ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-855

Hybrid Gaussian and von Mises Model-Based Clustering

Sergio Luengo-Sanchez [1] and **Concha Bielza** and **Pedro Larrañaga**

Abstract. Data collected about a phenomenon often measures its magnitude and direction. The most common approach to clustering this data assumes that directional data can be modeled as Gaussian. However, directional data has special properties that conventional statistics cannot handle. To deal with them, other approaches like the von Mises distribution must be applied. In this paper we present a new model based on mixtures of Bayesian networks to simultaneously cluster both linear and directional data.

1 Introduction

In a wide range of scientific fields, angle measurement is required to represent information about a phenomenon. Usually, this data comes together with its magnitude. Examples are in meteorology with wind direction and speed measurements [10], rhythmometry, medicine or demography [3, 4].

Typically, when this data is collected, an exploratory analysis is performed to reveal patterns. Cluster analysis partitions data into groups of homogeneous observations. A probabilistic clustering approach is model-based clustering [14, 31, 32]. Model-based clustering assumes that data was generated by a statistical model. Finite mixtures of Gaussians are the most commonly used distribution in model-based clustering because they can approximate any non-directional multivariate density given enough components [45].

However, mixtures of Gaussians which are based on classical statistics, are not suitable for clustering directional data because they cannot handle its periodicity. For example, given angles $1°$ and $359°$, the linear mean would be $180°$. This points in the opposite direction to the angular mean angle which is $0°$. To address this problem, a popular choice for the component distribution when data is positioned on the surface of a sphere or hypersphere is the von Mises-Fisher (vMF) distribution [2, 17]. The vMF distribution is the circular analogue of the multivariate normal distribution whose covariance matrix is a multiple of the identity matrix [26]. This model outperforms others based on linear distributions for problems such as text categorization and gene expression analysis [2, 47]. Another common choice when data is multimodal and is distributed on a torus or hypertorus are the mixtures of von Mises (vM) distributions. Mixtures of bivariate [29] and multivariate [25] vM distributions have been applied successfully in bioinformatics to characterize the structure of proteins.

Because mixtures of Gaussian, vM or vMF distributions partially solve the problem of simultaneously clustering directional and linear data, several distributions have been proposed to cluster cylindrical data, that is, a linear and a directional variable together [9, 16, 28, 38]. They all represent the joint probability density function of a univariate Gaussian and a univariate vM distribution. However, data is usually multidimensional and consists of a large number of linear and directional variables.

In our work we introduce a hybrid model inspired by a recent directional extension of the naive Bayes classifier [23], mixing multivariate Gaussian and multivariate vMF distributions. We adapt this model to learn a mixture where each component is the product of a multivariate Gaussian and several independent vM distributions. This is a learning from incomplete data problem, where variables are observed and the cluster membership is a hidden variable [22]. We exploit conditional independence assumptions encoded by the Bayesian network to factorize the joint probability distributions. This decomposition enables efficient model learning.

The rest of this paper is organized as follows. Section 2 introduces background material. Section 3 describes a clustering algorithm for naive Bayes, that is, a Bayesian network with structural constraints, where data are directional only. Under the naive Bayes assumption, Section 4 extends the previous model to the hybrid case mixing Gaussian and vM distributions. Additionally, the hybrid case is further improved by learning the Bayesian network structure from Gaussian variables to relax the topology of the Gaussian part. The approaches are evaluated and results are discussed in Section 5. Section 6 outlines the conclusions and future research.

2 Background

2.1 Von Mises distribution

The univariate vM distribution for an angle $\theta \in [0, 2\pi]$ defines a probability density function over points on a circle with a radius of one according to

$$f_{vM}(\theta; \mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)}, \tag{1}$$

where $\mu \in [0, 2\pi]$ is the mean direction, $\kappa \geq 0$ is the concentration parameter and $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero

$$I_0(\kappa) = \frac{1}{\pi} \int_0^{\pi} e^{\kappa \cos\theta} d\theta. \tag{2}$$

Parameters μ and $1/\kappa$ are analogous to the normal distribution μ and σ^2, so the vM distribution is also known as the circular normal distribution.

2.2 Model-based clustering

Consider the finite set $\mathbf{X} = \{X_1, X_2..., X_L\}$ of variables and let $\mathbf{x} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N\}$ be a dataset where each $\mathbf{x}^i$ assigns a value to all variables in $\mathbf{X}$. The goal of model-based clustering is to recover the statistical model that generated $\mathbf{x}$. Finite mixture models provide a formal setting for model-based clustering. In finite mixture models,

[1] Universidad Politécnica de Madrid, Spain, email: sluengo@fi.upm.es

each cluster is represented by a probability distribution. The linear superposition of the above distributions generates the finite mixture density function

$$f(\mathbf{x};\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x};\boldsymbol{\theta}_k), \tag{3}$$

where the mixing proportions π_k are nonnegative and sum to 1, $\boldsymbol{\theta}_k$ are the parameters for cluster k, and K is the number of mixture components. In our case, f_k denotes a joint probability distribution over $\mathbf{x}$ encoded as a Bayesian network.

A Bayesian network [22, 35] is a directed acyclic graph that represents the probabilistic relationships among variables in $\mathbf{X}$. It consists of a pair $B = (S, \boldsymbol{\theta})$. The first component, S, is the graph structure whose vertices correspond to the variables $X_1, X_2 ..., X_L$. The second component, $\boldsymbol{\theta}$, represents the parameters. We use $\mathbf{Pa}_l = \{U_{1l}, U_{2l}, ..., U_{Tl}\}$ to denote the parents of node X_l in S, where U_{tl} is its t-th parent. Structure S encodes the local Markov property, i.e., each variable X_l is independent of its non-descendants given its parents $\mathbf{Pa}_l$. Hence, the joint probability distribution can be factorized as

$$f(\mathbf{x}) = \prod_{l=1}^{L} f(X_l|\mathbf{Pa}_l;\boldsymbol{\theta}). \tag{4}$$

Naive Bayes (NB) [13] is the simplest Bayesian network structure and one of the most extended models for classification. All variables are assumed to be conditionally independent given the class variable. Accordingly, the class variable, or cluster variable Z in our case, is the only parent of all variables in the graph and no more arcs are allowed. Although there are few real-world cases where this strong assumption about the conditional dependencies holds, is simple, its accuracy is competitive and it has a small generalization error. One of its advantages is that the probability distribution of Z is efficiently computed when the data is complete because the maximum likelihood estimator (MLE) or maximum a posteriori (MAP) methods can be applied to estimate parameters. In a clustering problem, however, cluster Z is a latent or hidden variable. Thus, this is a parameter learning problem with missing data where the values of the class variable are unknown. When working with missing data, we need to estimate the optimum parameters of the model at the same time as we try to hypothesize the cluster of each instance. Because of its simplicity and efficiency the expectation-maximization (EM) algorithm [12] is the most popular method for dealing with this problem.

EM addresses the missing data problem selecting a starting point, which is either an initial set of parameters or an initial assignment to the cluster variable. Once we have a parameter set, we can apply inference to complete the data or, conversely, once we have the complete data, we can estimate the set of parameters from MLE. Thus, there are two separate steps: use parameters to complete the data (expectation step)

$$Q_i(z^i) = p(z^i|\mathbf{x}^i;\boldsymbol{\theta}) = \frac{f(\mathbf{x}^i|z^i;\boldsymbol{\theta})p(z^i;\boldsymbol{\theta})}{\sum_z f(\mathbf{x}^i|z;\boldsymbol{\theta})p(z;\boldsymbol{\theta})}, \tag{5}$$

and then estimate a new set of parameters from the complete data (maximization step)

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{z^i} Q_i(z^i) \log \frac{f(\mathbf{x}^i, z^i;\boldsymbol{\theta})}{Q_i(z^i)}. \tag{6}$$

Both steps are repeated iteratively, and the likelihood improves until convergence to a local maximum.

In clustering with Bayesian networks, the EM algorithm only optimizes one of the components of the pair $B = (S, \boldsymbol{\theta})$, namely the parameters $\boldsymbol{\theta}$. The structure S must be preset and fixed. However, finding the underlying structure of the data has advantages such as discovering dependence relations between variables and efficient factorization. There are several successful methods for structure and parameters learning when data is complete [18, 22]. If data is complete, a heuristic search [11, 22, 46] for an optimal structure may compare the scores [1, 39] of the current model and the model built after adding or removing arcs between variables. When data is incomplete, this search is no longer viable because the network score does not decompose, and inference is needed at every step of the learning process to evaluate the model. Structural EM [15, 36] is a method based on EM to learn structure and parameters when data is incomplete.

Structural EM introduces structural learning as an additional step in the EM algorithm. It starts with a given initial structure S and a set of parameters $\boldsymbol{\theta}$. Then, it iterates between a pair of steps. First, the parameters are maximized according to the standard EM algorithm because it is cheaper than searching for a better model. When it converges, the algorithm searches for the best structure completing data with the output of the expectation step. Both steps are repeated until convergence. Any general-purpose heuristic search algorithm for structural learning can be applied. A common choice for the score to be maximized is the Bayesian information criterion (BIC) [39] because, if the search procedure always finds a better structure at each iteration, BIC guarantees that the score increases monotonically. BIC is a measure that adds a penalty to the log-likelihood $\hat{L}$ based on the number of model parameters v. This score is used to choose the clustering model (parameterization, structure and number of clusters) to ensure that the selected models are not too complex. BIC is computed as

$$BIC = 2\hat{L} - v\log(N), \tag{7}$$

where N is the size of the dataset.

3 Clustering directional data

Directional data clustering has been addressed previously in the literature based on the EM framework and vM distribution. Several studies have modeled directional data with mixtures of univariate vM distribution [7, 30, 33]. The use of the EM algorithm to cluster mixtures of bivariate vM distributions is investigated in [29] . The maximization step was tackled by means of numerical optimization because the MLE does not have a closed-form solution. A multivariate vM mixture model is studied in [27] proposing an approximation of the intractable normalizing constant when data is highly concentrated. This approach computes MLE according to the method of moments and the EM. However, the likelihood function may not always be monotonically increasing because it is using an approximation, although it does usually stabilize to some local maximum.

In this section, we first introduce a mixture of Bayesian networks for clustering data when variables are only directional, i.e., $\mathbf{Y} = Y_1, Y_2 \cdots, Y_M$. In fact, the goal is to learn the parameters $\boldsymbol{\theta}$ of Bayesian networks given that they have a NB structure S. Figure 1 shows the graphical structure of the proposed model for each mixture component where, assuming that directional data is distributed according to the vM distribution, nodes represent vM variables and arcs encode the dependencies.

We choose the NB structure because its factorization can solve the above clustering problems for multivariate vM distributions. To

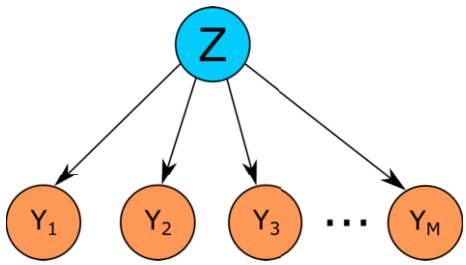

Figure 1. Graphical structure S for the naive Bayes model. The latent variable Z is the parent of all the variables, ruling out all other arcs. Thus, given Z, all the variables are conditionally independent of each other.

exploit the benefits of NB factorization, however, data must be complete. This is not the case in clustering because Z is hidden. Therefore, we need to apply the EM algorithm. First, we compute the expected values of Z according to the expectation step (Equation (5)). This completes the data so the joint distribution can be factorized to

$$f(\mathbf{Y}, Z; \boldsymbol{\theta}) = p(Z; \boldsymbol{\theta}) \prod_{m=1}^{M} f(Y_m | Z; \boldsymbol{\theta}), \tag{8}$$

which is a product of conditional probabilities such that each variable of the model contributes a factor of that product. This representation shows that NB naturally extends to M-dimensional data, which is one of the benefits of factorization.

Once the expectation step has been calculated, parameter estimation is carried out in the maximization step. Substituting the joint probability distribution of Equation (8) in Equation (6) results in

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{z^i} Q_i(z^i) \left[\sum_{m=1}^{M} \log f(y_m^i | z^i; \boldsymbol{\theta}) + \log p(z^i; \boldsymbol{\theta}) \right]. \tag{9}$$

Thus, the sum of the log-likelihood of each variable must be maximized for each cluster to compute the MLE of $\boldsymbol{\theta}$ in the mixture. Representing MLE as a summation simplifies parameter estimation so that each variable is optimized locally, i.e., independently of the others. The biggest advantage of this property is that MLE equations are closed-form and are computed efficiently for each variable Y_m of cluster k as [8]

$$\begin{aligned} \hat{\pi}_k &= \frac{\sum_{i=1}^{N} Q_i(k)}{N} \\ \hat{\mu}_k^m &= \arctan\left(\frac{\sum_{i=1}^{N} Q_i(k) \sin(y_m^i)}{\sum_{i=1}^{N} Q_i(k) \cos(y_m^i)} \right) \\ \hat{\kappa}_k^m &= A^{-1}\left(\frac{\sum_{i=1}^{N} Q_i(k) \cos(y_m^i - \hat{\mu}_k^m)}{\sum_{i=1}^{N} Q_i(k)} \right), \end{aligned} \tag{10}$$

where $A(\hat{\kappa}_k^m) = \frac{I_1(\hat{\kappa}_k^m)}{I_0(\hat{\kappa}_k^m)}$. An accurate approximation for A^{-1} is presented in [6].

This approach based on exploiting independence constraints avoids the numerical optimization needed for the mixtures of bivariate vM distributions. It also ensures that likelihood increases monotonically in each EM step until convergence to a local maximum even though data is not concentrated. However, NB does not capture information about the dependence between variables inside each cluster.

4 Hybrid Gaussian and von Mises clustering

Some practical scenarios involve several linear and directional variables. For example, geomagnetic and ionospheric signals are apparently associated with earthquake prediction [24, 42, 43, 44]. Thus, a dataset for this field of study is composed of measure such as magnitude and coordinates of an earthquake, speed of solar wind, ionospheric total electron content, magnetic activity, etc. A hybrid model for jointly clustering multivariate linear and multivariate directional variables can be achieved by means of mixtures of Bayesian networks. In this section we present a hybrid Gaussian and von Mises clustering from a preset NB structure. Then, we improve this model by learning the graph structure among Gaussian variables during the clustering process.

4.1 Conditionally independent variables

In Section 3, NB dependence constraints among variables were exploited to factorize the joint probability as a product where each factor corresponded to one variable. When they were combined using the EM algorithm to estimate the cluster parameters, the parameters of each variable were maximized independently of the others, resulting in closed-form equations. The advantages provided by the factorization of the NB structure on directional data are now extrapolated to achieve a model for multidimensional hybrid data.

Given a set of linear $\mathbf{X} = X_1, X_2, ..., X_L$ and directional $\mathbf{Y} = Y_1, Y_2, ..., Y_M$ variables, a NB structure S (see Figure 2) and the expected values of Z computed according to the expectation step, the joint probability distribution factorizes as

$$f(\mathbf{X}, \mathbf{Y}, Z; \boldsymbol{\theta}) = p(Z; \boldsymbol{\theta}) \prod_{l=1}^{L} f(X_l | Z; \boldsymbol{\theta}) \prod_{m=1}^{M} f(Y_m | Z; \boldsymbol{\theta}). \tag{11}$$

This introduces a new product of conditional probabilities of Gaussians with respect to Equation (8). Consequently, the model generalizes to higher dimensions and handles any number of linear and directional variables.

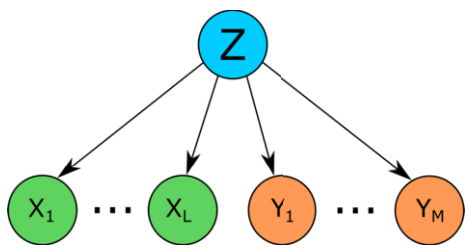

Figure 2. Graphical structure S for the NB hybrid model. Hybrid Gaussian and vM NB consists of a graphical structure where green nodes are Gaussian variables, orange nodes are vM variables and Z is the parent of all the variables.

The maximization step is computed by substituting the joint probability with the above factorization (11):

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{z^i} Q_i(z^i) \\ \left[\sum_{l=1}^{L} \log f(x_l^i | z^i; \boldsymbol{\theta}) + \sum_{m=1}^{M} \log f(y_m^i | z^i; \boldsymbol{\theta}) + \log p(z^i; \boldsymbol{\theta}) \right] \end{aligned} \tag{12}$$

Because of the NB structure assumption, parameter estimation involves the maximization of sums of log-likelihoods. Therefore, the parameters of Gaussian and vM variables are estimated locally. Hence, the maximization step for vM variables can be computed according to Equation (10) and for Gaussian variables according to the well-known equations

$$\hat{\mu}_k^l = \frac{\sum_{i=1}^N Q_i(k) x_l^i}{\sum_{i=1}^N Q_i(k)}$$
$$\hat{\sigma}_k^l = \sqrt{\frac{\sum_{i=1}^N Q_i(k)(x_l^i - \hat{\mu}_k^l)^2}{\sum_{i=1}^N Q_i(k)}}. \tag{13}$$

4.2 General hybrid model

Strong constraints were imposed on the structure of the Bayesian network for the above models. However, when the structure is preset and fixed, some beneficial properties of the Bayesian networks are lost. Discovering the graph topology provides information about the relations of dependence between variables and may improve the model's accuracy. The Structural EM algorithm [15, 36] defines a flexible approach to clustering, automatically learning the structure of the network during the clustering process.

The proposed multivariate model aims to fit hybrid data, so some relations between variables must be constrained in the learning structure step of the Structural EM algorithm to exploit factorization efficiently as we did in previous sections. First, we assume independence between Gaussian and vM variables given the parent node Z. Second, vM variables should be conditionally independent of each other to achieve closed-form equations for the maximization step. As a result, a new scenario is set where Gaussian dependencies are freely learned by Structural EM (without constraints), the structure of vM variables is fixed and dependencies between Gaussian and vM variables are ruled out (Figure 3).

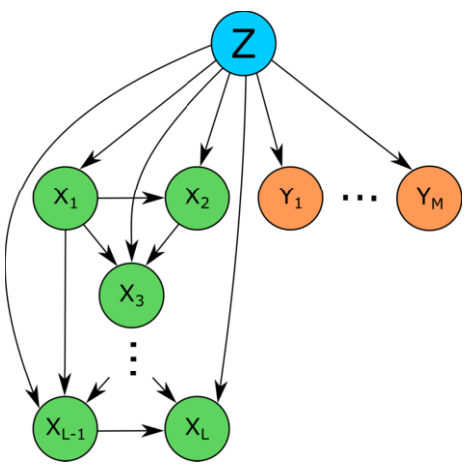

Figure 3. An example of the graphical structure S for the general hybrid model. The structure of Gaussian variables is learned during the clustering process. vM variables are independent given Z which is the parent of all the variables. There is no dependence between Gaussian and vM variables.

The first step of Structural EM algorithm, EM computation, is addressed as in previous sections by computing the expected values of Z according to the expectation step, we find that a new distribution is factorized as

$$f(\mathbf{X}, \mathbf{Y}, Z; \boldsymbol{\theta}) = p(Z; \boldsymbol{\theta}) \prod_{l=1}^{L} f(X_l | \mathbf{Pa}_l, Z; \boldsymbol{\theta}) \prod_{m=1}^{M} f(Y_m | Z; \boldsymbol{\theta}), \tag{14}$$

which is quite similar to the factorization shown in Equation (11). The difference lies in the decomposition of the conditional probability distribution of Gaussian variables because other Gaussian variables may be their parents ($\mathbf{Pa}_l$). In (14), $f(X_l|\mathbf{Pa}_l, Z; \boldsymbol{\theta})$ is a linear Gaussian, i.e., a linear combination of its Gaussian parents

$$f(X_l|\mathbf{Pa}_l, Z = k; \boldsymbol{\theta}) = \mathcal{N}(\beta_{0k} + \boldsymbol{\beta}_k \mathbf{Pa}_l, (\sigma_k^l)^2), \tag{15}$$

where $\boldsymbol{\beta}_k$ is the vector of coefficients of the linear Gaussian for cluster k. When the only parent of a Gaussian variable is Z, then $\beta_{0k} = \mu_k$.

Parameter estimation is tackled by the maximization step substituting the joint probability distribution by its factorization (14):

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \sum_{z^i} Q_i(z^i) \left[\sum_{l=1}^{L} \log f(x_l^i | \mathbf{Pa}_l^i, z^i; \boldsymbol{\theta}) + \sum_{m=1}^{M} \log f(y_m^i | z^i; \boldsymbol{\theta}) + \log p(z^i; \boldsymbol{\theta}) \right]. \tag{16}$$

As in previous cases, due to the independence assumption represented by the structure, MLE entails maximizing a sum of log-likelihoods to locally estimate vM and Gaussian variables without Gaussian parents according to Equations (10) and (13). Nevertheless, some Gaussian variables have Gaussian parents on which the computation of their MLE depends. For these variables, we set

$$\mathbb{E}_D[X] = \sum_{i=1}^{N} Q_i(k) x^i,$$

and we get the MLE of $\hat{\boldsymbol{\beta}}$ coefficients from the following system of equations:

$$\begin{aligned}
\mathbb{E}_D[X_l] &= \hat{\beta}_{0k}\mathbb{E}_D[\mathbf{1}] + \hat{\beta}_{1k}\mathbb{E}_D[U_{1l}] + \cdots + \hat{\beta}_{Tk}\mathbb{E}_D[U_{Tl}] \\
\mathbb{E}_D[X_l \cdot U_{1l}] &= \hat{\beta}_{0k}\mathbb{E}_D[U_{1l}] + \hat{\beta}_{1k}\mathbb{E}_D[U_{1l} \cdot U_{1l}] + \cdots \\
&\quad + \hat{\beta}_{Tk}\mathbb{E}_D[U_{1l} \cdot U_{Tl}] \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\mathbb{E}_D[X_l \cdot U_{Tl}] &= \hat{\beta}_{0k}\mathbb{E}_D[U_{Tl}] + \hat{\beta}_{1k}\mathbb{E}_D[U_{1l} \cdot U_{Tl}] + \cdots \\
&\quad + \hat{\beta}_{Tk}\mathbb{E}_D[U_{Tl} \cdot U_{Tl}].
\end{aligned} \tag{17}$$

Once the coefficients are known, the variance of X_l is computed as

$$(\hat{\sigma}_k^l)^2 = \frac{\sum_{i=1}^N Q_i(k)(x_l^i - \hat{\beta}_{0k} - \hat{\boldsymbol{\beta}}_{\boldsymbol{k}} \mathbf{Pa}_l^i)^2}{\sum_{i=1}^N Q_i(k)}. \tag{18}$$

The expectation and maximization steps iterate until convergence.

The EM algorithm outputs complete data and a set of parameters $\boldsymbol{\theta}$. Structural EM applies this outcome to learn the structure of the Bayesian network. When complete data is available, heuristic search algorithms optimize the score locally due to the decomposability property. Thus, part of the network topology can be optimized, while the rest remains unchanged. We exploit this point to search the structure for the Gaussian variables. We choose the BIC score to search for the best structure because it guarantees that the algorithm always converges in a local maximum.

5 Experiments and results

In this section, we report two experiments with different goals. On the one hand, we numerically evaluate all the proposed models by clustering artificial datasets and measuring the accuracy of the estimated parameters. On the other hand, we introduce an application of the general hybrid model for neuroscience to cluster neuronal dendritic spines.

5.1 Artificial datasets

To achieve a deeper insight into the suitability of the above models for clustering tasks, we evaluate them numerically. We study the performance of the vM and general hybrid models comparing their goodness of fit and their accuracy against the most common choice for multivariate directional data modeling, the Gaussian mixture model. This study should highlight the differences of applying a linear distribution in place of a directional distribution for directional data modeling. For all the experiments, data was simulated to find out beforehand the component of the mixture that generated each instance and the model parameters for comparison with the outcome of the experiment. For each experiment we rebooted the algorithm 10 times, changing the initial parameterization each time. We saved the model that maximized the BIC score.

5.1.1 Von Mises model

In the first place, we evaluated the goodness of fit of the model based on mixtures of NB for vM variables which we compare with the Gaussian mixture model. To do this, we simulated data from three clusters and two variables $\Theta \sim vM(\mu_\Theta, \kappa_\Theta), \Phi \sim vM(\mu_\Phi, \kappa_\Phi)$. We set the concentration parameter κ to low values so clusters overlap. We analyzed the goodness of fit of both models depending on the sample size ($N = 30, 300$).

Table 1. Comparison of parameter estimation between vM and Gaussian models changing the sample size. Each cluster is denoted by Cl., followed by its number. For the Gaussian mixture model κ was computed as $1/\sigma^2$. We use boldface to denote the value of the distribution that best approximates each parameter for each cluster.

Variable	Parameters	Original		
		Cl. 1	Cl. 2	Cl. 3
Θ	μ_Θ	0	$\pi/2$	π
	κ_Θ	1	1	1
Φ	μ_Φ	0	$\pi/2$	π
	κ_Φ	2	2	3

N = 30

Variable	Parameters	vM Clustering			Gaussian Clustering		
		Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3
Θ	$\hat{\mu}_\Theta$	**-0.53**	**1.68**	**2.77**	0.79	2.1	5.57
	$\hat{\kappa}_\Theta$	3.89	2.94	**1.44**	**2.44**	**1.26**	5.66
Φ	$\hat{\mu}_\Phi$	**0.54**	0.77	**3.19**	0.79	**1.71**	6.06
	$\hat{\kappa}_\Phi$	2.89	**2.22**	**3.18**	**2.87**	0.3	100

N = 300

Variable	Parameters	vM Clustering			Gaussian Clustering		
		Cl. 1	Cl. 2	Cl. 3	Cl. 1	Cl. 2	Cl. 3
Θ	$\hat{\mu}_\Theta$	**-0.36**	**1.59**	**2.87**	5.03	1.62	3.25
	$\hat{\kappa}_\Theta$	1.4	1.34	**0.58**	**1.13**	**1.06**	0.29
Φ	$\hat{\mu}_\Phi$	**0.08**	1.27	**3.25**	4.87	**1.48**	1.84
	$\hat{\kappa}_\Phi$	**1.85**	**1.47**	**3.33**	0.66	0.5	0.92

Results from Table 1 show that the mixtures of NB for vM variables yield better results for estimating the mean of the distributions, especially when the mean is 0, than the Gaussian mixture model, which fails due to the special properties of the directional data. When the sample size increases, the proposed model further improves the estimation of κ values.

Then, we evaluated the performance of the clustering algorithm by changing the number of clusters ($K = 3, 5, 10$) and variables ($M = 10, 25, 50$). Modifying the number of variables provides information about the accuracy of the model when data is concentrated or sparse. Varying the number of clusters in a bounded and fixed space we measure the performance of the method as more clusters overlap. For the experiment, complete data was available, i.e., variables and cluster labels were known. We started by hiding the cluster label of all instances and clustering the data. We crisply assigned each instance to the cluster with maximum membership probability. As a result, each instance belonged to one group. Then, we compared the real label with label provided by the clustering algorithm to get its hit rate. The accuracy of the proposed model was compared against the Gaussian mixture model, see Table 2.

Table 2. Hit rate of vM vs Gaussian mixture models. We simulated 100 instances from each cluster. The best results are denoted in boldface.

	vM clustering			Gaussian clustering		
N. Cl./N. Var.	10	25	50	10	25	50
3	**99%**	**100%**	**100%**	94.6%	99.6%	68.33%
5	**97%**	**100%**	**100%**	47.2%	59.2%	**100%**
10	**56.2%**	**99.1%**	**100%**	38.7%	40.4%	38.3%

Analyzing Table 2 we find that mixtures of vM distributions improve their accuracy as the number of variables increases. This is because clusters are further apart and consequently easily separated in higher dimensions. The opposite applies when the number of clusters grows. In this case the clusters overlap. Therefore, the boundaries between them are not clearly defined, and clustering algorithms are less accurate. However, the Gaussian mixture model behaves differently. Even though data sparsity increases when the number of variables is 50, the accuracy of Gaussian variables decays for 3 and 10 clusters with respect to the case when there are 25 variables. For all cases vM clustering achieves better results than Gaussian mixture models.

5.1.2 General hybrid model

To evaluate the general hybrid model, we adapted the above experiments to hybrid data. We started by validating the goodness of fit and the structure learned by the model. To do this, we manually defined a Bayesian network with five Gaussian nodes and two vM nodes (Figure 4). As before the number of clusters is 3. We simulated 100 instance of this Bayesian network for each cluster. Then, we applied the general hybrid model to learn the model parameters and the structure from Gaussian variables.

We measured the distance between the original and the learned structure according to the Hamming distance, i.e., the number of changes in a BN structure needed to turn it into another. The operations are add an arc, drop an arc or revert arc. Figure 4 shows that we only need to add one arc ($X_5 \rightarrow X_3$) to achieve the original structure, so the Hamming distance was one and the structure was an accurate approximation.

Table 3 shows the results of parameter estimation. First, we observe that X_3 has one parameter less because the learned structure missed an arc with respect to the original structure. The elimination of the coefficient β_5 is offset by the remaining coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. Despite this fact, the value of $\hat{\sigma}$ accurately approximates the original

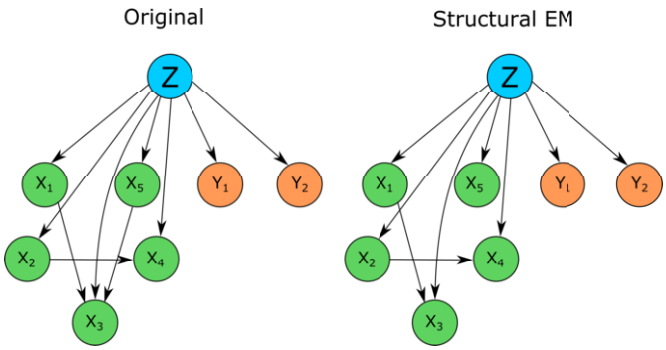

Figure 4. Original structure of the BN and structure learned by the general hybrid model. The Structural EM approximates the original structure quite well but drops the arc from X_5 to X_3.

value for that variable. Also note that good approximations were obtained for most of the estimated parameters, except in some cases like the mean of X_1 for cluster 2 and the mean of X_2 for cluster 3. Of particular note are the good results for the directional variables, especially for the means.

Table 3. Parameter estimation of general hybrid model with respect to the original model

Original				
Variable	Parameters	Cl. 1	Cl. 2	Cl. 3
X_1	β_0	0	1	0
	σ	1	2.27	2.27
X_2	β_0	0	0	1
	σ	1.4	2	2
X_3	$\beta_0, \beta_1, \beta_5$	0.04,1.05,0.11	-0.65,0.19,1.47	0,0.29,0.96
	σ	1.4	0.75	1.24
X_4	β_0, β_2	-0.01,0.77	-0.01,0.11	0.87,0.1
	σ	1.67	1.38	1.37
X_5	β_0	-0.01	0.99	0.04
	σ	2.3	0.99	1.03
Y_1	μ	0	$\pi/2$	π
	κ	1	1	1
Y_2	μ	0	$\pi/2$	π
	κ	2	2	3

General hybrid model				
Variable	Parameters	Cl. 1	Cl. 2	Cl. 3
X_1	$\hat{\beta}_0$	0.08	-0.1	-0.26
	$\hat{\sigma}$	1.05	2.41	2.35
X_2	$\hat{\beta}_0$	-0.12	0.56	0.14
	$\hat{\sigma}$	1.34	2.09	2.01
X_3	$\hat{\beta}_0, \hat{\beta}_1$	0.29,0.90	-0.06,-0.64	0.01,1.52
	$\hat{\sigma}$	1.56	0.79	1.30
X_4	$\hat{\beta}_0, \hat{\beta}_2$	0.18,0.77	0.05,0.02	0.85,0.00
	$\hat{\sigma}$	1.77	1.36	1.29
X_5	$\hat{\beta}_0$	-0.14	-0.08	-0.06
	$\hat{\sigma}$	2.31	1.07	1.05
Y_1	$\hat{\mu}$	0.11	1.43	2.91
	$\hat{\kappa}$	0.92	0.77	1.53
Y_2	$\hat{\mu}$	-0.19	1.56	3.13
	$\hat{\kappa}$	1.21	2.05	2.71

Next, we look at the performance of the general hybrid model by changing the proportional number of Gaussian and vM variables, as well as the number of clusters. We simulated three different datasets to evaluate the model and compare it with multivariate Gaussian mixture models. The first dataset had an equal number of linear and directional variables and consisted of five Gaussian and five vM variables. The second dataset had more linear variables: 15 Gaussian and 5 vM variables. The third dataset had 5 Gaussians and 15 vM variables. Again we hid the cluster label of the instances for data clustering.

Table 4. Hit rate of general hybrid and Gaussian mixture models. We simulate 100 instances for each cluster. We change the number of variables for the data. First we analyze 5 Gaussian and 5 vM, then 15 Gaussian and 5 vM and finally 5 Gaussian and 15 vM.

	General hybrid clustering			Gaussian clustering		
N. Cl./N. Var.	5-5	15-5	5-15	5-5	15-5	5-15
3	**99.6%**	**100%**	**100%**	99%	99.6%	**100%**
5	**95.4%**	**100%**	**100%**	89.2%	99.8%	99.6%
10	**94.6%**	**99.8%**	**100%**	81.9%	99.2%	95.5%

According to Table 4 general hybrid model overcomes Gaussian mixture model in all the proposed scenarios. General hybrid model yields better results when there is an equal number of linear and directional variables and a low dimensional space. However, when there are more linear variables than directional variables, the Gaussian mixture models turns competitive and almost tie our model. In the last trial, when the number of directional variables surpass the number of linear variables, the proposed model slightly outperforms the Gaussian mixture model. General hybrid model obtained better BIC score in all the cases.

5.2 Clustering of dendritic spines

Dendritic spines are small membranous protusions. They are receptors of excitatory synapses placed on the surface of some neuronal dendrites [34]. They have captured the attention of neuroscientists because their morphology has been associated with brain funcionality. For example, it has been claimed that thin spines contribute to learning, while the biggest and steady spines are linked to the memory process. Disturbances of their morphology or density have been related to mental disorders such as schizophrenia, dementia or mental retardation [21]. Therefore, the clustering of dendritic spines is attracting interest in neuroscience. A traditionally accepted categorization is described in [37], proposing four groups (Figure 5). There is also debate about whether morphologies constitute a continuum instead of discrete classes [19].

We present an application of the general hybrid model to cluster dendritic spines according to their morphology. For the experiment we used a set of 500 triangular meshes representing the surface of three-dimensional dendritic spines reconstructed from pyramidal neurons extracted from the cingular cortex of a human male (aged 40). Spines were provided by the Cajal Cortical Circuits Lab (UPM-CSIC). For details about spine acquisition, see [5].

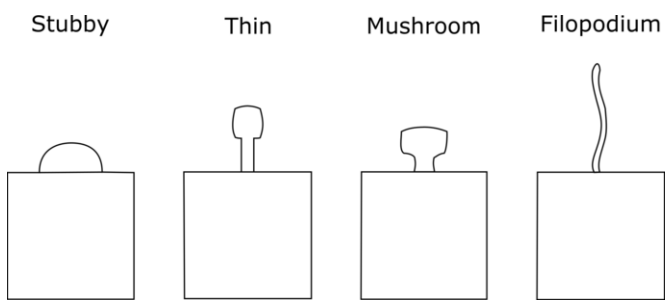

Figure 5. Traditional classification of spines proposed in [37], adapted from [40].

Meshes had to be previously transformed into data characterizing the morphology of the spines. This task was addressed using the multiresolutional Reeb graph (MRG) [20, 41] technique which constructs a graph from a 3D geometrical model to describe its topology. MRG partitions a triangular mesh into regions according to a function $\alpha(\cdot)$. In our case, this function was the geodesic distance because it is invariant to translation and rotation. Geodesic distance was computed from the mean point of the total surface that is in contact with the dendrite to each vertex of the mesh. The domain of $\alpha(\cdot)$ was divided into seven regions. For each region, we measured morphological characteristics, i.e, length, growth direction, eccentricity, flatness and size of the region (see Figure 6). There are a total of 35 linear variables and 30 directional variables.

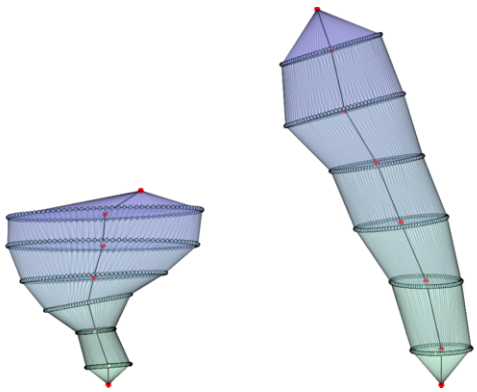

Figure 6. Examples of spines after computing and dividing the MRG it into regions. For each region, we measure morphological characteristics.

Since this experiment serves merely to illustrate an application of the proposed model and does not represent any valid neuroscientific result, we then jittered data with Gaussian and von Mises noise of zero mean. We ran the general hybrid model several times, modifying the number of clusters from two to ten. As a result, we managed to maximize the BIC score and AIC score for three clusters and seven clusters respectively. We analyzed exclusively the results provided by BIC score because its number of clusters is closest to the number of categories in the traditional classification.

To characterize clusters and compare them with the classification in [37], we performed a Welch t-test for linear variables and a Watson-Williams test for directional variables. We observed that almost all linear variables are significantly different between cluster 2 and the other two clusters. However, cluster 1 and cluster 3 only differ in so far as cluster 3 has a small neck at the base of the spine. We checked which cluster takes the maximum and minimum value for each measured feature to characterize the spine.

Thus, cluster 1 presents the shortest and flattest regions. Besides, all the regions are of the same size. This description fits the stubby class. Cluster 2 shows the longest and most elongated regions. Additionally, the size of the regions increases from the base to the top and the growth of the regions is less straight. Hence, this cluster groups filopodium and thin classes. Cluster 3 has short regions and a small base. This cluster grows backwards (in a $\frac{3\pi}{2}$ direction) while the other two clusters grow to the left (in a π direction). It apparently matches the mushroom class.

The Structural EM also provides some interesting information about the dependencies represented by the graph topology. For example, we find that the length of the next region depends on the length

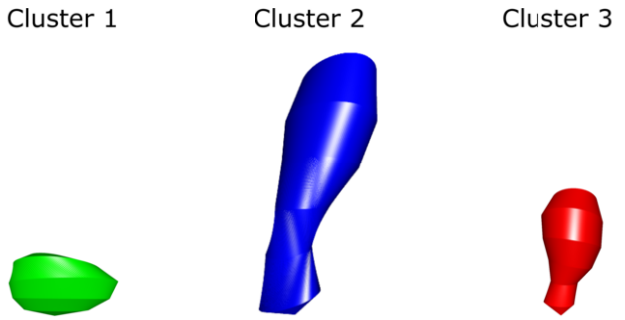

Figure 7. Examples of spines for each of the clusters. The spine representing cluster 1 is shaded green, the spine representing cluster 2 is shaded blue and the spine representing cluster 3 is shaded red.

of the previous regions. Also the size of the regions is related to the size of previous regions. We also observe connections between the eccentricity of the region and its length. These dependencies may be relevant for the electrophisiological behavior of the spine.

6 Conclusion

This paper investigated models for clustering hybrid (linear and directional) data. Although the most common approach for modeling this data is by means of Gaussian mixture models, directional data has some special properties that rule out the use of classical statistics. Assuming that directional data are Gaussian sometimes leads to poor approximations. In this paper, we reviewed previous models for clustering multivariate von Mises and multivariate hybrid data: current methods for clustering multivariate directional data are constrained to concentrated data and involve numerical optimization, whereas we did not find any specific clustering models for multivariate hybrid data.

This is why we proposed clustering models for multivariate directional and hybrid data based on Bayesian networks. To be precise, we exploited the benefits of factorization provided by the naive Bayes structure to get closed-form equations for the expectation-maximization algorithm. Additionally, we also improved the hybrid model by learning the graph structure from the linear variables according to the Structural EM framework.

We evaluated the proposed models against multivariate Gaussian distributions. The results provided by our models are better than the outcome of the Gaussian mixture model in almost all scenarios where directional data is involved. Besides, we applied a hybrid Structural EM algorithm to cluster dendritic spines with the aim of illustrating real applications of the model.

Future research includes the extension of the hybrid model to cover other distributions like von Mises-Fisher, Kent or discrete nodes, as well as relations of dependence between Gaussian and directional variables.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness through the Cajal Blue Brain (C080020-09; the Spanish partner of the Blue Brain initiative from EPFL) and TIN2013-41592-P projects, by the Regional Government of Madrid through the S2013/ICE-2845-CASI-CAM-CM project, and by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project).

REFERENCES

[1] H Akaike, 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control*, **19**(6), 716–723, (1974).

[2] A Banerjee, IS Dhillon, J Ghosh, and S Sra, 'Clustering on the unit hypersphere using von Mises-Fisher distributions', *Journal of Machine Learning Research*, **6**, 1345–1382, (2005).

[3] E Batschelet, *Circular Statistics in Biology*, Academic Press London, 1981.

[4] E Batschelet, D Hillman, M Smolensky, and F Halberg, 'Angular-linear correlation coefficient for rhythmometry and circannually changing human birth rates at different geographic latitudes', *International Journal of Chronobiology*, **1**(55), 183–202, (1973).

[5] R Benavides-Piccione, I Fernaud-Espinosa, V Robles, R Yuste, and J DeFelipe, 'Age-based comparison of human dendritic spine structure using complete three-dimensional reconstructions', *Cerebral Cortex*, **23**(8), 1798–1810, (2012).

[6] DJ Best and NI Fisher, 'The BIAS of the maximum likelihood estimators of the von Mises-Fisher concentration parameters', *Communications in Statistics-Simulation and Computation*, **10**(5), 493–502, (1981).

[7] S Calderara, R Cucchiara, and A Prati, 'Detection of abnormal behaviors using a mixture of von Mises distributions', in *IEEE Conference on Advanced Video and Signal Based Surveillance, 2007*, pp. 141–146, (2007).

[8] S Calderara, A Prati, and R Cucchiara, 'Mixtures of von Mises Distributions for People Trajectory Shape Analysis', *IEEE Transactions on Circuits and Systems for Video Technology*, **21**(4), 457–471, (2011).

[9] J A Carta, P Ramírez, and C Bueno, 'A joint probability density function of wind speed and direction for wind energy analysis', *Energy Conversion and Management*, **49**(6), 1309–1320, (2008).

[10] JA Carta, P Ramirez, and S Velazquez, 'A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands', *Renewable and Sustainable Energy Reviews*, **13**(5), 933–955, (2009).

[11] GF Cooper and E Herskovits, 'A Bayesian method for the induction of probabilistic networks from data', *Machine Learning*, **9**(4), 309–347, (1992).

[12] AP Dempster, NM Laird, and DB Rubin, 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B*, **1**, 1–38, (1977).

[13] RO Duda, PE Hart, and DG Stork, *Pattern Classification*, John Wiley & Sons, 2012.

[14] C Fraley and AE Raftery, 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association*, **97**(458), 611–631, (2002).

[15] N Friedman, 'Learning belief networks in the presence of missing values and hidden variables', in *ICML*, volume 97, pp. 125–133, (1997).

[16] R Gatto and SR Jammalamadaka, 'The generalized von Mises distribution', *Statistical Methodology*, **4**(3), 341–353, (2007).

[17] Kurt H and Bettina G, 'movMF: An R package for fitting mixtures of von Mises-Fisher distributions', *Journal of Statistical Software*, **58**(1), 1–31, (2014).

[18] D Heckerman, D Geiger, and DM Chickering, 'Learning Bayesian networks: The combination of knowledge and statistical data', *Machine Learning*, **20**, 197–243, (1995).

[19] A Herrera-Arellano, J Miranda-Sánchez, P Ávila-Castro, S Herrera-Álvarez, JE Jiménez-Ferrer, A Zamilpa, R Román-Ramos, H Ponce-Monter, and J Tortoriello, 'Clinical effects produced by a standardized herbal medicinal product of Hibiscus sabdariffa on patients with hypertension. A randomized, double-blind, lisinopril-controlled clinical trial', *Planta Medica*, **73**(1), 6–12, (2007).

[20] M Hilaga, Y Shinagawa, T Kohmura, and TL Kunii, 'Topology matching for fully automatic similarity estimation of 3D shapes', in *Proceedings of SIGGRAPH*, pp. 203–212. ACM, (2001).

[21] B Jacobs, L Driscoll, and M Schall, 'Life-span dendritic and spine changes in areas 10 and 18 of human cortex: A quantitative Golgi study', *Journal of Comparative Neurology*, **386**(4), 661–680, (1997).

[22] D Koller and N Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[23] PL López-Cruz, C Bielza, and P Larrañaga, 'Directional naive Bayes classifiers', *Pattern Analysis and Applications*, **18**(2), 225–246, (2013).

[24] JJ Love and JN Thomas, 'Insignificant solar-terrestrial triggering of earthquakes', *Geophysical Research Letters*, **40**, 1165–1170, (2013).

[25] KV Mardia, G Hughes, CC Taylor, and H Singh, 'A multivariate von Mises distribution with applications to bioinformatics', *The Canadian Journal of Statistics*, **36**(1), 99–109, (2008).

[26] KV Mardia and PE Jupp, *Directional Statistics*, John Wiley & Sons, 1999.

[27] KV Mardia, JT Kent, Z Zhang, CC Taylor, and T Hamelryck, 'Mixtures of concentrated multivariate sine distributions with applications to bioinformatics', *Journal of Applied Statistics*, **39**(11), 2475–2492, (2012).

[28] KV Mardia and TW Sutton, 'A model for cylindrical variables with applications', *Journal of the Royal Statistical Society. Series B*, **40**(2), 229–233, (1978).

[29] KV Mardia, CC Taylor, and GK Subramaniam, 'Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data', *Biometrics*, **63**, 505–512, (2007).

[30] N Masseran, AM Razali, K Ibrahim, and MT Latif, 'Fitting a mixture of von Mises distributions in order to model data on wind direction in Peninsular Malaysia', *Energy Conversion and Management*, **72**, 94–102, (2013).

[31] G McLachlan and KE Basford, *Mixture Models: Inference and Applications to Clustering*, Wiley, 1988.

[32] V Melnykov and R Maitra, 'Finite mixture models and model-based clustering', *Statistics Surveys*, **4**, 80–116, (2010).

[33] JA Mooney, PJ Helms, and IT Jolliffe, 'Fitting mixtures of von Mises distributions: A case study involving sudden infant death syndrome', *Computational Statistics and Data Analysis*, **41**(34), 505–513, (2003).

[34] EA Nimchinsky, BL Sabatini, and K Svoboda, 'Structure and function of dendritic spines', *Annual Review of Physiology*, **64**, 313–353, (2002).

[35] J Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., 1988.

[36] JM Peña, JA Lozano, and P Larrañaga, 'An improved Bayesian structural EM algorithm for learning Bayesian networks for clustering', *Pattern Recognition Letters*, **21**(8), 779–786, (2000).

[37] A Peters and IR Kaiserman-Abramof, 'The small pyramidal neuron of the rat cerebral cortex. The perikaryon, dendrites and spines', *American Journal of Anatomy*, **127**(4), 321–355, (1970).

[38] X Qin, SJ Zhang, and DX Yan, 'A new circular distribution and its application to wind data', *Journal of Mathematics Research*, **2**(3), 12–17, (2010).

[39] G Schwarz, 'Estimating the dimension of a model', *The Annals of Statistics*, **6**, 461–464, (1978).

[40] N Spruston, 'Pyramidal neurons: Dendritic structure and synaptic integration', *Nature Reviews Neuroscience*, **9**(3), 206–221, (2008).

[41] JWH Tangelder and RC Veltkamp, 'A survey of content based 3D shape retrieval methods', *Multimedia Tools and Applications*, **39**(3), 441–471, (2008).

[42] JN Thomas, JJ Love, and MJS Johnston, 'On the reported magnetic precursor of the 1989 Loma Prieta earthquake', *Physics of the Earth and Planetary Interiors*, **173**(3), 207–215, (2009).

[43] JN Thomas, JJ Love, MJS Johnston, and K Yumoto, 'On the reported magnetic precursor of the 1993 Guam earthquake', *Geophysical Research Letters*, **36**, (2009).

[44] JN Thomas, JJ Love, A Komjathy, OP Verkhoglyadova, M Butala, and N Rivera, 'On the reported ionospheric precursor of the 1999 Hector Mine, California earthquake', *Geophysical Research Letters*, **39**, (2012).

[45] DM Titterington, AFM Smith, and UE Makoy, *Statistical Analysis of Finite Mixture Distributions*, Wiley, 1985.

[46] I Tsamardinos, LE Brown, and CF Aliferis, 'The max-min hill-climbing Bayesian network structure learning algorithm', *Machine Learning*, **65**(1), 31–78, (2006).

[47] S Zhong and J Ghosh, 'A comparative study of generative models for document clustering', in *Proceedings of the Workshop on Clustering High Dimensional Data and Its Applications in SIAM Data Mining Conference*, (2003).