) Gaussian mixture model.

Popular parametric models for circular data are the Wrapped Normal (WN) and the von Mises (vM) distributions [2]. Initially we fitted mixtures of both vM and Normal distributions, and verified the extent to which the modalities of the mixtures repeated at the grid points. For the vM mixtures, we followed a stepwise procedure (testing for n components against more than n) [2] to determine the number of components in the mixture and used a recent Expectation-Maximization (EM) algorithm [3] to estimate their parameters. For the Gaussian mixtures, we used the fact that WNs can be obtained by wrapping Normal distributions around the circle, which enabled us to apply for Gaussian Mixture Models (GMMs) an EM algorithm that has been shown to perform well at finding the correct number of spherical Gaussian mixture components (see the next section for details).

Mixture models suffer the disadvantage that, depending on the initial values of the parameters, the solution may converge to a local optimum and yield different structures. To alleviate this problem, we repeat the fitting process several times with different starting values and choose the best solution. The solutions are assessed via silhouettes [4] – a validation technique suitable for algorithms that have random initial guesses, where the performed analysis depends only on the resulting partitions and not on the algorithm that generated them. As the best solution, we select the one with the highest average silhouette from those whose modalities are obtained most often during the repeated fittings.

We tested the two mixture models on the data sets presented in Section 7 (Results) of the paper, applying the algorithms (including the repetitions) ten times at every grid point of the domain. For each of the runs we derived a scalar field containing the modalities (number of components) of the fitted mixtures and analyzed for both procedures the standard deviations of these fields. While a detailed comparison of the two algorithms is beyond the scope of this appendix, we noticed that, for the data sets under test, the GMM yielded much more consistent results. Namely, the number of modes repeated consistently at approximately 98% of the total number of grid points, as opposed to 50% for the vM mixtures. Given the small number of non-repetitive modality instances and that users can also use the spatial neighborhood of a grid point to infer on a possible under- or overestimated number of modes, we decided to use, without loss of generality, the Gaussian mixtures in the rest of our analyses.

Modelling Data using GMMs

Circular data consists of observations that can be regarded either as unit vectors in the plane or as points on the unit circle. Choosing an initial direction (the x-axis) and an orientation (counterclockwise) allows specifying the observations by the angle from the axis to the point on the unit circle. We model the angular data using GMMs [5]. For a sample $\theta_1,...,\theta_n$, the pdf is

$$f(\theta) = \sum_{i=1}^{N} \alpha_i \mathcal{N}\left(\mu_i, \sigma_i^2\right)(\theta), \quad \alpha_i > 0, \quad \sum_{i=1}^{N} \alpha_i = 1, \qquad (1)$$

where N is the number of Gaussian components parameterized by the mean vectors μ_i and variances σ_i^2, and α_i are the weights of the components. For a given N, the $3N$ parameters $\{\alpha_i, \mu_i, \sigma_i^2\}$ can be estimated using the EM algorithm. The true number of mixture components is, however, unknown and must be inferred. Despite a large number of publications on the topic, there is no optimal solution to this problem. Typical methods start with a large number of components and merge similar components (cf. Hennig [6] for a survey), or perform fittings with an increasing number of components up to a certain threshold and then use different criteria to select an optimal number (cf. Oliveira-Brochado and Martins [7] for an overview).

We determine the modality of a pdf and its structure automatically, without a priori knowledge on the number of modes. To this purpose, we first perform a statistical test of randomness, so that we do not attempt to find clusters when the data is distributed uniformly. Uniform distributions may occur, for instance, around criti-

Accepted for publication in Proc. IEEE Conf. Visual Analytics Science and Technology (VAST 2015).

cal points, where the corresponding vectors are zero. While we did not have any instances of zero vectors emerging exactly at the grid points in our data sets, since no directions can be defined at such vectors, these observations should be modelled by components of their own and not be considered further when fitting mixtures. We use the omnibus test [2] to test against any non-uniform alternative model. If the hypothesis of randomness can be rejected at the 5% significance level, we follow a procedure similar to the one proposed by Hamerly and Elkan [8], which has been shown to work well to determine the correct number of modes. The k-means algorithm can be applied in a recursive manner until the data assigned to each of the k centers can be modeled by a single Gaussian component. Instead of the k-means, however, we use the less restrictive EM algorithm, which allows a different full covariance matrix for every component rather than assume that data points are distributed spherically around the k-means centers, and fit a GMM to the given data set.

Specifically, we start with one component and test whether it can be modeled by a single Gaussian distribution. If this is not the case, we run the EM algorithm to fit two Gaussians to the data set and repeat this procedure recursively, until no component needs to be split anymore (either because it is well approximated by a Gaussian or its cardinality is too small). Finally, we use the identified components as initial conditions to the EM algorithm and run it on the whole data set to refine the solution.

For splitting, we run a Lilliefors test for smaller samples (cardinality below 25), and the Anderson-Darling test (a very powerful normality test, also used in [8]) otherwise to test whether the null hypothesis that the sample comes from a normal distribution can be rejected at the 1% significance level. To address the problem of multiple comparisons, which appears when multiple hypotheses are tested on a single set of data, we use the Bonferroni correction. Performing n tests on a single set of data increases the likelihood that the null hypothesis is rejected due to chance even if it is true. Testing each of the n tests at a significance level $1/n$ times as low as that for testing only one hypothesis reduces the chances of obtaining false positives. As the cardinality of our ensembles is around 50 and typically just a few splittings are performed, we use a significance level of 0.1%.

Note that GMMs cannot be applied directly to directional data, since, depending on the location where the circle is cut to unwrap it to an interval of length 2π on the real line, a mode, e.g., around 0, on the circle may split into two modes, e.g., around 0 and 2π (cf. Fig. 2(a) and (b)), on the line. To perceive the modality better, instead of cutting the circle, Mardia and Jupp [9] recommend repeating a complete cycle of the data, yielding an interval of length 4π on the real line. Wu et al. [10] proceed in this way, padding circular data to fit mixture models to wave direction data using a standard variational Bayesian technique and an initial overestimated number of components. The original data set defined over $[0,2\pi]$ is padded at both ends to obtain an extended data set over $[-\pi,3\pi]$, by duplicating the observations in the $(0,\pi]$ interval to $(2\pi,3\pi]$ and those in $[\pi,2\pi)$ to $[-\pi,0)$ (cf. Fig. 2(c)). We follow the same approach of padding the data, which allows using GMMs with the standard EM algorithm.

Once a GMM has been fitted to the augmented data set, the next step is to restrict the interval $[-\pi,3\pi]$ to the initial interval $[0,2\pi]$ and summarize the final components on the circle. Thus, for modes like those in Fig. 2(c), where the data values cluster around 0 and 2π on the line, we want to group the initial observations into one component. For modes around other values, e.g., $\pi/2$, we want to discard the mode around $5\pi/2$ and keep only the one in the original set. To this purpose, we perform a merging step: For every pair of observations in the initial data set and their repeated counterparts, we determine the pairs of components to which both observations can belong to with posterior probabilities over a given threshold. Then, for each unique pair of components $(\mathcal{N}(\mu_i,\sigma_i^2),\mathcal{N}(\mu_j,\sigma_j^2))$, we compute the distance between their means

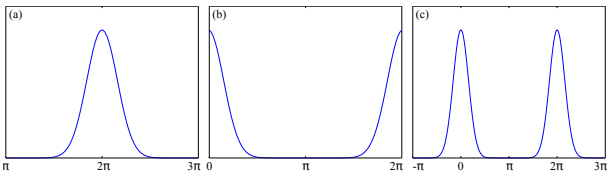

Figure 2: Cutting a circular distribution with a mode at 0 at the antimode π (a) conveys the mode clearly, whereas cutting it at the mode (b) creates the impression of bimodality. The ambiguity is resolved by repeating the complete cycle and padding each half to the left and right (c).

$$d(\mu_i,\mu_j) = \pi - \left|\pi - |\mu_i - \mu_j|\right| \tag{2}$$

and consider for merging the components with means that are situated close to each other on the unit circle. Components that are not merged are kept only if they contain initial observations. All means are restricted to values in the interval $[0,2\pi]$.

The resulting components can be summarized via sample trigonometric moments. For a component with m angles, we consider the corresponding unit vectors and compute the mean direction μ as that of the resultant vector after vector addition [2]. More specifically,

$$\mu = \begin{cases} \tan^{-1}(S/C) & \text{if } S \geq 0 \text{ and } C \geq 0, \\ \tan^{-1}(S/C) + \pi & \text{if } C \leq 0, \\ \tan^{-1}(S/C) + 2\pi & \text{if } S \leq 0 \text{ and } C \geq 0, \end{cases} \tag{3}$$

where

$$C = \Sigma_{i=1}^{m} \cos(\theta_i) \text{ and } S = \Sigma_{i=1}^{m} \sin(\theta_i). \tag{4}$$

The mean resultant length ρ associated to the mean direction μ is computed as the length of the resultant vector, normalized by m,

$$\rho = \frac{\sqrt{C^2+S^2}}{m}. \tag{5}$$

ρ takes values in the range $[0,1]$, higher values showing increased concentrations around the mean direction [9]. The two parameters characterize the Wrapped Normal distribution $WN(\mu,\rho)$, the result of wrapping a Normal distribution $\mathcal{N}(\mu,\sigma^2)$ (given on the line) around the circle:

$$\rho = \exp\left(-\frac{1}{2}\sigma^2\right), \text{ where } \sigma^2 = -2\log(\rho). \tag{6}$$

These formulae can also be used to switch between the WN and the corresponding Normal distribution.

A2. Local Similarity Measures for Directional Data

Assume that a sample $\theta_1,\ldots,\theta_n$ is modelled using a mixture of K Gaussian components. The posterior probabilities that each observation θ_i belongs to component k can be summarized in a $n \times K$ matrix P with elements $p_{i,k}$.

Two observations θ_i and θ_j are considered similar according to the *modality* measure if both have a posterior probability over a certain threshold τ of belonging to the same mode k

$$\mathrm{m}(\theta_i,\theta_j) = \begin{cases} 1 & \text{if } p_{i,k} \geq \tau \text{ and } p_{j,k} \geq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

If two observations θ_i and θ_j are assessed as similar according to the modality measure, we compute two other similarity measures to refine the pairwise characterizations. To this purpose, we first identify the component k which both observations are most likely to belong to. Thus, assuming there are one or more components $k_1,\ldots,k_q$ to which the two observations can belong with a posterior probability over the threshold τ, we select the mode k with the

highest sum of posterior probabilities $p_{i,k} + p_{j,k}$.

The observations are classified as similar according to the *scaled angular* measure if the smallest angle between them (computed as in Eq. 2) is less than or equal to the maximum sample circular standard deviation (cf. Eq. 6) in the domain

$$\mathrm{sm}(\theta_i, \theta_j) = \begin{cases} 1 & \text{if } \mathrm{m}(\theta_i, \theta_j) = 1 \text{ and } d(\theta_i, \theta_j) \leq \sigma_{\max}, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

For the *locally scaled angular* measure, the smallest angle is compared to the local sample circular standard deviation instead of to the maximum one

$$\mathrm{lsm}(\theta_i, \theta_j) = \begin{cases} 1 & \text{if } \mathrm{m}(\theta_i, \theta_j) = 1 \text{ and } d(\theta_i, \theta_j) \leq \sigma_k, \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

A3. Comparing Mixture Models

There is no best measure to assess the degree to which two directional pdfs are similar. We were looking for a measure that has a closed form for GMMs and is thus easily computable on the fly. While the Kullback-Leibler (KL) distance is a popular tool, it is analytically tractable only for two Gaussian distributions, reason for which in this paper we build upon the concordance coefficient [11], which fulfills the aforementioned criteria.

For two GMMs $f_1 = \sum_{i=1}^{N_1} \alpha_{1i} \mathcal{N}(\mu_{1i}, \sigma_{1i}^2)$ and $f_2 = \sum_{j=1}^{N_2} \alpha_{2j} \mathcal{N}(\mu_{2j}, \sigma_{2j}^2)$, the concordance coefficient reads

$$C(f_1, f_2) = \frac{2 \int f_1(x) f_2(x) \mathrm{d}x}{\int f_1^2(x) \mathrm{d}x + \int f_2^2(x) \mathrm{d}x} = \frac{2F(f_1, f_2)}{F(f_1, f_1) + F(f_2, f_2)}, \quad (10)$$

where

$$F(f, g) = \sum_{i=1}^{N_f} \sum_{j=1}^{N_g} \frac{\alpha_{fi} \alpha_{gj}}{\sqrt{\sigma_{fi}^2 + \sigma_{gj}^2}} \exp\left(-\frac{1}{2} \frac{(\mu_{fi} - \mu_{gj})^2}{\sigma_{fi}^2 + \sigma_{gj}^2} \right). \quad (11)$$

For circular data, we compute the variance as in Eq. 6 and the distance between two means as in Eq. 2. The concordance coefficient takes values between zero – when the two pdfs have completely dissimilar support sets – and one – when they are identical. In order to get values on $[0, \infty)$, we use the coefficient $D(f_1, f_2) = -\log(C(f_1, f_2))$ instead, so that, for constant variation, this measure grows instead of decreasing with the distance between the means. It should be noted that we only use this measure for two pdfs of the same type, i.e., the similarity coefficient between a uniform pdf and a mixture model is infinite, whereas any two uniform pdfs are considered identical.

References

[1] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26, CRC press, 1986

[2] N. Fisher. *Statistical analysis of circular data.* Cambridge Univ. Press, 1995.

[3] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions. *J. Mach. Learn. Res.*,6:1345–1382, Dec. 2005.

[4] P. J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, 20: 5365, 1987.

[5] G. McLachlan and D. Peel. *Finite mixture models*. Wiley & Sons, 2000.

[6] C. Hennig. Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1):334, 2010.

[7] A. Oliveira-Brochado and F. V. Martins. Assessing the number of components in mixture models: A review. Technical report, Univ. of Porto, 2005.

[8] G. Hamerly and C. Elkan. Learning the K in K-Means. In *Neural Information Processing Systems*,2003.

[9] K. Mardia and P. Jupp. *Directional statistics*, Wiley & Sons, 2000.

[10] B. Wu, C.A. McGrory, and A. N. Pettitt. The variational bayesian approach to fitting mixture models to circular wave direction data. *J. Applied Meteorology and Climatology*, 51(10):17501762, 2012.

[11] S. Ray. *Distance-based model-selection with application to the analysis of gene expression data*. PhD thesis, PSU, 2003.