A Directional-Linear Bayesian Network and Its Application for Clustering and Simulation of Neural Somas

SERGIO LUENGO-SANCHEZ[iD], PEDRO LARRAÑAGA, AND CONCHA BIELZA

Computational Intelligence Group, Department of Artificial Intelligence, Universidad Politécnica de Madrid, 28660 Madrid, Spain

Corresponding author: Sergio Luengo-Sanchez (sluengo@fi.upm.es)

ABSTRACT Neural somas perform most of the metabolic activities in the neuron and support the chemical process that generates the basic elements of the synapses, and consequently the brain activity. The morphology of the somas is one of the fundamental features for classifying neurons and their functionality. In this paper, we characterize the morphology of the 39 three-dimensional reconstructed human pyramidal somas in terms of their multiresolutional Reeb graph representation, from which we extract a set of directional and linear variables to perform model-based clustering. To deal with this dataset, we introduce the novel Extended Mardia-Sutton mixture model whose mixture components are distributed according to a newly proposed multivariate probability density function that is able to capture the directional-linear correlations. We exploit the capabilities of Bayesian networks in combination with the Structural Expectation-Maximization algorithm to learn the finite mixture model that clusters the neural somas by their morphology and the conditional independence constraints between variables. We also derive the Kullback–Leibler divergence of the Extended Mardia-Sutton distribution to be used as a measure of similarity between soma clusters. The proposed finite mixture model discovered three subtypes of human pyramidal somas. We performed Weltch t-tests and Watson-Williams tests, as well as rule-based identification of clusters to characterize each group by its most prominent features. Furthermore, the resulting model allows us to simulate the 3D virtual representations of somas from each cluster, which can be a useful tool for neuroscientists to reason and suggest new hypotheses.

INDEX TERMS Clustering morphology soma, directional-linear data, Extended Mardia-sutton mixture model, Structural Expectation-Maximization algorithm.

I. INTRODUCTION

The study of a plethora of phenomena requires the measurement of their magnitude and direction. Examples include meteorology [1], rhythmometry, medicine, demography [2], [3] and neuroscience [4]. Usually, the first step after collecting these data is to perform an exploratory analysis to reveal patterns. A popular statistical tool to accomplish this task is cluster analysis, i.e., data division into homogeneous groups describing their main characteristics. A probabilistic approach is model-based clustering [5]–[7], which assumes that the data are generated by an underlying mixture of probability distributions. Finite mixture models [8]

The associate editor coordinating the review of this manuscript and approving it for publication was Bora Onat.

provide a formal setting for model-based clustering where each cluster is represented by a distribution. The most well-known method for probabilistic clustering is the Gaussian mixture model [9], which is widely applied because of its computational tractability and its suitability to approximate any linear multivariate density given enough components. However, Gaussian mixture models are not capable of handling periodicity of directional data, and consequently, they generally underperform in these datasets [10], [11].

Clustering of directional data has been broadly addressed in the literature. Mixtures of circular [12], [13], hyperspherical [14] and hypertoroidal [15] probability distributions have been successfully applied in problems such as text categorisation, gene expression analysis and characterization of the structure of proteins, overcoming those models based on

linear distributions. Nevertheless, clustering of joint directional-linear data with parametric models is challenging because of the lack of efficient density estimation methods and identifiability problems [16].

In this paper, we propose an extension of the Mardia-Sutton distribution [17], named the Extended Mardia-Sutton (EMS) distribution, to accommodate the joint presence of directional and linear variables, and it allows directional-linear and linear-linear correlations. Partially inspired by [10], we use Bayesian networks (BNs) to encode the new distribution, enabling efficient clustering of multivariate directional-linear data. To the best of our knowledge, this is the first model for clustering directional-linear data with multiple correlations between directional and linear variables. Additionally, we derive the Kullback-Leibler divergence expression between two EMS distributions as a measure of similarity between the clustering outcomes.

We present an application of our method to the neuroscientific problem of clustering three-dimensional neural somas by their morphology. Somas were characterized according to their multiresolutional Reeb graph representations [18], [19], which provide a geometrical description of their morphology based on a combination of both linear and directional variables. Based on this representation of the soma, we performed model-based clustering using the EMS mixture model and analyzed the morphology of the resulting groups and the similarity between them. We describe the process of simulating three-dimensional virtual somas from the model learned by the clustering algorithm and give some examples of simulated somas.

The rest of the paper is organized as follows. Section II briefly overviews model based-clustering with BNs and clustering with directional-linear data. Section III defines the Extended Mardia-Sutton distribution and its Kullback-Leibler divergence. Section IV describes the procedure to cluster and simulate neural somas. Section V draws conclusions and proposes future research.

II. THEORETICAL BACKGROUND

A. MODEL-BASED CLUSTERING WITH BAYESIAN NETWORKS

A BN $\mathcal{B}$ is a probabilistic graphical model [20], [21] that represents the joint probability density function (pdf) among sets of random variables. Concretely, in the directional-linear framework, the random variables are linear $\mathbf{X} = \{X_1, X_2, \ldots, X_L\}$ and directional $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_D\}$. A BN consists of a pair of components $\mathcal{B} = (\mathcal{G}, \boldsymbol{\theta})$, where $\mathcal{G}$ is the structure and $\boldsymbol{\theta}$ are the parameters of the model. The structure $\mathcal{G}$ is represented as a directed acyclic graph that encodes conditional independences among triplets of variables in the network. The set of parameters $\boldsymbol{\theta}$ comprises the sufficient statistics of the conditional pdf of each variable given its parents in $\mathcal{G}$. BNs satisfy the local Markov property, i.e., each variable is independent of its non-descendants given its parents in the graph. Hence, the joint pdf

decomposes as

$$f(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) = \prod_{l=1}^{L} f(X_l | \mathbf{Pa}_{X_l}^{\mathcal{G}}; \boldsymbol{\theta}) \prod_{d=1}^{D} f(Y_d | \mathbf{Pa}_{Y_d}^{\mathcal{G}}; \boldsymbol{\theta}). \quad (1)$$

We use $\mathbf{Pa}_{X_l}^{\mathcal{G}}$ and $\mathbf{Pa}_{Y_d}^{\mathcal{G}}$ to denote the parents of variables X_l and Y_d in $\mathcal{G}$ respectively.

Score and search-based BN learning can be approached as an optimisation problem [22], [23]. For a given structure $\mathcal{G}$, the parameters $\boldsymbol{\theta}$ are usually estimated by the maximum likelihood estimation (MLE) method. Structure optimisation is a search procedure over a set of candidate structures according to a scoring function that measures how well a network fits the observed data $\mathcal{D} = \{(\mathbf{x}^1, \mathbf{y}^1), \ldots, (\mathbf{x}^N, \mathbf{y}^N)\}$, where each instance $(\mathbf{x}^i, \mathbf{y}^i)$ assigns a value to all variables in $\mathbf{X}$ and $\mathbf{Y}$, respectively. Several heuristics have been proposed in the literature based on local search algorithms [22]–[24] to cope with the superexponential nature of the problem of searching for the highest-scoring network structure. These methods are based on scoring functions as the Bayesian information criterion (BIC) [25] defined as

$$\text{BIC}(\mathcal{D}, \mathcal{B}) = \ell(\mathcal{B} | \mathcal{D}) - \frac{v \log(N)}{2}, \quad (2)$$

where $\ell(\mathcal{B} | \mathcal{D})$ is the log-likelihood of the model and v is the number of parameters in $\mathcal{B}$. The evaluation of the candidate structures with BIC becomes efficient given that the log-likelihood and the penalization term decompose according to Equation (1), and consequently, each node of the BN is scored locally.

Model-based clustering [5]–[7] is generally defined as a finite mixture model [8], where each cluster is a component and represents a probability distribution. The convex combination of all probability distributions generates the mixture density function

$$f(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) = \sum_{k=1}^{K} p(Z; \boldsymbol{\theta}^k) f(\mathbf{X}, \mathbf{Y} | Z; \boldsymbol{\theta}^k), \quad (3)$$

where Z is a discrete latent variable with K latent states, $p(Z; \boldsymbol{\theta}^k)$ are the mixing weights, and $f(\mathbf{X}, \mathbf{Y} | Z; \boldsymbol{\theta}^k)$ is the distribution of the mixture component k. Learning the parameters of the mixture components $\boldsymbol{\theta}^k$ for clustering is a challenging task given that conditional independence assumptions encoded by Equation (1) do not apply when Z is unobserved or missing and that the evaluation of the BIC score requires inferred to be performed. Moreover, the search for an optimal set of parameters becomes a non-linear optimization problem that is usually addressed using the expectation-maximization (EM) algorithm [26], [27].

The EM algorithm is the most widely used algorithm for estimating the parameters of a model in the presence of incomplete data. The EM algorithm addresses the missing data problem by selecting a starting point, which is either an initial set of parameters or an initial assignment to the latent variable Z. Once we have a parameter set, we can apply

inference to complete the data; conversely, once we have the complete data, we can estimate the set of parameters from the MLE method. Thus, it is an iterative method comprising two steps. The expectation step (E-step) assigns to each instance i a probability of belonging to the mixture component k according to

$$p(z^i = k|\mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k) = \frac{f(\mathbf{x}^i, \mathbf{y}^i|z^i; \boldsymbol{\theta}^k)p(z^i; \boldsymbol{\theta}^k)}{\sum_{k=1}^{K} f(\mathbf{x}^i, \mathbf{y}^i|z^i; \boldsymbol{\theta}^k)p(z^i; \boldsymbol{\theta}^k)}, \quad (4)$$

where $i = 1, .., N$. As a result, the completed dataset $\mathcal{D}^+$ is obtained. The maximisation step (M-step) estimates a new set of parameters for the mixture component k from $\mathcal{D}^+$ as

$$\widehat{\boldsymbol{\theta}^k} = \underset{\boldsymbol{\theta}^k}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{k=1}^{K} p(z^i|\mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k) \log \frac{f(\mathbf{x}^i, \mathbf{y}^i, z^i; \boldsymbol{\theta}^k)}{p(z^i|\mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k)}. \quad (5)$$

The EM algorithm iterates between both steps monotonically, improving the likelihood of the model ($\ell(\mathcal{B}|\mathcal{D})$ in Equation (2)) until convergence.

The EM algorithm only optimises the parameters $\boldsymbol{\theta}$, assuming a fixed structure $\mathcal{G}$ in the BN. The Structural Expectation-Maximization (SEM) algorithm [28] extends the EM algorithm including structural learning to simultaneously optimize the structure and the parameters of a BN from incomplete data. SEM starts with a specified initial structure and an initial set of parameters or an initial assignment to Z. From the execution of the EM algorithm, a completed dataset $\mathcal{D}^+$ for a given BN structure is obtained. Once the data are completed, a score evaluating the fitness of the model decomposes, thus allowing score and search methods to efficiently learn an optimal structure for $\mathcal{D}^+$. SEM algorithm iterates between the EM algorithm and the structure optimisation. The BIC score is the common choice for the score to be maximized because it avoids overfitting, and if the search procedure always finds a better structure at each iteration, SEM ensures the convergence of the score function to a local optimum.

B. CLUSTERING DIRECTIONAL-LINEAR DATA

Directional data analysis is challenging because conventional statistics are not suitable to handle periodicity. Several circular distributions have been proposed in the literature to model angles as the wrapped normal, the wrapped Cauchy or the projected normal [12]. However, the von Mises (vM) distribution is the most prominent among the univariate circular distributions because of its analogy to the Gaussian distribution on the real line. In fact, when the fluctuations in the circular variable are small, it reduces to the Gaussian distribution. Given a circular random variable $0 \le Y \le 2\pi$, the vM pdf is defined as

$$f_{\mathcal{VM}}(Y; \mu_Y, \kappa_Y) = \frac{e^{\kappa_Y \cos(Y - \mu_Y)}}{2\pi I_0(\kappa_Y)}, \quad (6)$$

where $0 \le \mu_Y \le 2\pi$ is the location parameter representing the mean angle, $\kappa_Y > 0$ is the scale or concentration

parameter and, $I_0(\kappa_Y)$ is the modified Bessel function of the first kind and order zero, where

$$I_n(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos(Y)} \cos(nY) dY, \quad (7)$$

is the modified Bessel function of the first kind and order n.

The construction of a joint directional and linear pdf is a non-trivial problem. The literature regarding cylindrical distributions is scarce. Johnson and Wehrly [29] presented several cylindrical distributions, invoking maximum entropy principles and a general method based on copulas to construct *bivariate* cylindrical distributions with specific circular and linear marginals. This method has inspired new cylindrical distributions that provide higher tractability and more flexibility [30], [31] but suffers from some drawbacks. Because of the complicated theoretical results, copulas are suitable for the bivariate case but are difficult to extend to higher dimensions. Additionally, it is also arduous to find closed-form expressions of the MLE equations for copulas.

A different approach to define a *bivariate* cylindrical distribution was proposed by Mardia and Sutton [17]. They conditioned a Gaussian distribution to be a *bivariate* Gaussian. Then, they transformed the *bivariate* Gaussian from Cartesian to polar coordinates and restricted their parameters to construct the von Mises distribution (see Equation (8)). Thus, given the random variables $-\infty < X < \infty$ and $0 \le Y \le 2\pi$, they defined the joint pdf of the Mardia-Sutton distribution as

$$\begin{aligned} f_{\mathcal{MS}}(X, Y; \boldsymbol{\beta}, \sigma, \mu_Y, \kappa_Y) \\ = f_{\mathcal{VM}}(Y; \mu_Y, \kappa_Y) \\ \cdot f_{\mathcal{N}}(X; \beta_0 + \beta_1 \cos Y + \beta_2 \sin Y, \sigma) \end{aligned} \quad (8)$$

such that

$$\begin{aligned} \beta_0 &= \mu_X - \beta_1 \cos \mu_Y - \beta_2 \sin \mu_Y \\ \beta_1 &= \kappa_Y \operatorname{cov}(X, \cos Y), \quad \beta_2 = \kappa_Y \operatorname{cov}(X, \sin Y) \\ \sigma &= \sigma_X^2 - \kappa_Y \operatorname{cov}^2(X, \cos Y) - \kappa_Y \operatorname{cov}^2(X, \sin Y) \end{aligned}$$

where $f_{\mathcal{N}}(X; \beta_0 + \beta_1 \cos Y + \beta_2 \sin Y, \sigma)$ is a conditional linear Gaussian probability distribution, $\operatorname{cov}(\cdot, \cdot)$ is the covariance between two random variables, μ_X is the mean of X, σ_X is the standard deviation of X, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$. The resulting pdf has some desirable properties, as the marginal distribution for the angular variable is vM and the conditional distribution of the linear variable is Gaussian. However, the marginal distribution for the linear variable is complicated.

Clustering of fully correlated multivariate directional-linear data is an unsolved problem. The main reason is that, even when directional and linear variables are independent, multivariate directional distributions can hardly be extended beyond the bivariate case. The normalization constant of high-dimensional multivariate directional distributions is usually intractable, and only under certain circumstances may it be approximated [32]. Thus, little is known

TABLE 1. Summary of previous works involving clustering of directional-linear data with their limitations. The column named Dimensions denotes the maximum number of variables considered by each distribution.

Reference	Dimension	Limitations
Carta et al. [1]	2	As a copula-based distribution, it cannot be directly extended to higher dimensions
Roy et al. [33]	2	As a copula-based distribution, it cannot be directly extended to higher dimensions
Mastrantonio et al. [16]	2	Only considers one circular variable
Lagona et al. [34], [35]	> 2	Assume independence of all the variables given the latent variable
Bulla et al. [36]	> 2	Bivariate directional and bivariate linear distributions are conditionally independent given the latent variable
Roy et al. [11]	> 2	Full correlation among one circular and several linear variables
Luengo-Sanchez et al. [10]	> 2	Given the latent variable, it only allows correlations between linear variables

about efficient estimation methods for most of the multivariate directional-linear distributions. In the presence of latent variables, parameter estimation is even more challenging given the iterative nature of the EM algorithm. Numerical optimisation methods for estimating parameters can be prohibitive from a computational point of view when they are embedded inside the EM algorithm. These difficulties cause the literature regarding clustering directional-linear data to be limited to bivariate pdfs or models that impose strong conditional independence assumptions (see Table 1). Recently, Luengo-Sanchez et al. [10] exploited the conditional independence assumptions encoded by a BN to enable efficient running of the SEM algorithm for clustering, considering several directional and linear variables and allowing dependencies among the linear variables. However, the BN structure was constrained so that every directional variable was independent of all the other variables in the BN given the latent variable.

III. EXTENDED MARDIA-SUTTON MIXTURE MODEL BASED ON BNs

The aim of this section is to relax the limitations of all the directional-linear models shown in Table 1 by developing a multivariate distribution that accommodates more than one directional variable and extending [10] to allow correlations among directional and linear variables. We omit correlations between directional variables due to the intractability of the normalisation constant of the multivariate directional distributions, as discussed in Section II-B. Thus, given Z, the independence assumption between directional variables is mandatory to design an efficient clustering algorithm.

In model-based clustering, distributions of the mixture components whose MLE equations are closed-form are preferable over distributions that require numerical optimisation methods for parameter estimation for obvious computational efficiency reasons. One of the few cylindrical distributions whose MLE expressions are closed-form is the Mardia-Sutton distribution, which also has the advantage of being defined according to the maximum entropy distributions for directional and linear variables, i.e., the vM and the Gaussian distributions. To model directional-linear data, we propose the Extended Mardia-Sutton (EMS) distribution, an extension of the Mardia-Sutton distribution from the bivariate (Equation (8)) to the multivariate case, which is

defined as

$$
f_{\mathcal{EMS}}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\beta}, \mathbf{Q}, \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\kappa}_{\mathbf{Y}})
$$
$$
= \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d, \kappa_d)
$$
$$
\cdot f_{\mathcal{N}}(\mathbf{X}; \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top \cos \mathbf{Y} + \boldsymbol{\beta}_2^\top \sin \mathbf{Y}, \mathbf{Q}), \quad (9)
$$

where $\mathbf{X}$ has dimension L, $\mathbf{Y}$ has dimension D, $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)$, $\boldsymbol{\beta}_0$ is a vector of length L, $\boldsymbol{\beta}_1^\top$ and $\boldsymbol{\beta}_2^\top$ are matrices of size $L \times D$, $\mathbf{Q}$ is a covariance matrix of dimension L, and $\cos \mathbf{Y}$, $\sin \mathbf{Y}$, $\boldsymbol{\mu}_{\mathbf{Y}}$ and $\boldsymbol{\kappa}_{\mathbf{Y}}$ are vectors of length D. The detailed derivation and estimation of the parameters can be found in Appendix V.

Assuming that the dataset $\mathcal{D}$ has no missing values, a mixture model whose mixture components are distributed according to the EMS distribution explicitly imposes some constraints on the relations between the variables. First, Z is the only parent of the directional variables, i.e., $\mathbf{Pa}_{\mathbf{Y}}^{\mathcal{G}} = Z$. Thus, directional variables should be conditionally independent given the latent variable. Second, directional-linear correlations must be represented by conditioning linear variables to directional variables (and not vice versa). As shown by [10], BNs in combination with the SEM algorithm are suitable tools for learning generative models that satisfy conditional independence constraints between variables. Both restrictions can be encoded in the BN structure $\mathcal{G}$ by fixing the latent variable Z as the unique parent of the directional variables during the learning process. Fig. 1 shows an example of a BN structure representing a mixture of EMS distributions.

Next, the directional-linear data clustering procedure for mixtures of EMS distributions according to the SEM algorithm is described via the pseudocode of Algorithm 1. In line 1, SEM is initialised according to a given structure $\mathcal{G}_0$ and a set of parameters $\boldsymbol{\theta}_0$. Then, in lines 4-8, the EM algorithm iterates optimising the parameters until convergence. The dataset $\mathcal{D}$ is probabilistically completed according to the E-step (see Equation (4)) in line 6, giving the completed dataset $\mathcal{D}_s^+$ as a result, where s denotes the iteration of the EM algorithm. Once the data is complete and, consequently, the latent variable Z is observed, the joint pdf decomposes according to Equation (1) as

$$
f(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) = \sum_{k=1}^{K} p(Z; \boldsymbol{\theta}^k) \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d | Z; \boldsymbol{\theta}^k)
$$
$$
\cdot \prod_{l=1}^{L} f_{\mathcal{N}}(X_l | \mathbf{Pa}_{X_l}^{\mathcal{G}}; \boldsymbol{\theta}^k), \quad (10)
$$
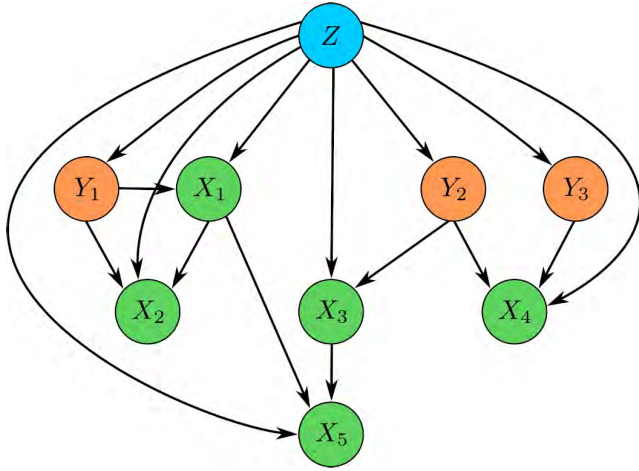
FIGURE 1. An example of a BN structure representing a mixture of EMS distributions. Green nodes are Gaussian variables, orange nodes are vM variables and Z is the latent variable. The only restriction on $\mathcal{G}$ is that the only parent of the vM nodes must be Z.

where $\mathbf{Pa}^{\mathcal{G}}_{X_l} \subset \{\mathbf{X}, \mathbf{Y}, Z\}$ and $Z \in \mathbf{Pa}^{\mathcal{G}}_{X_l}$. The conditional Gaussian distribution $f_{\mathcal{N}}(X_l | \mathbf{Pa}^{\mathcal{G}}_{X_l}; \boldsymbol{\theta}^k)$ is defined as

$$
\begin{aligned}
&f_{\mathcal{N}}(X_l | \mathbf{Pa}^{\mathcal{G}}_{X_l}; \boldsymbol{\beta}^k_l, \sigma^{2,k}_l) \\
&= f_{\mathcal{N}}(\beta^k_{0l} + \boldsymbol{\beta}^{k\top}_{1l} \mathbf{X} + \boldsymbol{\beta}^{k\top}_{2l} \cos \mathbf{Y} + \boldsymbol{\beta}^{k\top}_{3l} \sin \mathbf{Y}, \sigma^{2,k}_l) \\
&= f_{\mathcal{N}}(\beta^k_{0l} + \sum_{t=1}^{T} \beta^k_{tl} U_{tl}, \sigma^{2,k}_l).
\end{aligned}
$$

where $\boldsymbol{\beta}^k_l = (\beta^k_{0l}, \boldsymbol{\beta}^k_{1l}, \boldsymbol{\beta}^k_{2l}, \boldsymbol{\beta}^k_{3l})^{\top}$ are the regression coefficients and $\sigma^{2,k}_l$ is the variance of variable X_l for cluster k, β^k_{tl} are the non-zero coefficients in $\boldsymbol{\beta}^k_{1l}, \boldsymbol{\beta}^k_{2l}, \boldsymbol{\beta}^k_{3l}$, and T are the number of β^k_{tl} coefficients. Also, we substitute the random variables $(\mathbf{X}, \cos \mathbf{Y}, \sin \mathbf{Y})$ by $\mathbf{U}_l = (U_{1l}, \ldots, U_{Tl})$ for the sake of simplicity. Note that, for those variables $\mathbf{X}, \mathbf{Y} \notin \mathbf{Pa}^{\mathcal{G}}_{X_l}$, their regression coefficients are zero.

Parameter estimation is tackled in line 8 by the M-step (see Equation (5)). The decomposition of the joint pdf reduces the MLE computation to a set of local optimisations, one for each variable. For each latent state ($k = 1, \ldots, K$) of Z, the MLE equations for a directional variable Y_d are

$$
\hat{\mu}^k_d = \arctan \left(\frac{\sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k) \sin y^i_d}{\sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k) \cos y^i_d} \right),
$$

$$
\hat{\kappa}^k_d = A^{-1} \left(\frac{\sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k) \cos(y^i_d - \hat{\mu}^k_d)}{\sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k)} \right),
$$

where $A(\hat{\kappa}^k_d) = \frac{I_1(\hat{\kappa}^k_d)}{I_0(\hat{\kappa}^k_d)}$ (see Equation (7)). An accurate approximation for function $A^{-1}(\cdot)$ is presented in [37]. The non-zero coefficients for a linear variable X_l are estimated by solving the following system of equations:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[X_l] &= \hat{\beta}^k_{0l} \mathbb{E}_{\mathcal{D}}[\mathbf{1}] + \cdots + \hat{\beta}^k_{Tl} \mathbb{E}_{\mathcal{D}}[U_{Tl}] \\
\mathbb{E}_{\mathcal{D}}[X_l \cdot U_{1l}] &= \hat{\beta}^k_{0l} \mathbb{E}_{\mathcal{D}}[U_{1l}] + \cdots + \hat{\beta}^k_{Tl} \mathbb{E}_{\mathcal{D}}[U_{1l} \cdot U_{Tl}] \\
&\;\;\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\mathbb{E}_{\mathcal{D}}[X_l \cdot U_{Tl}] &= \hat{\beta}^k_{0l} \mathbb{E}_{\mathcal{D}}[U_{Tl}] + \cdots + \hat{\beta}^k_{Tl} \mathbb{E}_{\mathcal{D}}[U_{Tl} \cdot U_{Tl}],
\end{aligned}
$$

where $\mathbb{E}_{\mathcal{D}}[X] = \sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k) x^i$. The variance of the linear variable X_l is estimated from

$$
\hat{\sigma}^{2,k}_l = \frac{\sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k)(x^i_l - \hat{\beta}^k_{0l} - \sum_{t=1}^{T} \beta^k_{tl} u^i_{tl})^2}{\sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k)},
$$

where u^i_{tl} denotes the i-th instance of t-th random variable in $\mathbf{U}_l$. Finally, the prior probability of cluster k is computed as

$$
p(Z; \boldsymbol{\theta}^k) = \frac{1}{N} \sum_{i=1}^{N} p(z^i | \mathbf{x}^i, \mathbf{y}^i; \boldsymbol{\theta}^k).
$$

Algorithm 1 describes in lines 10-17 the hill climbing procedure for BN structure learning. It is a greedy method that iteratively computes a score function on all of the legal networks resulting from the application of a single operator to $\mathcal{G}_{j+1}$ (line 12). Usually, the operators considered are arc additions, deletions and reversions. At the end of each iteration, the hill-climbing procedure applies the operation that most improves the BIC score on the structure $\mathcal{G}_{j+1}$ (lines 16-17). The search for the optimal structure ends when there are no more local changes on the structure that improve the BIC score.

Algorithm 1 Pseudocode of the SEM Algorithm

Input: Dataset $\mathcal{D}$
Output: Best BN structure $\mathcal{G}^*$ and parameters $\boldsymbol{\theta}^*$
1   select $\mathcal{G}_0$ and $\boldsymbol{\theta}_0$;
2   **loop** for $j = 0, 1, \ldots$ until convergence
3     $\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_{j+1}$;
4   **loop** for $s = 0, 1, \ldots$ until convergence
5   // E-step
6   let $\mathcal{D}^+_s$ be the completed dataset inferred from $\mathcal{D}$ and $\boldsymbol{\theta}_s$;
7   // M-step
8   let $\boldsymbol{\theta}_s$ be arg max$_{\boldsymbol{\theta}}$ $\ell((\mathcal{G}_j, \boldsymbol{\theta}) | \mathcal{D}^+_s)$;
9     $\mathcal{G}_{j+1} \leftarrow \mathcal{G}_s, \boldsymbol{\theta}_{j+1} \leftarrow \boldsymbol{\theta}_s, \mathcal{D}^+_{j+1} \leftarrow \mathcal{D}^+_s$;
10   // hill-climbing procedure
11   **loop** for $s = 0, 1, \ldots$ until convergence
12   let $\mathbf{c}$ be the set of local changes that can be applied to $\mathcal{G}_j$;
13   **loop** for each c in $\mathbf{c}$
14   let $\mathcal{G}'$ be the result of applying c to $\mathcal{G}_j$;
15   let $\boldsymbol{\theta}'$ be arg max$_{\boldsymbol{\theta}}$ $\ell((\mathcal{G}', \boldsymbol{\theta}) | \mathcal{D}^+_{j+1})$;
16   **if** BIC$(\mathcal{D}^+_{j+1}, (\mathcal{G}', \boldsymbol{\theta}')) >$ BIC$(\mathcal{D}^+_{j+1}, (\mathcal{G}_{j+1}, \boldsymbol{\theta}_{j+1}))$ **then**
17         $\mathcal{G}_{j+1}, \boldsymbol{\theta}_{j+1} \leftarrow \mathcal{G}', \boldsymbol{\theta}'$;
18   $\mathcal{G}^*, \boldsymbol{\theta}^* \leftarrow \mathcal{G}_{j+1}, \boldsymbol{\theta}_{j+1}$;

A. KULLBACK-LEIBLER DIVERGENCE OF THE EXTENDED MARDIA-SUTTON DISTRIBUTION

The performance of clustering algorithms usually depends on the separability of the mixture components [38]. In addition, the identification and interpretation of clusters are easier when groups are homogeneous, i.e., when the

instances ascribed to each cluster belong to their cluster with a high probability. The overlap between probability distributions provides a quantitative description of these desirable properties, but its computation is often intractable analytically. For this reason, overlapping is usually replaced by similarity measures between distributions. Among them, one of the most widely used is the relative entropy or Kullback-Leibler divergence (KL) [39], which is defined as a measure of the difference between two distributions

$$D_{\text{KL}}(P||Q) = \int_{\mathbf{X},\mathbf{Y}} P(\mathbf{X},\mathbf{Y}) \log \frac{P(\mathbf{X},\mathbf{Y})}{Q(\mathbf{X},\mathbf{Y})} d\mathbf{X} d\mathbf{Y},$$

where $P(\mathbf{X},\mathbf{Y})$ and $Q(\mathbf{X},\mathbf{Y})$ denote, respectively, the probability density functions of distributions P and Q for a set of random variables.

The KL divergence formula can be expressed in closed-form for the EMS distribution, decomposing the joint pdf according to the independence assumptions represented by the BN structure (see Equation (1)) and applying the chain rule of relative entropy [40]

$$D_{\text{KL}}(P||Q) = \sum_{d=1}^{D} D_{\text{KL}}(P(Y_d)||Q(Y_d))$$
$$+ D_{\text{KL}}(P(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})||Q(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})).$$

Thus, the KL divergence for the joint pdf decomposes as a sum of KL divergences between univariate vM distributions and a conditional relative entropy of the multivariate density of the linear variables given their parents.

We define the two univariate vM distributions $P(Y_d)$ and $Q(Y_d)$ (Equation (6)) for the directional variable Y_d as

$$P(Y_d) = f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P) \quad \text{and} \quad Q(Y_d) = f_{\mathcal{VM}}(Y_d; \mu_d^Q, \kappa_d^Q)$$

respectively. Then, for the sake of simplicity in the calculations, distributions $P(Y_d)$ and $Q(Y_d)$ are rotated according to μ_d^P, giving as results the means $\mu_d^p = \mu_d^P - \mu_d^P = 0$ and $\mu_d^q = \mu_d^Q - \mu_d^P$ and the concentration parameters $\kappa_d^p = \kappa_d^P$ and $\kappa_d^q = \kappa_d^Q$. Note that the rotation does not change the concentration of the distributions. The KL divergence for the univariate vM distribution after the rotation is

$$D_{\text{KL}}(P(Y_d)||Q(Y_d)) = \log I_0(\kappa_d^q) - \log I_0(\kappa_d^p)$$
$$+ A(\kappa_d^p)\left(\kappa_d^p - \kappa_d^q \cos(\mu_d^q)\right),$$

where $A(\kappa_d^p) = \frac{I_1(\kappa_d^p)}{I_0(\kappa_d^p)}$. A detailed derivation of this KL divergence between two univariate vM distributions can be found in Appendix B-A.

The conditional relative entropy between distributions P and Q for the Extended Mardia-Sutton distribution is

$$D_{\text{KL}}(P(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})||Q(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}}))$$
$$= \int_{\mathbf{Y}} \prod_{d=1}^{D} P(Y_d) D_{\text{KL}}(P(\mathbf{X}|\mathbf{Y})||Q(\mathbf{X}|\mathbf{Y})) d\mathbf{Y}.$$

Given that $P(\mathbf{X}|\mathbf{Y})$ and $Q(\mathbf{X}|\mathbf{Y})$ are distributed according to a multivariate normal distribution (see Equation (9)),

the KL divergence is computed according to the well-known equation

$$D_{\text{KL}}(P(\mathbf{X}|\mathbf{Y})||Q(\mathbf{X}|\mathbf{Y}))$$
$$= \frac{1}{2}\Bigg[\text{Tr}(\mathbf{\Sigma}^{-1,Q}\mathbf{\Sigma}^P) + (\boldsymbol{\mu}^Q - \boldsymbol{\mu}^P)^\top \mathbf{\Sigma}^{-1,Q}(\boldsymbol{\mu}^Q - \boldsymbol{\mu}^P)$$
$$- L + \ln \frac{|\mathbf{\Sigma}^Q|}{|\mathbf{\Sigma}^P|}\Bigg],$$

where $\boldsymbol{\mu}^P$ and $\boldsymbol{\mu}^Q$ are the means and $\mathbf{\Sigma}^P$ and $\mathbf{\Sigma}^Q$ are the covariance matrices of the multivariate conditional Gaussian distributions represented by distributions $P(\mathbf{X}|\mathbf{Y})$ and $Q(\mathbf{X}|\mathbf{Y})$, L is the number of linear variables, and $|\cdot|$ is the determinant. For the sake of simplicity, we define $\boldsymbol{\mu}^R = \boldsymbol{\mu}^Q - \boldsymbol{\mu}^P$ as

$$\boldsymbol{\mu}^R = (\boldsymbol{\beta}_0^Q - \boldsymbol{\beta}_0^P) + (\boldsymbol{\beta}_1^Q - \boldsymbol{\beta}_1^P)^\top \cos \mathbf{Y} + (\boldsymbol{\beta}_2^Q - \boldsymbol{\beta}_2^P)^\top \sin \mathbf{Y}$$
$$= \boldsymbol{\beta}_0^R + \boldsymbol{\beta}_1^{R\top} \cos \mathbf{Y} + \boldsymbol{\beta}_2^{R\top} \sin \mathbf{Y}.$$

The conditional relative entropy is then computed according to

$$D_{\text{KL}}(p(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})||q(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}}))$$
$$= \frac{1}{2}\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\Bigg[\beta_{0i}^R\beta_{0j}^R + 2\beta_{0i}^R\sum_{d=1}^{D}\beta_{1jd}^R A(\kappa_d^P)$$
$$+ \sum_{d=1}^{D}\frac{\beta_{1id}^R\beta_{1jd}^R}{2}\left(1 + \frac{I_2(\kappa_d^P)}{I_0(\kappa_d^P)}\right)$$
$$+ \sum_{d=1}^{D}\sum_{m\neq d}^{D}\beta_{1id}^R\beta_{1jm}^R A(\kappa_d^P)A(\kappa_m^P)$$
$$+ \sum_{d=1}^{D}\frac{\beta_{2id}^R\beta_{2jd}^R}{2}\left(1 - \frac{I_2(\kappa_d^P)}{I_0(\kappa_d^P)}\right)\Bigg]$$
$$+ \frac{1}{2}\Bigg[\text{Tr}(\mathbf{\Sigma}^{-1,Q}\mathbf{\Sigma}^P) - L + \ln\frac{|\mathbf{\Sigma}^Q|}{|\mathbf{\Sigma}^P|}\Bigg],$$

where β_{1id}^R and β_{2id}^R are the d-th element of vectors $\boldsymbol{\beta}_{1i}^R$ and $\boldsymbol{\beta}_{2i}^R$ of variable X_i; β_{0i}^R, $\boldsymbol{\beta}_{1i}^R$ and $\boldsymbol{\beta}_{2i}^R$ are the coefficients of the conditional mean corresponding to the linear variable X_i, i.e., the i-th elements of vectors $\boldsymbol{\beta}_0^R$, $\boldsymbol{\beta}_1^R$ and $\boldsymbol{\beta}_2^R$; $\Sigma_{ij}^{-1,Q}$ is the element at the i-th row and j-th column of the matrix $\mathbf{\Sigma}^{-1,Q}$, and κ_d^P is the concentration parameter of the distribution $P(Y_d) = f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)$. A more detailed description of the procedure to obtain the above expression is provided in Appendix B-B.

IV. CLUSTERING AND SIMULATION OF NEURAL SOMAS
The soma is the component of a neuron in which its cell nucleus is placed. As the cell body, it contains the organelles common to living cells (mitochondria, Golgi apparatus, ribosomes, lysosomes, etc) that perform most of the metabolic activities in the neuron. These components also support the chemical processing of the neuron that origins the neurotransmitters, which are the basic elements of the synapses and consequently of the brain activity. The morphology of the soma

is one of the fundamental features for discriminating between different types of neurons [41]. However, in neuron classification and clustering, somas are usually ignored [42], or their morphometric analysis is limited to general two-dimensional somatic aspects, such as perimeter, surface area, elongation and roundness [43], [44]. Thus, the characterisation and clustering of three-dimensional reconstructed somas could provide a useful tool for achieving a more precise description of the morphology of the soma and discovering new subtypes of neurons.

In this section, we perform model-based clustering on the morphology of neural somas using the EMS mixture model. We also simulate 3D virtual representations of somas. To the best of our knowledge this is the first attempt to cluster and simulate three-dimensional neuronal somas according to their morphology. For the study, we used a set of three-dimensional triangular meshes representing the soma surface of pyramidal cells, which get their names from the pyramidal shape of their soma, from layer III of the cingulate (25 somas), temporal (16 somas) and frontal (18 somas) cortex of a 40-year-old human male. Further information regarding data collection and treatment is given at [45].

A. PREPROCESSING

During the neuron acquisition with confocal microscopy, some artefacts such as holes or cavities are usually present in the surface of most of the somas. Additionally, there is not a clear line demarcating the end of the soma and the origin of the dendrites and axon. Therefore, the morphometric analysis of the soma highly depends on the neuroanatomist to select the cutting point. Fig. 2 displays the preprocessing steps, originally proposed in [46], to repair and univocally find the morphology of a neural soma. Those somas whose surfaces were extremely damaged and could not be repaired were discarded. Thus, from the original dataset of 59 somas, we obtained 39 repaired somas.

For clustering purposes, 3D meshes representing the surface of the somas must be transformed into a set of morphological features that unambiguously captures the geometry of the somas, i.e., there must be a unique correspondence between an assignment to the features and a three-dimensional soma. If this condition is fulfilled, then the features should capture all of the relevant geometrical information of the soma, and consequently any morphometric measure can be computed from the set of features. Recently, in [49], a characterisation method was proposed based on this premise. It partitions the surface of the mesh into regions from a multiresolutional Reeb graph representation [18], [19] and computes a set of features for each region that locally characterizes the topology of the object, while the combination of all of the features provides a complete description of the soma morphology. Fig. 3 summarizes this characterisation of the somas. As a result of computing the multiresolutional Reeb graph, each soma was represented as a set of six regions and seven ellipses. Then, for each region i, we measured the following set of linear and directional features (see Fig. 4):
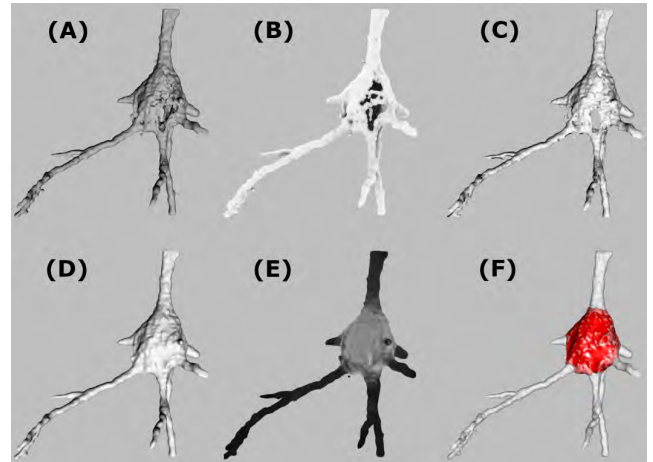

FIGURE 2. Summary of the soma preprocessing. (A) Example of a soma with holes and cavities in its surface. (B) Soma coloured according to light exposure resulting from the application of an ambient occlusion algorithm to the mesh. (C) Neuron after automatically identifying and removing the vertices of the mesh, forming holes and cavities by means of an unsupervised learning method. (D) Repair of the mesh surface with the Poisson surface reconstruction algorithm [47]. (E) Vertices of the mesh coloured according to the shape diameter function [48]. (F) Univocal definition of the soma (in red) after the segmentation of the dendrites (in white). Images adapted from [46].

- $|\mathbf{h}_i|$: Height of region i. It is the length of the vector $\mathbf{h}_i$ between the centroids of the ellipses bounding region i.
- $|B_i^R|$: Length of the major axis of ellipse B_i, where B_i is the closest ellipse to the apical dendrite of the pair of ellipses that bound region i.
- $|B_i^r|$: Length of the minor axis of ellipse B_i.
- $\cos\theta_i$: Curvature of the soma at region i. Taking vector $\mathbf{h}_i$ as the zenith of a spherical coordinate system, vectors $\mathbf{h}_i$ and $\mathbf{h}_{i+1}$ define a direction that can be expressed in spherical coordinates, i.e., the azimuth angle ϕ_i and elevation angle θ_i. The curvature is computed from the dot product $\cos\theta_i = \frac{\mathbf{h}_i \cdot \mathbf{h}_{i+1}}{|\mathbf{h}_i||\mathbf{h}_{i+1}|}$. Note that, although θ_i is an angle, and it is not periodical because its domain is $[0, \pi]$. In [50], it is discussed that the suitability of modelling random angles is clearly restricted to an interval smaller than 2π as circular variates, concluding that these angles should be treated like ordinary linear variables. Hence, we considered θ_i as a linear variable.
- ϕ_i: Growing direction of region i. It is the azimuth angle computed from vectors $\mathbf{h}_i$ and $\mathbf{h}_{i+1}$ that, combined with θ_i, describes the direction of a vector $\mathbf{h}_{i+1}$ in spherical coordinates.
- Θ_i: Direction of ellipse B_i. It is the polar angle or colatitude in the spherical coordinate system defined by the perpendicular vector to ellipse B_i, assuming the centroid of the ellipse as the origin. It is considered the instantaneous curvature. It is obtained from the vector $\frac{B_i^R}{|B_i^R|} \times \frac{B_i^r}{|B_i^r|}$. It was considered as a linear variable for the same reason as θ_i.
- Φ_i: Direction of ellipse B_i. The azimuth or azimuthal angle in the spherical coordinate system defined by
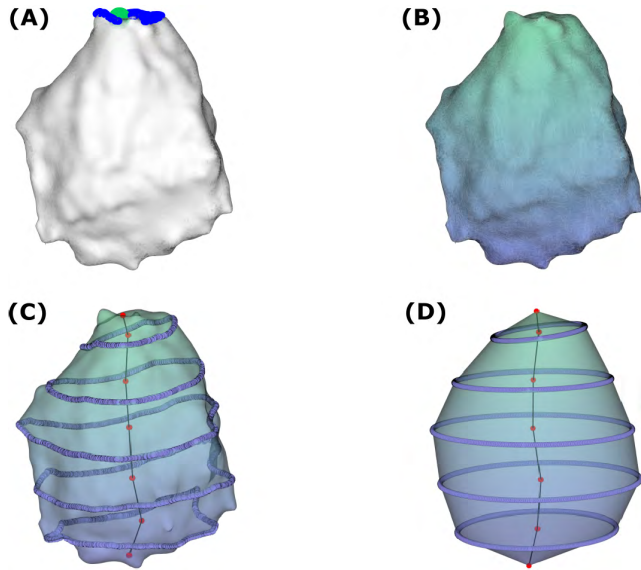
FIGURE 3. Characterisation of the soma morphology. **(A)** Computation of the insertion point. We obtained the blue points on top by projecting the vertices that represented the apical dendrite on the surface of the soma. The insertion point denoted by the colour green was the result of averaging all the blue points and searching for the closest vertex of the mesh to that mean. **(B)** Computation of the geodesic distance [51] from the insertion point. The soma is coloured with a gradient whereby the closest vertices to the insertion point were coloured green and the furthest were coloured purple. **(C)** Multiresolutional Reeb graph. We discretised the surface of the soma into equal-length regions according to the geodesic distance. All of the points in a curve are equidistant with respect to the insertion point (isolines or contour lines). **(D)** Each curve was approximated by the ellipse contained in the best fitting plane computed using principal component analysis.
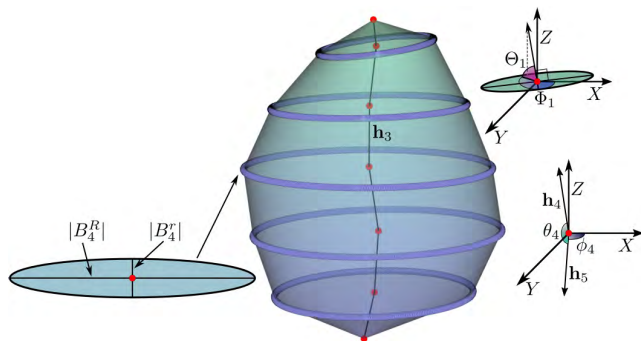
FIGURE 4. Feature extraction from the multiresolutional Reeb graph representation. Each ellipse B_i is defined by its centroid and major $|B_i^R|$ and minor $|B_i^r|$ axes. The height of each region is given by the length of the vector $\mathbf{h}_i$ between the centroids of the ellipses. Vectors $\mathbf{h}_i$ and $\mathbf{h}_{i+1}$ define a direction in spherical coordinates from which ϕ_i and θ_i are obtained. Φ_i and Θ_i are computed from the perpendicular vector to each ellipse B_i.

the perpendicular vector to the ellipse B_i assuming the centroid of the ellipse as the origin. It is obtained from the vector $\frac{B_i^R}{|B_i^R|} \times \frac{B_i^r}{|B_i^r|}$. Both Θ_i and Φ_i together describe the direction of the perpendicular vector to B_i.

B. CLUSTERING
The morphology of a soma was approximated with 43 variables, where 12 of them are directional and 31 are linear.

According to this characterisation, the number of variables was larger than the number of repaired somas. When the number of parameters to estimate is larger than the amount of data available, the model can overfit the data, or the covariance matrix can even become singular for some clusters. This problem gets worse in model-based clustering, as the number of parameters of the model increases linearly with the number of components of the mixture. Hence, we had to constrain the degrees of freedom of the model by introducing an upper bound to the maximum number of parents for each node, as well as the number of clusters. Another implementation detail is related to the SEM algorithm, which guarantees the convergence to a stationary point (local optimum, global optimum or a saddle point), which can be non-optimal in some cases. Because SEM is a deterministic algorithm, the starting point dictates the convergence point. To reduce the probability of converging to undesirable stationary points, Algorithm 1 was initialised from several random uniformly distributed starting points. We empirically set the maximum number of parents to two for the structure learning and executed the SEM algorithm 300 times from randomly selected starting points for two and three clusters.

From the SEM algorithm outcome (see Figure 5), we selected the model that maximised the BIC score (Equation (2)) and found three clusters as the best result. For each soma, we computed its probability of belonging to each cluster (p_1, p_2, p_3), where p_i is the membership probability of a soma to cluster i and $\sum_i^3 p_i = 1$. All of the somas were clearly ascribed to their most probable cluster as it was fulfilled that $\max\{p_1, p_2, p_3\} > 0.99$ in all of the cases. We then assigned each soma to its most probable cluster; the 39 somas that made up the complete dataset were distributed so that five somas belonged to Cluster 1, 17 somas were attributed to Cluster 2, and the remaining 17 somas were ascribed to Cluster 3. Examples of the somas assigned to each one of the three clusters are shown in Fig. 6. We also include three-dimensional representations of all the somas ascribed to each cluster as Supplementary Material.[1]

To identify the features that characterised each cluster, we performed the Welch t-test [52] on the linear variables and the Watson-Williams test [53] on the directional variables. Given a pair of clusters, the null hypothesis of both tests determined if both clusters had equal means. Table 2 shows that for each cluster, the features for which the null hypothesis was rejected with a p-value < 0.05 in all of the hypothesis tests performed between a given cluster and the rest of the clusters.

Table 2 is useful for distinguishing the clusters. Nevertheless, evaluating all of the characteristics at the same time is an arduous task for a neuroanatomist who wants to identify the most prominent properties of each group to determine possible functionalities. Using the rule-based learner

[1]The source code in R, the software documentation and the three-dimensional representations of the somas grouped by their cluster with higher membership probability are freely available at https://github.com/sergioluengosanchez/EMS_clustering.
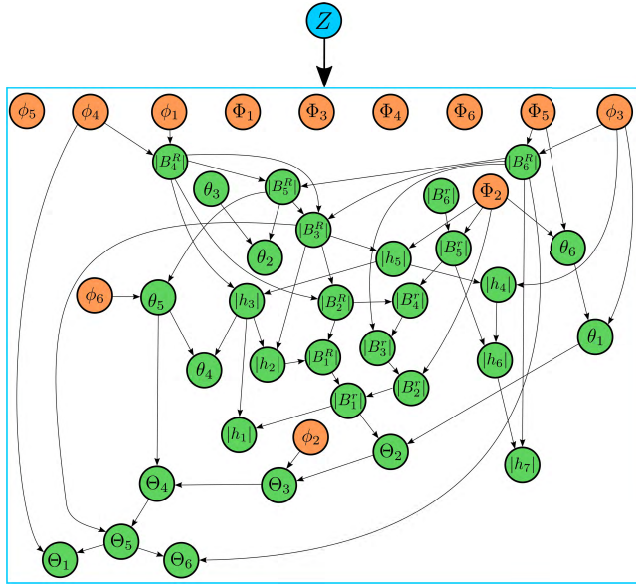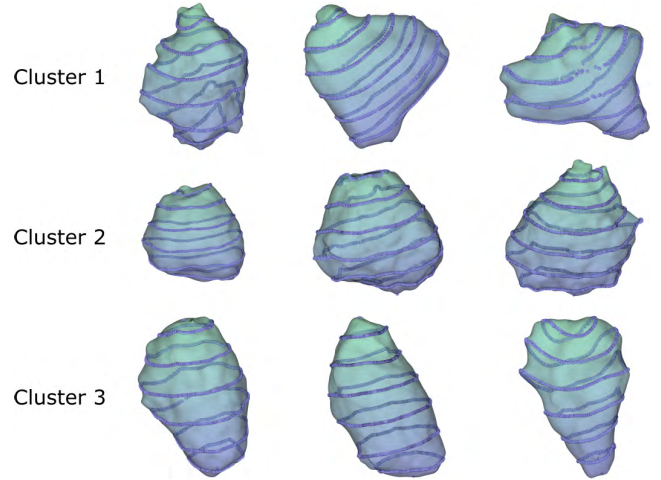
FIGURE 5. The BN structure learned by the SEM algorithm during the clustering process. To avoid cluttering the BN with many arcs, all the arcs from the latent variable *Z* (top) to each variable are represented as only one arc from *Z* to the group (inside the box). The BN structure shows that linear variables (green) are interrelated in consecutive regions, such as $|B_4^r| \rightarrow |B_3^r| \rightarrow |B_2^r|$. Also, curvature variables θ and Θ (orange) are mostly correlated with directional variables or other curvature variables.

TABLE 2. Results from the Welch t-test and the Watson-Williams test, which checked for significant differences between the means of the cluster and the rest of the clusters. The first column shows the names of the variables (a total of 20 out of 43) for which their mean was significantly different (*p*-value < 0.05) from the mean of the same variable in the rest of the clusters. The symbol < denotes that the mean of the variable was significantly smaller than it was for the other clusters, > denotes that the mean was significantly larger and = means that the mean was neither larger nor smaller and was significantly different.

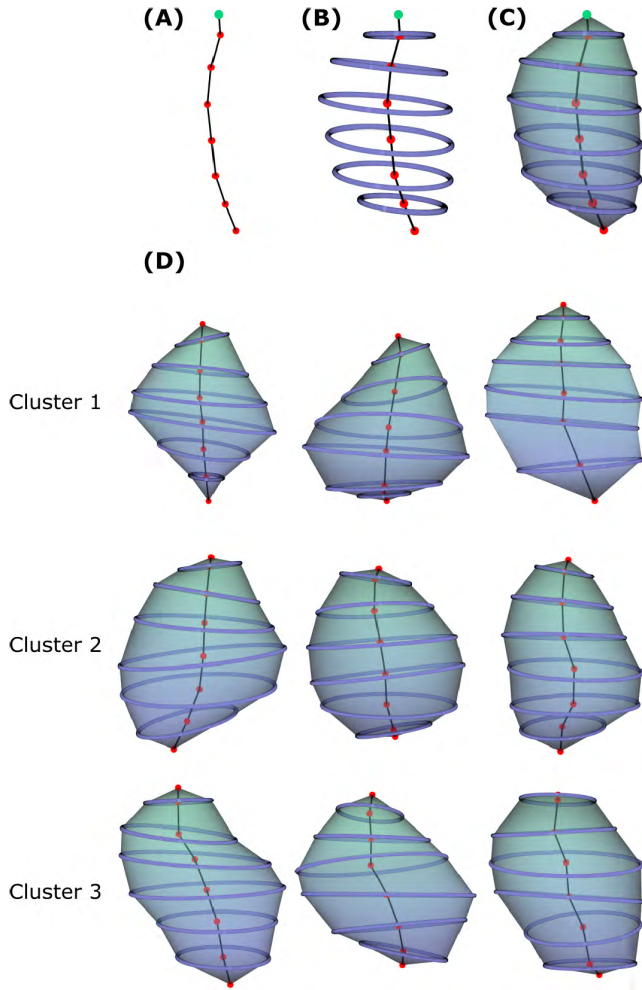| Variables	Cluster 1	Cluster 2	Cluster 3
$|\mathbf{h}_3|$			>
$|\mathbf{h}_4|$			>
$|\mathbf{h}_5|$			>
$|B_1^r|$	<	=	>
$|B_1^R|$			>
$|B_2^R|$			>
$|B_4^r|$	<		>
$|B_6^r|$	<		>
$\cos\theta_2$	>		
$\cos\theta_5$	<		
$\cos\theta_7$	<		
ϕ_6	>		
Θ_2		<	
Θ_3		<	
Θ_4		<	
Θ_5		<	
Θ_6		<	
Φ_2	<		
Φ_3	<		
Φ_6	<		

FIGURE 6. Examples of somas attributed to their most probable cluster.

each cluster with their accuracy between parentheses and a short description are:

- Cluster 1: $|B_3^r| \leq 4.59$ (89.8%). Somas whose short axis of the third ellipse is extremely small.
- Cluster 2: $\Theta_5 \geq 0$ and $\Theta_5 \leq 1.36$ and $|\mathbf{h}_3| \leq 3.62$ (71.8%). Somas whose fifth ellipse is slightly tilted and the farthest region from the apical dendrite is very short.
- Cluster 3: $|\mathbf{h}_3| \geq 3.66$ (76.9%). Somas whose third region is long.

To gain insights on the complete morphology of the somas, we extend the study based on the features extracted from the regions defined by the multiresolutional Reeb graph representation that describes locally the geometry of the somas. More concretely, we analyse the full set of variables checked as significantly different among clusters (Table 2) and the variables identified by RIPPER as a whole. We observe that the somas in Cluster 1 are mainly characterised by short axes in their ellipses. Therefore, the somas in this group are narrower than the rest. Cluster 2 can be distinguished because the variables related to the instantaneous curvature take lower values than for the other clusters. In consequence these somas tend to be more curved farther from the apical dendrite. Finally, the length of the regions as well as the length of the ellipse axes are significantly longer for Cluster 3, so the largest somas are grouped within this cluster.

From a neuroanatomical point of view, neurons with similar morphologies perform analogous brain functions. Therefore, it is interesting to find out which clusters of morphologies were more similar to each other. For this purpose, we computed the KL divergence between the three subtypes of pyramidal somas uncovered by our clustering approach. Thus, we obtained that the most similar clusters were Cluster 1 and Cluster 3 with a KL divergence of 869.4. Cluster 2 brought together the most different morphologies, as its KL divergences with respect to Cluster 1 and Cluster 3 were 2,078.3 and 1,629.0, respectively.

RIPPER [54], we summarized a unique rule for each cluster the set of characteristics needed to best discriminate between clusters. The rules generated by the RIPPER algorithm for

FIGURE 7. Simulation of virtual somas. (A) The skeleton is created. First, the insertion point (green) is placed at the origin of the coordinates. For each ellipse, compute the coordinates of its centroid from the centroid of the previous ellipse using the height $|\mathbf{h}_i|$, the curvature $\cos\theta_i$ and the growing direction ϕ_i of region i. (B) An ellipse for each centroid is generated. Given the centroid at the bottom of region i, 360 points are sampled from ellipse B_i defined by the length of its axes, $|B_i^R|$ and $|B_i^r|$, and its inclination given by Θ_i and Φ_i. (C) Finally, consecutive ellipses are triangulated to obtain a closed mesh. (D) Examples of virtual somas simulated from each cluster.

C. SIMULATION OF THREE-DIMENSIONAL SOMAS

One of the main challenges faced by neuroscience is the simulation of the human brain circuitry based on mathematical models. Given that ethical limitations prevent acquisition of the data directly from human brains, statistical models present an opportunity to reason, make predictions and suggest new hypotheses. The generative model implemented in this study allowed us to simulate virtual somas in a two-step process. First, new datasets were sampled from the joint pdf represented by the learnt BN. Then, for each instance of the new dataset, the three-dimensional representation of the soma was generated. Note that the univocal correspondence between an assignment to the variables and the geometry of the soma enabled the three-dimensional reconstruction. The procedure to obtain a virtual soma and some examples of virtual somas simulated from each cluster are shown in Fig. 7.

V. CONCLUSION AND FUTURE RESEARCH

In this paper, we presented a finite mixture model based on Bayesian networks for clustering directional-linear data. We developed a multivariate extension of the Mardia-Sutton distribution, as well as the Kullback-Leibler divergence between two EMS distributions in a closed-form. The proposed multivariate distribution relaxes the independence constraints among directional and linear variables of previous directional-linear models applied in clustering, allowing for any number of directional-linear correlations and avoiding approximate estimation of the parameters. This approach has several advantages over the models previously presented in the literature for clustering directional-linear data. First, our model can capture directional-linear correlations without limitations. Second, we avoid computationally costly optimisation algorithms for the estimation of the parameters during the EM algorithm, which can be extremely inefficient limiting the clustering to low-dimensional directional-linear datasets. Additionally, optimisation algorithms do not ensure the convergence of the EM algorithm to a local optima. Finally, learning the Bayesian network structure during clustering we reduce the complexity of the model while discovering the conditional independencies between variables.

We applied the finite mixture model to the neuroscientific problem of clustering neural somas by their morphology. The characterisation of the somas according to the adapted multiresolution Reeb graph representation enabled three-dimensional simulation of virtual somas from the three groups found by the SEM algorithm. Therefore, we describe for the first time the complete process for the repair, segmentation, clustering and simulation of 3D somas. The resulting model can be a useful tool for reasoning and suggesting new hypotheses regarding the function of the somas from a neuroscientific perspective.

The proposed model could be improved by conditioning the directional variables to linear variables or to other directional variables using, for example, the projected normal distribution as in [16]. This would increase the expressiveness of the model and would probably simplify the structures of the BNs learnt by the SEM algorithm, but numerical optimisation would be needed to estimate the parameters of the model. Additionally, most of the regression models in the cylindrical framework are based on bivariate distributions that assume a linear relation between the directional and the linear variables [12] or on non-parametric regression procedures, which are more flexible but also more difficult to extend to multivariate data [55]. Inspired by [56], where a Gaussian mixture regression is proposed, future research would include the development of an EMS mixture regression model for non-linear regression analysis when the independent variables are directional and linear, and the response variable is linear. Regarding the neuroscientific problem of somas, we plan to gather more data from different cerebral cortex layers and subjects and repeat the experiment while relaxing the constraints on the model as increase the maximum number of parents per node and the maximum number of clusters

for the SEM algorithm. We also consider to perform neuron classification and clustering using the set of features proposed in this work given that they could provide a better description of the soma morphology than the usual characterization of the literature, which consists of bidimensional measures as the perimeter, the area, the elongation or the sphericity of the soma.

APPENDIX A
DERIVATION OF THE EXTENDED MARDIA-SUTTON DISTRIBUTION

To obtain a cylindrical distribution, Mardia and Sutton [17] decomposed a *trivariate* normal distribution into a Gaussian distribution conditioned to a *bivariate* Gaussian. Then, they transformed the *bivariate* Gaussian from Cartesian to polar coordinates and restricted their parameters to construct the von Mises distribution (see Equation (8)). We define the Extended Mardia-Sutton distribution following a similar procedure.

First, we consider two disjoint sets of linear random variables $\mathbf{X}_a \in \mathbb{R}^L$ and $\mathbf{X}_b \in \mathbb{R}^{2D}$, where L is the number of linear variables and D is the number of directional variables. We assume that $\mathbf{X}_a$ and $\mathbf{X}_b$ are distributed according to the following joint probability density function:

$$f\begin{pmatrix}\mathbf{X}_a \\ \mathbf{X}_b\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}\mathbf{X}_a \\ \mathbf{X}_b\end{pmatrix}; \begin{pmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}\right],$$

where $\boldsymbol{\mu}_a \in \mathbb{R}^L$ and $\boldsymbol{\mu}_b \in \mathbb{R}^{2D}$ are the means of $\mathbf{X}_a$ and $\mathbf{X}_b$, respectively, $\boldsymbol{\Sigma}_{aa}$ is a matrix of dimension $L \times L$, $\boldsymbol{\Sigma}_{ab}$ is a matrix of dimension $L \times 2D$, $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^T$, and $\boldsymbol{\Sigma}_{bb}$ is a matrix of dimension $2D \times 2D$.

Applying the chain rule of probability to the joint probability density function of the multivariate normal distribution, the well-known expression for the conditional normal distribution is obtained

$$f\begin{pmatrix}\mathbf{X}_a \\ \mathbf{X}_b\end{pmatrix} = f(\mathbf{X}_b)f(\mathbf{X}_a|\mathbf{X}_b)$$
$$= f_{\mathcal{N}}(\mathbf{X}_b; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb})f_{\mathcal{N}}(\mathbf{X}_a; \boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top\mathbf{X}_b, \mathbf{Q}), \quad (11)$$

where

$$\boldsymbol{\beta}_0 = \boldsymbol{\mu}_a - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\mu}_b,$$
$$\boldsymbol{\beta}^\top = \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1},$$
$$\mathbf{Q} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}.$$

The next step transforms the multivariate normal distribution on variables $\mathbf{X}_b$ from Cartesian to polar coordinates through a Jacobian transformation. The components of the transformation are

$$\mathbf{X}_{b1} = \mathbf{r} \circ \cos\mathbf{Y}$$
$$\mathbf{X}_{b2} = \mathbf{r} \circ \sin\mathbf{Y},$$

where $\mathbf{X}_b = (\mathbf{X}_{b1}, \mathbf{X}_{b2})^\top$, $\mathbf{X}_{b1}$ and $\mathbf{X}_{b2} \in \mathbb{R}^D$, $\mathbf{r} = (r_1, \ldots, r_D)^\top$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_D)^\top$ are the vector of directional variables and $0 \leq Y_d \leq 2\pi$ for all $d = 1, \ldots, D$. The Jacobian determinant for the transformation is

given by

$$J(\mathbf{r}, \mathbf{Y}) = \begin{vmatrix} \dfrac{\partial\mathbf{X}_{b1}}{\partial\mathbf{r}} & \dfrac{\partial\mathbf{X}_{b1}}{\partial\mathbf{Y}} \\ \dfrac{\partial\mathbf{X}_{b2}}{\partial\mathbf{r}} & \dfrac{\partial\mathbf{X}_{b2}}{\partial\mathbf{Y}} \end{vmatrix} = \begin{vmatrix} \cos\mathbf{Y} & -\mathbf{r}\circ\sin\mathbf{Y} \\ \sin\mathbf{Y} & \mathbf{r}\circ\cos\mathbf{Y} \end{vmatrix} = \prod_{d=1}^{D} r_d.$$

Hence, the resulting expression of applying the Jacobian transformation to Equation (11) is

$$f(\mathbf{X}_a|\mathbf{Y}, \mathbf{r})f(\mathbf{Y}, \mathbf{r})$$
$$= f_{\mathcal{N}}((\mathbf{r}\circ\cos\mathbf{Y}, \mathbf{r}\circ\sin\mathbf{Y}); \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb})$$
$$\times f_{\mathcal{N}}(\mathbf{X}_a; \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top(\mathbf{r}\circ\cos\mathbf{Y}) + \boldsymbol{\beta}_2^\top(\mathbf{r}\circ\sin\mathbf{Y}), \mathbf{Q})\prod_{d=1}^{D} r_d.$$
$$(12)$$

Given the independence assumption between the directional variables and that $\cos Y_d$ and $\sin Y_d$ are orthogonal with the same variance $\sigma_d^2 = \dfrac{1}{\kappa_d}$

$$\boldsymbol{\Sigma}_{bb} = \begin{pmatrix} \boldsymbol{\Sigma}_{b1b1} & \boldsymbol{\Sigma}_{b1b2} \\ \boldsymbol{\Sigma}_{b2b1} & \boldsymbol{\Sigma}_{b2b2} \end{pmatrix}$$

is a diagonal matrix such that

$$\boldsymbol{\Sigma}_{b1b1} = \boldsymbol{\Sigma}_{b2b2} = \begin{pmatrix} \dfrac{1}{\kappa_d} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \dfrac{1}{\kappa_D} \end{pmatrix},$$

and $\boldsymbol{\Sigma}_{b1b2} = \boldsymbol{\Sigma}_{b2b1} = \mathbf{0}$.

The last step to construct the Extended Mardia-Sutton distribution is to restrict the distribution over the directional variables to the unit circle. For this purpose, we condition Equation (12) so that $\mathbf{r} = \mathbf{1}$ obtaining

$$f(\mathbf{X}_a, \mathbf{Y}|\mathbf{r} = \mathbf{1}) = f(\mathbf{X}_a|\mathbf{Y}, \mathbf{r} = \mathbf{1})f(\mathbf{Y}|\mathbf{r} = \mathbf{1}).$$

The expression $f(\mathbf{X}_a|\mathbf{Y}, \mathbf{r} = \mathbf{1})$ is obtained from the conditional multivariate normal distribution given in Equation (12)

$$f(\mathbf{X}_a|\mathbf{Y}, \mathbf{r} = \mathbf{1}) = f_{\mathcal{N}}(\mathbf{X}_a; \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top\cos\mathbf{Y} + \boldsymbol{\beta}_2^\top\sin\mathbf{Y}, \mathbf{Q}).$$
$$(13)$$

The computation of the expression $f(\mathbf{Y}|\mathbf{r} = \mathbf{1})$ is not immediate and has to be obtained using the Bayes' theorem, i.e., $f(\mathbf{Y}|\mathbf{r} = \mathbf{1}) = \dfrac{f(\mathbf{Y}, \mathbf{r}=\mathbf{1})}{f(\mathbf{r}=\mathbf{1})}$. The numerator is computed according to

$$f(\mathbf{Y}, \mathbf{r} = \mathbf{1}) = \frac{1}{|2\pi\boldsymbol{\Sigma}_{bb}|^{1/2}}$$
$$\cdot e^{-\frac{1}{2}\sum_{d=1}^{D}\left[(\cos Y_d - \cos\mu_d)^2\kappa_d + (\sin Y_d - \sin\mu_d)^2\kappa_d\right]}$$
$$= \prod_{d=1}^{D} e^{\kappa_d\cos(Y_d - \mu_d)}\frac{\prod_{d=1}^{D} e^{-D\kappa_d}}{|2\pi\boldsymbol{\Sigma}_{bb}|^{1/2}}. \quad (14)$$

The normalisation term $f(\mathbf{r} = \mathbf{1})$ is obtained by marginalizing $\mathbf{Y}$ in Equation (14) and applying the modified Bessel function (see Equation (7)):

$$f(\mathbf{r} = \mathbf{1}) = \int_{\mathbf{Y}} \prod_{d=1}^{D} e^{\kappa_d\cos(Y_d - \mu_d)}\frac{\prod_{d=1}^{D} e^{-D\kappa_d}}{|2\pi\boldsymbol{\Sigma}_{bb}|^{1/2}}d\mathbf{Y}$$
$$= \prod_{d=1}^{D} 2\pi I_0(\kappa_d)\frac{\prod_{d=1}^{D} e^{-D\kappa_d}}{|2\pi\boldsymbol{\Sigma}_{bb}|^{1/2}}. \quad (15)$$

Thus, from Equation (14) and Equation (15) we have

$$f(\mathbf{Y}|\mathbf{r}=\mathbf{1}) = \prod_{d=1}^{D} \frac{e^{\kappa_d \cos(Y_d - \mu_d)}}{2\pi I_0(\kappa_d)}$$

$$= \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d, \kappa_d). \quad (16)$$

Finally, the Extended Mardia-Sutton distribution among liner variables $\mathbf{X}_a$ and directional variables $\mathbf{Y}$ is defined by the product of Equation (13) and Equation (16) as

$$f_{\mathcal{EMS}}(\mathbf{X}_a, \mathbf{Y}; \boldsymbol{\beta}, \mathbf{Q}, \boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\kappa}_{\mathbf{Y}})$$
$$= f(\mathbf{X}_a, \mathbf{Y}|\mathbf{r}=\mathbf{1})$$
$$= f(\mathbf{X}_a|\mathbf{Y}, \mathbf{r}=\mathbf{1})f(\mathbf{Y}|\mathbf{r}=\mathbf{1}) = \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d, \kappa_d)$$
$$\cdot f_{\mathcal{N}}(\mathbf{X}_a; \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top \cos \mathbf{Y} + \boldsymbol{\beta}_2^\top \sin \mathbf{Y}, \mathbf{Q}), \quad (17)$$

where

$$\boldsymbol{\beta}_0 = \boldsymbol{\mu}_{\mathbf{X}_a} - \boldsymbol{\beta}_1^\top \cos \boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\beta}_2^\top \sin \boldsymbol{\mu}_{\mathbf{Y}},$$
$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{ab1} \boldsymbol{\Sigma}_{b1b1}^{-1},$$
$$\boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{ab2} \boldsymbol{\Sigma}_{b2b2}^{-1},$$
$$\mathbf{Q} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab1} \boldsymbol{\Sigma}_{b1b1}^{-1} \boldsymbol{\Sigma}_{b1a} - \boldsymbol{\Sigma}_{ab2} \boldsymbol{\Sigma}_{b2b2}^{-1} \boldsymbol{\Sigma}_{b2a},$$
$$\boldsymbol{\Sigma}_{ab1} = \text{cov}(\mathbf{X}_a, \cos \mathbf{Y}),$$
$$\boldsymbol{\Sigma}_{ab2} = \text{cov}(\mathbf{X}_a, \sin \mathbf{Y}).$$

APPENDIX B
KL DIVERGENCE OF THE EXTENDED MARDIA-SUTTON DISTRIBUTION

The Kullback-Leibler divergence or relative entropy between two Extended Mardia-Sutton distributions P and Q is defined as

$$D_{\text{KL}}(P(\mathbf{X}, \mathbf{Y})||Q(\mathbf{X}, \mathbf{Y})) = \mathbb{E}_P \left(\log \frac{P(\mathbf{X}, \mathbf{Y})}{Q(\mathbf{X}, \mathbf{Y})} \right), \quad (18)$$

which can be factorised according to Equation (17) as

$$D_{\text{KL}}(P(\mathbf{X}, \mathbf{Y})||Q(\mathbf{X}, \mathbf{Y})) = \sum_{d=1}^{D} D_{\text{KL}}(P(Y_d)||Q(Y_d))$$
$$+ D_{\text{KL}}(P(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})||Q(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})). \quad (19)$$

Hence, the KL divergence decomposes as a sum of independent KL divergences between univariate von Mises distributions and the KL divergence of the linear variables conditioned to the directional variables. In the next subsections, we derive the two types of KL terms.

A. KL DIVERGENCE FOR THE VON MISES DISTRIBUTION
We start by introducing the definitions of the modified Bessel function of the first kind and order n for a directional variable Y_d

$$I_n(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos(Y_d)} \cos(nY_d) dY_d \quad (20)$$

and the ratio between modified Bessel functions of order one and order zero

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}. \quad (21)$$

Assume that we have two univariate von Mises distributions (Equation (6)) for the directional variable Y_d, i.e.,

$$P(Y_d) = f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)$$

and

$$Q(Y_d) = f_{\mathcal{VM}}(Y_d; \mu_d^Q, \kappa_d^Q).$$

For the sake of simplicity, we rotate the directional variable Y_d and the distributions $P(Y_d)$ and $Q(Y_d)$ according to μ_d^P. The directional variable after the rotation is defined as $Y_d^* = Y_d - \mu_d^P$, and the rotated distributions p and q are defined as

$$p(Y_d^*) = f_{\mathcal{VM}}(Y_d^*; \mu_d^p, \kappa_d^p)$$

and

$$q(Y_d^*) = f_{\mathcal{VM}}(Y_d^*; \mu_d^q, \kappa_d^q),$$

where the means of both distributions are $\mu_d^p = \mu_d^P - \mu_d^P = 0$ and $\mu_d^q = \mu_d^Q - \mu_d^P$ and the concentration parameters are obtained from $\kappa_d^p = \kappa_d^P$ and $\kappa_d^q = \kappa_d^Q$. Note that the rotation does not change the concentration of the distributions. Then, the KL divergence between univariate vM distributions is given by

$$D_{\text{KL}}(p(Y_d^*)||q(Y_d^*))$$
$$= \int_0^{2\pi} p(Y_d^*) \log \frac{p(Y_d^*)}{q(Y_d^*)} dY_d^*$$
$$= \int_0^{2\pi} p(Y_d^*) \log \frac{e^{\kappa_d^p \cos(Y_d^*)}}{I_0(\kappa_d^p)} \frac{I_0(\kappa_d^q)}{e^{\kappa_d^q \cos(Y_d^* - \mu_d^q)}} dY_d^*.$$

Simplifying and applying the logarithm, we obtain

$$\int_0^{2\pi} p(Y_d^*) \left[\log I_0(\kappa_d^q) - \log I_0(\kappa_d^p) + \kappa_d^p \cos(Y_d^*) \right.$$
$$\left. - \kappa_d^q \cos(Y_d^* - \mu_d^q) \right] dY_d^*.$$

Thus, we have to compute four integrals. The first integral is:

$$\int_0^{2\pi} p(Y_d^*) \log I_0(\kappa_d^q) dY_d^* = \log I_0(\kappa_d^q) \int_0^{2\pi} p(Y_d^*) dY_d^*$$
$$= \log I_0(\kappa_d^q),$$

as $\int_0^{2\pi} p(Y_d^*) dY_d^* = 1$. The second integral is obtained similarly:

$$\int_0^{2\pi} p(Y_d^*) \log I_0(\kappa_d^p) dY_d^* = \log I_0(\kappa_d^p) \int_0^{2\pi} p(Y_d^*) dY_d^*$$
$$= \log I_0(\kappa_d^p).$$

To compute the third integral, we use Equations (7) and (21):

$$\int_0^{2\pi} p(Y_d^*)\kappa_d^p \cos(Y_d^*)dY_d^*$$
$$= \frac{\kappa_d^p}{2\pi I_0(\kappa_d^P)} \int_0^{2\pi} e^{\kappa_d^p \cos(Y_d^*)} \cos(Y_d^*)dY_d^* = \kappa_d^p A(\kappa_d^P).$$

To compute the last integral, we again apply Equations (7) and (21):

$$\int_0^{2\pi} p(Y_d^*)\kappa_d^q \cos(Y_d^* - \mu_d^q)dY_d^*$$
$$= \frac{\kappa_d^q}{2\pi I_0(\kappa_d^P)}$$
$$\int_0^{2\pi} e^{\kappa_d^p \cos(Y_d^*)} \cos(Y_d^* - \mu_d^q)dY_d^*$$
$$= \frac{\kappa_d^q}{2\pi I_0(\kappa_d^P)}$$
$$\int_0^{2\pi} e^{\kappa_d^p \cos(Y_d^*)}(\cos(Y_d^*)\cos(\mu_d^q) + \sin(Y_d^*)\sin(\mu_d^q))dY_d^*$$
$$= \frac{\kappa_d^p}{2\pi I_0(\kappa_d^P)}\left[\cos(\mu_d^q)\int_0^{2\pi} e^{\kappa_d^p \cos(Y_d^*)}\cos(Y_d^*)dY_d^*\right.$$
$$\left. + \sin(\mu_d^q)\int_0^{2\pi} e^{\kappa_d^p \cos(Y_d^*)}\sin(Y_d^*)dY_d^*\right]$$
$$= \frac{\kappa_d^q}{I_0(\kappa_d^P)}\cos(\mu_d^q)I_1(\kappa_d^p) = \kappa_d^q \cos(\mu_d^q)A(\kappa_d^P).$$

Finally, joining the results of all the integrals, we obtain the closed-form

$$D_{\mathrm{KL}}(p(Y_d^*)||q(Y_d^*)) = \log I_0(\kappa_d^q) - \log I_0(\kappa_d^p)$$
$$+ A(\kappa_d^p)\left(\kappa_d^p - \kappa_d^q \cos(\mu_d^q)\right). \quad (22)$$

B. CONDITIONAL KL DIVERGENCE OF THE EXTENDED MARDIA-SUTTON DISTRIBUTION

The KL divergence of the linear variables conditioned to the directional variables between the Extended Mardia-Sutton distributions P and Q is defined as

$$D_{\mathrm{KL}}(P(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}})||Q(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}}))$$
$$= \int_{\mathbf{Y}} \prod_{d=1}^{D} P(Y_d)D_{\mathrm{KL}}(P(\mathbf{X}|\mathbf{Y})||Q(\mathbf{X}|\mathbf{Y}))d\mathbf{Y}. \quad (23)$$

Given that the linear variables are distributed according to a multivariate normal distribution (see Equation (17)), the multivariate normal KL divergence can be computed according to the well-known equation

$$D_{\mathrm{KL}}(P(\mathbf{X}|\mathbf{Y})||Q(\mathbf{X}|\mathbf{Y}))$$
$$= \frac{1}{2}\left[\mathrm{Tr}(\boldsymbol{\Sigma}^{-1,Q}\boldsymbol{\Sigma}^P) + (\boldsymbol{\mu}^Q - \boldsymbol{\mu}^P)^\top \boldsymbol{\Sigma}^{-1,Q}(\boldsymbol{\mu}^Q - \boldsymbol{\mu}^P)\right.$$
$$\left. - L + \ln\frac{|\boldsymbol{\Sigma}^Q|}{|\boldsymbol{\Sigma}^P|}\right], \quad (24)$$

where L is the number of linear variables.

There are four additive terms. The first, the third and the fourth terms are constant with respect to $\mathbf{Y}$, so

$$\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)$$
$$\cdot \frac{1}{2}\left[\mathrm{Tr}(\boldsymbol{\Sigma}^{-1,Q}\boldsymbol{\Sigma}^P) - L + \ln\frac{|\boldsymbol{\Sigma}^Q|}{|\boldsymbol{\Sigma}^P|}\right]d\mathbf{Y}$$
$$= \frac{1}{2}\left[\mathrm{Tr}(\boldsymbol{\Sigma}^{-1,Q}\boldsymbol{\Sigma}^P) - L + \ln\frac{|\boldsymbol{\Sigma}^Q|}{|\boldsymbol{\Sigma}^P|}\right]$$

Let us define $\boldsymbol{\mu}^R = \boldsymbol{\mu}^Q - \boldsymbol{\mu}^P$ as

$$\boldsymbol{\mu}_R = (\boldsymbol{\beta}_0^Q - \boldsymbol{\beta}_0^P) + (\boldsymbol{\beta}_1^Q - \boldsymbol{\beta}_1^P)^\top \cos\mathbf{Y}$$
$$+ (\boldsymbol{\beta}_2^Q - \boldsymbol{\beta}_2^P)^\top \sin\mathbf{Y} = \boldsymbol{\beta}_0^R + \boldsymbol{\beta}_1^{R\top} \cos\mathbf{Y} + \boldsymbol{\beta}_2^{R\top} \sin\mathbf{Y}.$$

The second term in the multivariate normal KL divergence (see Equation (24)) is a quadratic form that can be written as

$$\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)\left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\mu_i^R \mu_j^R\right)d\mathbf{Y}$$
$$= \int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)\left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}(\beta_{0i}^R + \boldsymbol{\beta}_{1i}^{R\top} \cos\mathbf{Y}\right.$$
$$\left. + \boldsymbol{\beta}_{2i}^{R\top} \sin\mathbf{Y})(\beta_{0j}^R + \boldsymbol{\beta}_{1j}^{R\top} \cos\mathbf{Y} + \boldsymbol{\beta}_{2j}^{R\top} \sin\mathbf{Y})\right)d\mathbf{Y}.$$

where $\Sigma_{ij}^{-1,Q}$ is the element at the i-th row and j-th column in $\boldsymbol{\Sigma}^{-1,Q}$ and μ_i^R and μ_j^R are the i-th and j-th components of vector $\boldsymbol{\mu}^R$. Then, we compute the integrals over each additive term, applying Equation (7) and some well-known trigonometric identities to yield

$$\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)\left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\beta_{0i}^R \beta_{0j}^R\right)d\mathbf{Y}$$
$$= \sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\beta_{0i}^R \beta_{0j}^R,$$

$$\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)$$
$$\cdot \left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\beta_{0i}^R(\boldsymbol{\beta}_{1j}^{R\top} \cos\mathbf{Y})\right)d\mathbf{Y}$$
$$= \sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\beta_{0i}^R \sum_{d=1}^{D} \beta_{1jd}^R A(\kappa_d^P),$$

$$\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)$$
$$\cdot \left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\beta_{0j}^R(\boldsymbol{\beta}_{1i}^{R\top} \cos\mathbf{Y})\right)d\mathbf{Y}$$
$$= \sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q}\beta_{0j}^R \sum_{d=1}^{D} \beta_{1id}^R A(\kappa_d^P),$$

$$
\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)
$$
$$
\cdot \left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q} (\boldsymbol{\beta}_{1i}^{R\top} \cos \mathbf{Y})(\boldsymbol{\beta}_{1j}^{R\top} \cos \mathbf{Y}) \right) d\mathbf{Y}
$$
$$
= \sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q} \left(\sum_{d=1}^{D} \frac{\beta_{1id}^{R} \beta_{1jd}^{R}}{2} \cdot \left(1 + \frac{I_2(\kappa_d^P)}{I_0(\kappa_d^P)} \right) \right.
$$
$$
\left. + \sum_{d=1}^{D} \sum_{m \neq d}^{D} \beta_{1id}^{R} \beta_{1jm}^{R} A(\kappa_d^P) A(\kappa_m^P) \right),
$$

and

$$
\int_{\mathbf{Y}} \prod_{d=1}^{D} f_{\mathcal{VM}}(Y_d; \mu_d^P, \kappa_d^P)
$$
$$
\cdot \left(\sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q} (\boldsymbol{\beta}_{2i}^{R\top} \sin \mathbf{Y})(\boldsymbol{\beta}_{2j}^{R\top} \sin \mathbf{Y}) \right) d\mathbf{Y}
$$
$$
= \sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q} \sum_{d=1}^{D} \frac{\beta_{2id}^{R} \beta_{2jd}^{R}}{2} \left(1 - \frac{I_2(\kappa_d^P)}{I_0(\kappa_d^P)} \right).
$$

We omit those terms whose result of solving the integral was always zero. Finally, grouping all of the terms in one equation, we obtain the expression for the conditional KL divergence

$$
D_{\mathrm{KL}}(P(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}}) || Q(\mathbf{X}|\mathbf{Pa}_{\mathbf{X}}^{\mathcal{G}}))
$$
$$
= \frac{1}{2} \sum_{i,j=1}^{L} \Sigma_{ij}^{-1,Q} \Big[\beta_{0i}^{R} \beta_{0j}^{R}
$$
$$
+ 2\beta_{0i}^{R} \sum_{d=1}^{D} \beta_{1jd}^{R} A(\kappa_d^P) + \sum_{d=1}^{D} \frac{\beta_{1id}^{R} \beta_{1jd}^{R}}{2} \left(1 + \frac{I_2(\kappa_d^P)}{I_0(\kappa_d^P)} \right)
$$
$$
+ \sum_{d=1}^{D} \sum_{m \neq d}^{D} \beta_{1id}^{R} \beta_{1jm}^{R} A(\kappa_d^P) A(\kappa_m^P)
$$
$$
+ \sum_{d=1}^{D} \frac{\beta_{2id}^{R} \beta_{2jd}^{R}}{2} \left(1 - \frac{I_2(\kappa_d^P)}{I_0(\kappa_d^P)} \right) \Big]
$$
$$
+ \frac{1}{2} \left[\mathrm{Tr}(\boldsymbol{\Sigma}^{-1,Q} \boldsymbol{\Sigma}^{P}) - L + \ln \frac{|\boldsymbol{\Sigma}^{Q}|}{|\boldsymbol{\Sigma}^{P}|} \right].
$$

REFERENCES

[1] J. A. Carta, P. Ramírez, and C. Bueno, "A joint probability density function of wind speed and direction for wind energy analysis," *Energy Convers. Manage.*, vol. 49, no. 6, pp. 1309–1320, 2008.

[2] 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," *Int. J. Chronobiol.*, vol. 1, no. 3, pp. 183–202, 1973.

[3] E. Batschelet, *Circular Statistics in Biology*. London, U.K.: Academic, 1981.

[4] I. Leguey, "Directional-linear Bayesian networks and applications in neuroscience," Ph.D. dissertation, Univ. Politécnica de Madrid, Madrid, Spain, 2018.

[5] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Statist. Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.

[6] G. McLachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering*. Hoboken, NJ, USA: Wiley, 1988.

[7] V. Melnykov and R. Maitra, "Finite mixture models and model-based clustering," *Stat. Survey*, vol. 4, pp. 80–116, Apr. 2010.

[8] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2000.

[9] D. M. Titterington, A. Smith, and U. Makoy, *Statistical Analysis of Finite Mixture Distributions*. Hoboken, NJ, USA: Wiley, 1985.

[10] S. Luengo-Sanchez, C. Bielza, and P. Larrañaga, "Hybrid Gaussian and von Mises model-based clustering," in *Proc. 22nd Eur. Conf. Artif. Intell. (ECAI)*, 2016, pp. 855–862.

[11] A. Roy, S. K. Parui, and U. Roy, "SWGMM: A semi-wrapped Gaussian mixture model for clustering of circular–linear data," *Pattern Anal. Appl.*, vol. 19, no. 3, pp. 631–645, 2016.

[12] K. Mardia and P. Jupp, *Directional Statistics*. Hoboken, NJ, USA: Wiley, 1999.

[13] S. R. Jammalamadaka and A. Sengupta, *Topics in Circular Statistics*. Cleveland, OH, USA: World Scientific, 2001.

[14] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Sep. 2005.

[15] K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh, "A multivariate von Mises distribution with applications to bioinformatics," *Can. J. Statist.*, vol. 36, no. 1, pp. 99–109, 2008.

[16] G. Mastrantonio, A. Maruotti, and G. Jona-Lasinio, "Bayesian hidden Markov modelling using circular-linear general projected normal distribution," *Environmetrics*, vol. 26, no. 2, pp. 145–158, 2015.

[17] K. V. Mardia and T. W. Sutton, "A model for cylindrical variables with applications," *J. Roy. Stat. Soc., B*, vol. 40, no. 2, pp. 229–233, 1978.

[18] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, "Topology matching for fully automatic similarity estimation of 3D shapes," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, 2001, pp. 203–212.

[19] J. W. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools Appl.*, vol. 39, no. 3, pp. 441–471, Sep. 2008.

[20] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann, 1988.

[21] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[22] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, 2006.

[23] J. A. Gámez, J. L. Mateo, and J. M. Puerta, "Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood," *Data Mining Knowl. Discovery*, vol. 22, nos. 1–2, pp. 106–148, 2011.

[24] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, 1992.

[25] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B*, vol. 36, no. 1, pp. 1–22, 1977.

[27] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. Hoboken, NJ, USA: Wiley, 2008.

[28] N. Friedman, "Learning belief networks in the presence of missing values and hidden variables," in *Proc. 14th Int. Conf. Mach. Learn. (ICML)*. San Mateo, CA, USA: Morgan Kaufmann, 1997, pp. 125–133.

[29] R. A. Johnson and T. E. Wehrly, "Some angular-linear distributions and related regression models," *J. Amer. Stat. Assoc.*, vol. 73, no. 363, pp. 602–606, 1978.

[30] T. Abe and C. Ley, "A tractable, parsimonious and flexible model for cylindrical data, with applications," *Econometrics Statist.*, vol. 4, pp. 91–104, Oct. 2017.

[31] S. Kato and K. Shimizu, "Dependent models for observations which include angular ones," *J. Stat. Planning Inference*, vol. 138, no. 11, pp. 3538–3549, 2008.

[32] K. V. Mardia, J. T. Kent, Z. Zhang, C. C. Taylor, and T. Hamelryck, "Mixtures of concentrated multivariate sine distributions with applications to bioinformatics," *J. Appl. Statist.*, vol. 39, no. 11, pp. 2475–2492, 2012.

[33] A. Roy, A. Pal, and U. Garain, "JCLMM: A finite mixture model for clustering of circular-linear data and its application to psoriatic plaque segmentation," *Pattern Recognit.*, vol. 66, pp. 160–173, Jun. 2017.

[34] F. Lagona, A. Maruotti, and M. Picone, "A non-homogeneous hidden Markov model for the analysis of multi-pollutant exceedances data," in *Hidden Markov Models, Theory and Applications*, P. Dymarski, Ed. Rijeka, Croatia: InTech, 2011, ch. 10, pp. 207–222.

[35] F. Lagona and M. Picone, "A latent-class model for clustering incomplete linear and circular data in marine studies," *J. Data Sci.*, vol. 9, no. 4, pp. 585–605, 2011.

[36] J. Bulla, F. Lagona, A. Maruotti, and M. Picone, "A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series," *J. Agricult., Biol., Environ. Statist.*, vol. 17, no. 4, pp. 544–567, 2012.

[37] D. J. Best and N. I. Fisher, "The BIAS of the maximum likelihood estimators of the von Mises-Fisher concentration parameters," *Commun. Statist.-Simul. Comput.*, vol. 10, no. 5, pp. 493–502, 1981.

[38] H. Sun and S. Wang, "Measuring the component overlapping in the Gaussian mixture model," *Data Mining Knowl. Discovery*, vol. 23, no. 3, pp. 479–502, 2011.

[39] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[41] K. Svoboda, "The past, present, and future of single neuron reconstruction," *Neuroinformatics*, vol. 9, nos. 2–3, pp. 97–98, 2011.

[42] B. Mihaljević, R. Benavides-Piccione, L. Guerra, J. DeFelipe, P. Larrañaga, and C. Bielza, "Classifying GABAergic interneurons with semi-supervised projected model-based clustering," *Artif. Intell. Med.*, vol. 65, no. 1, pp. 49–59, 2015.

[43] R. Benavides-Piccione, F. Hamzei-Sichani, I. Ballesteros-Yáñez, J. DeFelipe, and R. Yuste, "Dendritic size of pyramidal neurons differs among mouse cortical regions," *Cerebral Cortex*, vol. 16, no. 7, pp. 990–1001, 2005.

[44] A. Karagiannis, T. Gallopin, C. Dávid, D. Battaglia, H. Geoffroy, J. Rossier, E. M. Hillman, J. F. Staiger, and B. Cauli, "Classification of NPY-expressing neocortical interneurons," *J. Neurosci.*, vol. 29, no. 11, pp. 3642–3659, 2009.

[45] 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*, vol. 23, no. 8, pp. 1798–1810, 2013.

[46] S. Luengo-Sanchez, C. Bielza, R. Benavides-Piccione, I. Fernaud-Espinosa, J. DeFelipe, and P. Larrañaga, "A univocal definition of the neuronal soma morphology using Gaussian mixture models," *Frontiers Neuroanatomy*, vol. 9, p. 137, Nov. 2015.

[47] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. 4th Eurograph. Symp. Geometry Process.*, 2006, pp. 61–70.

[48] L. Shapira, A. Shamir, and D. Cohen-Or, "Consistent mesh partitioning and skeletonisation using the shape diameter function," *Vis. Comput.*, vol. 24, no. 4, pp. 249–259, 2008.

[49] S. Luengo-Sanchez, I. Fernaud-Espinosa, C. Bielza, R. Benavides-Piccione, P. Larrañaga, and J. DeFelipe, "3D morphology-based clustering and simulation of human pyramidal cell dendritic spines," *PLOS Comput. Biol.*, vol. 14, no. 6, 2018, Art. no. e1006221.

[50] K. V. Mardia, "Statistics of directional data," *J. Roy. Stat. Soc., B*, vol. 37, no. 3, pp. 349–393, 1975.

[51] S.-Q. Xin and G.-J. Wang, "Improving Chen and Han's algorithm on the discrete geodesic problem," *ACM Trans. Graph.*, vol. 28, no. 4, 2009, Art. no. 104.

[52] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, nos. 1–2, pp. 28–35, 1947.

[53] G. S. Watson and E. J. Williams, "On the construction of significance tests on the circle and the sphere," *Biometrika*, vol. 43, nos. 3–4, pp. 344–352, 1956.

[54] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn. (ICML)*. San Mateo, CA, USA: Morgan Kaufmann, 1995, pp. 115–123.

[55] M. Di Marzio, A. Panzera, and C. C. Taylor, "Non-parametric regression for circular responses," *Scand. J. Statist.*, vol. 40, no. 2, pp. 238–255, 2013.

[56] H. G. Sung, "Gaussian mixture regression and classification," Ph.D. dissertation, Dept. Statist., Rice Univ., Houston, TX, USA, 2004.

SERGIO LUENGO-SANCHEZ received the B.Sc. degree in computer science from the Universidad de Alcalá de Henares, in 2013, and the M.Sc. degree in artificial intelligence from the Universidad Politécnica de Madrid (UPM), in 2014. He is currently pursuing the Ph.D. degree with the Artificial Intelligence Department, UPM, where he was involved in the Cajal Blue Brain Project. He also collaborates with the Human Brain Project, developing software modules.

PEDRO LARRAÑAGA received the M.Sc. degree in mathematics (statistics) from the University of Valladolid and the Ph.D. degree in computer science from the University of the Basque Country (excellence award). He has been a Full Professor in computer science and artificial intelligence with the Universidad Politécnica de Madrid (UPM), since 2007. Before moving to UPM, his academic career developed at the University of the Basque Country (UPV-EHU) at several faculty ranks: Assistant Professor, from 1985 to 1998, Associate Professor, from 1998 to 2004, and a Full Professor, from 2004 to 2007. He received the distinction for Full Professor, in 2003. He has published over 200 papers in high-impact factor journals, He has supervised over 25 Ph.D. theses. His research interests include the areas of probabilistic graphical models, metaheuristics for optimization, data mining, classification models, and real applications, such as biomedicine, bioinformatics, neuroscience, industry 4.0, and sports. He is Fellow of the European Association for Artificial Intelligence, since 2012, and a Fellow of the Academia Europaea, since 2018. He has received the 2013 Spanish National Prize in computer science and the prize of the Spanish Association for Artificial Intelligence, in 2018.

CONCHA BIELZA received the M.Sc. degree in mathematics from the Universidad Complutense de Madrid, in 1989, and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid (UPM), in 1996 (extraordinary doctorate award). She has been a Full Professor of statistics and operations research with the Departamento de Inteligencia Artificial, UPM, since 2010. Her research interests include primarily in the areas of probabilistic graphical models, decision analysis, metaheuristics for optimization, data mining, classification models, and real applications, such as biomedicine, bioinformatics, neuroscience, industry, and sports. She has published over 100 papers in hifh impact factor journals. She has supervised 13 Ph.D. theses. She has received the 2014 UPM Research Prize.

• • •