Regularized Multivariate von Mises Distribution

Luis Rodriguez-Lujan, Concha Bielza, and Pedro Larrañaga

Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid,
Campus de Monteganceda, 28660 Boadilla del Monte, Madrid, Spain
luis.rodriguezl@alumnos.upm.es
{mcbielza,pedro.larranaga}@fi.upm.es
http://cig.fi.upm.es

Abstract. Regularization is necessary to avoid overfitting when the number of data samples is low compared to the number of parameters of the model. In this paper, we introduce a flexible L_1 regularization for the multivariate von Mises distribution. We also propose a circular distance that can be used to estimate the Kullback-Leibler divergence between two circular distributions by means of sampling, and also serves as goodness-of-fit measure. We compare the models on synthetic data and real morphological data from human neurons and show that the regularized model achieves better results than non regularized von Mises model.

1 Introduction

Directional data is ubiquitous in science, from the direction of the wind to the branching angles of the trees. However, directional data has been traditionally treated as regular linear data despite of its different nature. Directional statistics [4] provides specific tools for modelling directional data. If the normal distribution is the most famous distribution for linear data, the von Mises distribution [6] is its analogue for directional data.

If we extend the von Mises as a multivariate distribution [3], we face the problem that no closed formulation is known for the normalization term when the number of variables is greater than two, and, therefore it cannot be easily fitted nor compared to other distributions. We introduce a computationally optimized version of the full pseudo-likelihood as well as a circular distance to address these problems.

Another problem in some application areas, like neuroscience, is that data is scarce and expensive. In these situations regularization is needed to prevent overfitting. We propose a L_1 regularization for the multivariate von Mises distribution that allows us to introduce prior beliefs on the relation between the variables.

This paper is organized as follows. Section 2 reviews the univariate and multivariate von Mises distributions. In Sect. 3 we propose a circular distance that is applied to estimate the KL divergence between two distributions. Then, in Sect. 4 we compare the von Mises distribution to the Gaussian distribution over synthetic data using the approximated KL divergence as the evaluation metric. We repeat the same process on real data from human neurons in Sect. 5, this time using the approximated KL-divergence as a two-sample test, and show that the regularized multivariate von Mises distribution always achieves better results. We conclude the paper in Sect. 6 with a final discussion and some proposals for future work.

2 The Multivariate von Mises Distribution

Directional statistics is a field within statistics that deals with angles, or equivalently, directions in space. Among the variety of directional distributions, the von Mises distribution is particularly noteworthy since it is considered the circular analogue of the normal distribution but having better mathematical properties than the wrapped-normal distribution [3,5,8].

The univariate von Mises distribution belongs to the exponential family and its density function is given by:

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

where μ is the mean angle and κ the concentration parameter, i.e. the inverse of the variance, and I_0 is the modified Bessel function of order 0.

Based on its exponential definition, we can define a multivariate von Mises distribution [3] analogous to the multivariate normal distribution. For $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ the density function is defined as:

$$f_{MVM}(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\Lambda}) = \frac{1}{Z(\boldsymbol{\kappa},\boldsymbol{\Lambda})} \exp\{\boldsymbol{\kappa}\cos(\boldsymbol{\theta}-\boldsymbol{\mu})^T + \frac{1}{2}\sin(\boldsymbol{\theta}-\boldsymbol{\mu})\boldsymbol{\Lambda}\sin(\boldsymbol{\theta}-\boldsymbol{\mu})^T\} \tag{2}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ and $\boldsymbol{\kappa} = (\kappa_1, ..., \kappa_p)$ are the multidimensional equivalents of μ and κ in the univariate von Mises respectively, and $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a $p \times p$ symmetric matrix with $\lambda_{ii} = 0$ and $\lambda_{ij} \geq 0$.

Unfortunately, the normalization term $Z(\boldsymbol{\kappa},\boldsymbol{\Lambda})$ does not have a known closed-form formula for any p greater than two, so it has to be approximated numerically, making the calculation of the density function intractable computationally.

2.1 Pseudo-Likelihood

Due to the complexity of computing the normalization term in the density function of Eq. (2) it is not practical to use the likelihood as the target function to fit the multivariate von Mises distribution given a set of data samples [3]. In this same article, the authors propose to use the pseudo-likelihood as a consistent

approximation of the likelihood term. Since each marginal conditional term for the multivariate von Mises is a univariate von Mises, the full pseudo-likelihood for a p-dimensional $\boldsymbol{\theta} = (\theta_{i,j})$ that contains N independent samples can be expressed as:

$$P\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = (2\pi)^{-Np} \prod_{i=1}^{N} \prod_{j=1}^{p} \frac{1}{I_0(\kappa_j^i)} \exp\left\{\kappa_j^i \cos(\theta_{i,j} - \mu_j^i)\right\} \tag{3}$$

where μ_j^i and κ_j^i are, respectively, the j-th marginal mean and concentration given the i-th data sample:

$$\mu_j^i = \mu_j + \arctan\left(\frac{\sum_{l \neq j} \lambda_{j,l} \sin(\theta_{i,l} - \mu_l)}{\kappa_j}\right) \tag{4}$$

$$\kappa_j^i = \sqrt{\kappa_j^2 + (\sum_{l \neq j} \lambda_{j,l} \sin(\theta_{i,l} - \mu_l))^2} \tag{5}$$

2.2 Optimization

Given a set of samples, to compute the parameters of the multivariate von Mises distribution that maximize the pseudo-likelihood we define a minimization problem where the loss function is minus the natural logarithm of the pseudo-likelihood defined in (3). This loss function can be written as:

$$L(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = (pN)\log(2\pi) + \sum_{i=1}^{N} \sum_{j=1}^{p} \left\{\log(I_0(\kappa_j^i) - \kappa_j^i \cos(\theta_{i,j} - \mu_j^i)\right\} \tag{6}$$

We simplified the loss function (6) to reduce its complexity (from a computational point of view) by expressing sums as matrix products and by applying trigonometric properties to reduce the number of operations to be computed, specifically, to avoid the computation of the tangent inverse function. To do so, the first step is to define the $N \times p$ matrix $\boldsymbol{\Phi}$ as:

$$\boldsymbol{\Phi} = \sin(\boldsymbol{\theta} - \boldsymbol{\mu})\boldsymbol{\Lambda}$$

Then, we can reduce the second term in the sum using some simple trigonometric identities:

$$\begin{aligned} \kappa_j^i \cos(\theta_{i,j} - \mu_j^i) &= \kappa_j^i(\cos(\theta_{i,j})\cos(\mu_j^i) + \sin(\theta_{i,j})\sin(\mu_j^i)) \\ &= \kappa_j^i \left(\frac{\cos(\theta_{i,j} - \mu_j)\kappa_j}{\kappa_j^i} + \frac{\sin(\theta_{i,j} - \mu_j)\phi_{i,j}}{\kappa_j^i}\right) \end{aligned}$$

As result we obtain a more compact version of the loss function, that do not require to compute the tangent inverse:

$$L_c(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = \sum_{i=1}^{N} \sum_{j=1}^{p} \left[\log(I_0(\kappa_j^i)) - \cos(\theta_{i,j} - \mu_j)\kappa_j - \sin(\theta_{i,j} - \mu_j)\phi_{i,j}\right] \tag{7}$$

To find the minima for function (7) we will use the quasi-newton L-BFGS-B algorithm [10] which is an extension of the well-known L-BFGS method that supports simple constraints such as $\kappa_j > 0$. This method only requires to evaluate the loss function and its partial derivatives. Since the optimal $\boldsymbol{\mu}$ parameter is the vector formed by the marginal means, only partial derivatives with respect to $\boldsymbol{\kappa}$ and $\boldsymbol{\Lambda}$ need to be computed in order to use this method. Please note that A_0 stands for $\frac{I_1}{I_0}$ where I_1 is the modified Bessel function of order 1.

$$\frac{\partial L_c}{\partial \kappa_j} = \sum_{i=1}^{N} \left[A_0(\kappa_j^i) \frac{\kappa_j}{\kappa_j^i} - \cos(\theta_{i,j} - \mu_j) \right] \tag{8}$$

$$\frac{\partial L_c}{\partial \lambda_{j,k}} = \sum_{i=1}^{N} \left[\sin(\theta_{i,k} - \mu_k) \left(\frac{A_0(\kappa_j^i)}{\kappa_j^i} \phi_{j,k} - \sin(\theta_{i,j} - \mu_j) \right) \right] \tag{9}$$

A comparison of the fitting execution time between the regular loss function provided by [8] based on the Eq. (6) and the optimized version in (7) was performed for a 5-dimensional von Mises distribution. Both methods were implemented in *ANSI C* and executed in similar conditions. The results in Fig. 1 show that the optimized version is significantly faster as the number of samples increase.

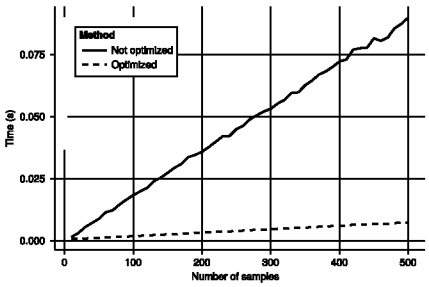


Fig. 1. Mean fitting time per number of samples of a 5-dimensional von Mises function. Each execution was repeated 100 times with 2 additional warm-up iterations.

2.3 Regularization

The regularized learning of the multivariate von Mises distribution has already been proposed by other authors [8]. However, from a Bayesian point of view, if we penalize equally all components in matrix $\boldsymbol{\Lambda}$ as it is done in the standard L_1 regularization, we are adding the prior belief that all components $\lambda_{i,j}$ are similar (in scale terms) which may not correspond with our previous knowledge of the problem as it is studied in [9] for the multivariate normal distribution.

To add prior knowledge about the structure we propose a generalized version of the L_1 penalization for the multivariate von Mises distribution where each

component $\lambda_{i,j}$ is individually weighted. To do so, a symmetric penalization matrix $\boldsymbol{\Psi}$ is defined with the only restriction that all elements should be positive. Then, the function to minimize is:

$$g(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = L_c(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) + \sum_{j=1}^{p} \sum_{k \neq j} |\lambda_{j,k}| \psi_{j,k} \tag{10}$$

Although the absolute value of $\lambda_{i,j}$ is not a differentiable at 0, we can find a differentiable function that approximates the absolute value with arbitrary precision. Bearing in mind that the number of real values that can be represented in a computer is finite, i.e. the minimum distance between real values is given by what is known as the machine epsilon, we can treat the absolute value function as if it was differentiable at any point. Then we just need to add a new term to the partial derivative with respect to $\lambda_{j,k}$:

$$\frac{\partial g}{\partial \lambda_{j,k}} = \frac{\partial L_c}{\partial \lambda_{j,k}} + sgn(\lambda_{j,k}) \psi_{j,k} \tag{11}$$

where sgn is the sign function that evaluates $sgn(0) = 0$.

3 Evaluation

The impossibility to express the normalization term of the multivariate von Mises distribution (12) as a closed formula for any p restrains the use of typical measures of divergence between distributions such as the Kullback-Leibler divergence since we cannot evaluate the density function in any point, which also impedes the use more powerful goodness of fit tests.

Other authors have used the angle between original and fitted parameters [8] or the pseudo-likelihood value [3] as evaluation metrics, but these approaches are either only applicable to synthetic data from a known distribution or rely on the approximated likelihood.

To overcome these drawbacks, we propose to use the approximation of the KL divergence for multivariate distributions [7] as evaluation measure. This approach takes two sets of samples as input (one from each distribution or one from the real data and the other sampled from the learned model) and uses the distance to the k-th nearest neighbor to approximate the KL divergence. Given two sets of samples $\{\boldsymbol{X_i}\}_{i=1}^{n}$ and $\{\boldsymbol{Y_i}\}_{i=1}^{m}$ from two p-dimensional distributions P and Q, the approximated KL divergence [7] between P and Q is computed as:

$$\hat{D}_k(P||Q) = \frac{p}{n} \sum_{i=1}^{n} \left[\log \left(\frac{r_k(\mathbf{x_i})}{s_k(\mathbf{x_i})} \right) \right] + \log \frac{m}{n-1} \tag{12}$$

where $r_k(\mathbf{x_i})$ and $s_k(\mathbf{x_i})$ are the distance to the k-th nearest neighbour of $\mathbf{x_i}$ in $\mathbf{X} \setminus \mathbf{x_i}$ and $\mathbf{Y}$ respectively.

To generate samples from the multivariate von Mises distribution we can either use a rejection sampling algorithm [5] for small or moderate p or use a

Gibbs sampler for higher p [8]. However, we still need to define a distance that computes the distance between two multivariate circular points. We defined a distance between two points $\boldsymbol{a}$,$\boldsymbol{b} \in [0, 2\pi)^p$ in Eq. (13) that takes into account the periodicity of circular data.

$$d(\boldsymbol{a}, \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}^*||_2$$

where:

$$\boldsymbol{b}^* = (b_i^*)_{i=1}^p = \begin{cases} b_i & \text{if } |a_i - b_i| \leq \pi \\ b_i + 2\pi & \text{if } |a_i - b_i| > \pi \text{ and } a_i > \pi \\ b_i - 2\pi & \text{if } |a_i - b_i| > \pi \text{ and } a_i \leq \pi \end{cases} \tag{13}$$

4 Multivariate von Mises vs. Multivariate Gaussian

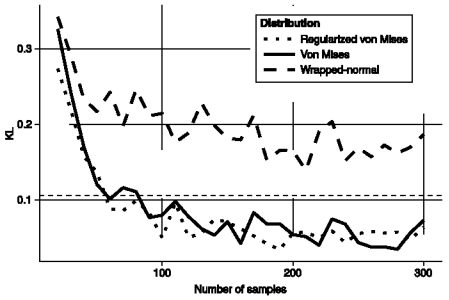
For high values of the concentration parameter, the univariate von Mises distribution approximates a normal distribution on the circumference. This behaviour extends to the multivariate case. However, it is not clear yet how this behaviour is affected either by the $\boldsymbol{\Lambda}$ parameter or the dimension p [8].

We used the empiric KL divergence defined in [7] along with the distance proposed in Sect. 3 to design a set of experiments with the aim of studying the behaviour of the multivariate normal and von Mises distribution when fitting circular data with different configurations. In addition, a regularized von Mises distribution is included in the comparison with penalization matrix $\boldsymbol{\Psi} = (\psi_{i,j}) = |i - j|$, which is similar to the $\boldsymbol{\Lambda}$ band matrix configuration in the experiments, and it also matches the penalization used in Sect. 5.

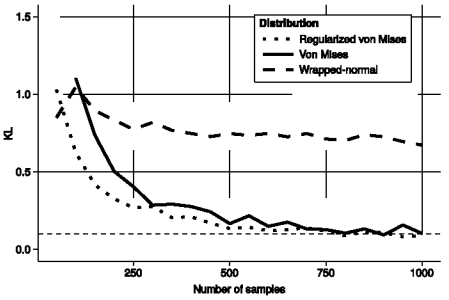
We generated random samples from different configurations of parameters, varying: (a) The number of variables from 4 to 50; (b) the number of samples from 10 to 500 in the simplest case (4 variables) and from 50 to 1000 in the most complex (50 variables); (c) the concentration vector $\boldsymbol{\kappa}$ from a vector where all values were equal to 0.1 to a vector where all values were 7.0; and (d) the $\boldsymbol{\Lambda}$ matrix from a very sparse configuration where all elements were equal to zero to a dense configuration where all elements were distinct of zero. For all configurations and variables the mean value $\boldsymbol{\mu}$ was fixed at π.

The procedure below is repeated for each combination of parameters 20 times to compute the approximate KL divergence as the average of all 20 results:

1. A set of N p-dimensional samples are generated from a multivariate von Mises distribution with parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\kappa}_0$ and $\boldsymbol{\Lambda_0}$
2. Multivariate normal and von Mises (regularized and non-regularized) distributions parameters are fitted from the N samples
3. Another set of m p-dimensional samples are generated from both original distribution and learned ones. Please note that m can be different from N. Then, the empiric approximation of the KL divergence is computed using these m samples



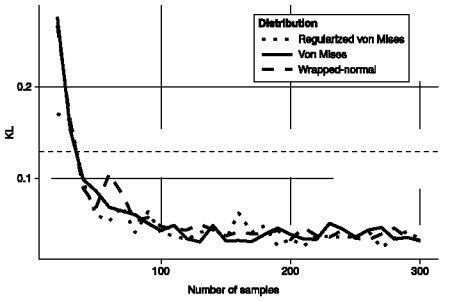
(a) $p = 4$, $m = 500$



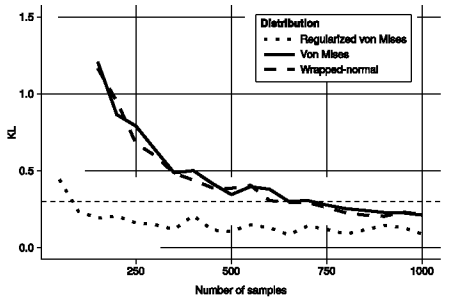
(b) $p = 50$, $m = 1000$

Fig. 2. Approximated KL divergence for low concentration $\boldsymbol{\kappa}$ and very sparse $\boldsymbol{\Lambda}$.

The results for a sparse $\boldsymbol{\Lambda}$ matrix with low concentration $\boldsymbol{\kappa}$ can be seen in Fig. 2. Both von Mises distributions obtain better results than the multivariate normal distribution. If the number of samples is high enough, the regularized and non regularized von Mises perform similarly. As expected, the number of samples needed to obtain the same fit in both regularized and non regularized distributions is higher as the number of variables increases.



(a) $p = 4$, $m = 500$

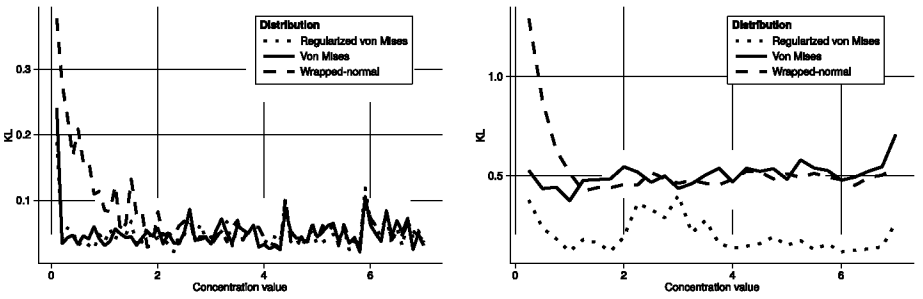


(b) $p = 50$, $m = 1000$

Fig. 3. Approximated KL divergence for high concentration $\boldsymbol{\kappa}$ and dense $\boldsymbol{\Lambda}$.

In Fig. 3 we can see the results for the opposite case, high concentration $\boldsymbol{\kappa}$ and dense $\boldsymbol{\Lambda}$ matrix. In this case with high concentration the multivariate normal distribution obtains similar or better results than the von Mises distribution. It is important to note that although the penalization matrix does not exactly match the real Λ_0 matrix structure, i.e. we do not have a perfect prior, the regularized version still performs equally or better than the non regularized one. In all cases we observe that the regularized von Mises distribution produces a better fit when the number of samples is low.

Plots in Fig. 4 depict how the variation of the concentration parameter affects each of the distributions under evaluation for a fairly high number of samples



(a) $p = 4$, $N = 100$, $m = 500$ (b) $p = 50$, $N = 300$, $m = 1000$

Fig. 4. Approximated KL divergence for concentration $\boldsymbol{\kappa}$ varying and banded $\boldsymbol{\Lambda}$.

($N = 100$ and $N = 300$ respectively). In Fig. 4a we observe that both versions of the multivariate von Mises obtain similar results, independently of the concentration, due to the high number of samples with respect to the number of variables. It is also interesting to note that as the concentration parameter increases, the normal distribution approaches the von Mises, getting similar KL values for $\kappa > 2.0$ in both Fig. 4a and b. From Fig. 4b we can preliminarily say that the improvement of the regularized distribution it is not affected by the concentration parameter $\boldsymbol{\kappa}$.

5 Validation on Morphological Data from Human Neurons

Neurons, the very basic component of the nervous system, can be divided into the cell body, dendrites and axon. Pyramidal cells is one of the most important types of neurons that have special basal dendrites, a set of dendrites that grows from the base of the cell body. In order to understand the differences between pyramidal neurons from different genders, species, brain regions, etc. it is important to characterize the grow direction of these basal dendrites, which also helps to simulate and understand the functionality of these neurons.

We downloaded a set of 3D reconstructions of human pyramidal neurons [2] from NeuroMorpho.Org [1], a public repository of neural reconstructions. The data includes gender and age, as well as other metadata related to the brain region or the reconstruction method. We restricted our selection to reconstructions from adults and with neurons belonging to the occipital lobe or the frontal lobe.

The original data in plain text format was parsed and the angles between dendrites were measured. To establish a criteria on variable ordering, the longest dendrite was taken as the principal and angles where numbered following a counter-clockwise ordering as depicted in Fig. 5. It is noteworthy that for a neuron with p dendrites, we have $p - 1$ angles since the last one is completely determined by the rest.

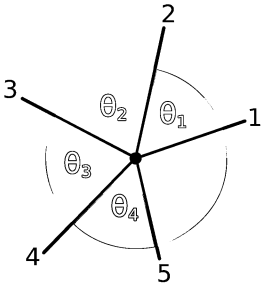


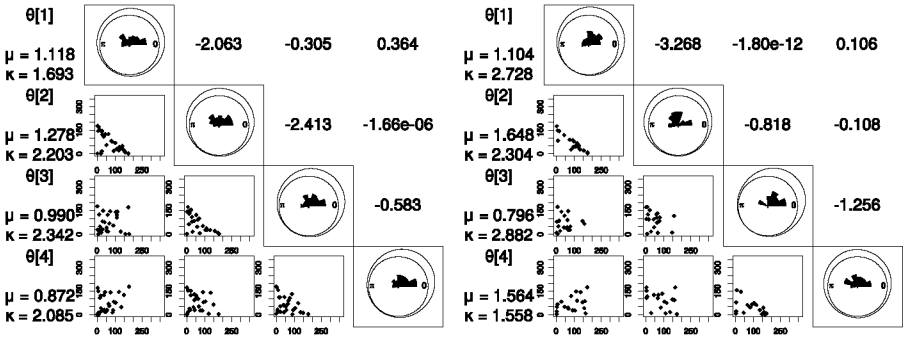
Fig. 5. Inter-dendrite angles.

We fitted a von Mises regularized distribution with a penalization matrix that grows with respect to the distance between angles. e.g. in Fig. 5 the value of $\psi_{1,4}$ is 2 (the shortest path between angles 1 and 4 is 1-5-4). We performed repeated train and test validation with 100 repetitions and use the empiric approximation KL divergence defined in Sect. 3 as evaluation metric. In each repetition, a 75 % of the original samples were selected at random as training set, leaving the remaining portion as the test set. In addition we also fitted a regular von Mises distribution for comparison purposes. Results are displayed in Table 1. In every case the regularized distribution obtains similar or better results.

Table 1. KL - divergence results for inter-dendrite angles.

Dendrites $(p+1)$	Gender	Brain Region	Samples (N)	vM	vM Regularized
5	Male	Occipital lobe	19	0.95	0.80
5	Female	Occipital lobe	28	0.76	0.74
5	All	Occipital lobe	47	0.58	0.57
5	Male	Frontal lobe	21	0.86	0.76
5	Female	Frontal lobe	21	0.81	0.65
5	All	Frontal lobe	42	0.50	0.49
6	Male	Frontal lobe	16	1.28	1.17
6	Female	Frontal lobe	12	1.41	1.33
6	All	Frontal lobe	28	1.13	0.96

A summary of the multivariate von Mises function fitted is shown in Fig. 6. The rose plots in the diagonal depict the marginal distributions in the original data, the numbers in the upper triangle are the $\lambda_{i,j}$ parameters of the multivariate von Mises while the values in the first column correspond to the mean and concentration parameters.



(a) Female (b) Male

Fig. 6. Angles between dendrites from adult occipital lobe neurons with 5 basal dendrites (Color figure online).

6 Conclusion and Future Work

This paper introduces a computationally optimized formulation of the pseudo-likelihood for the multivariate von Mises distribution that reduces fitting time and provides better scalability. We also propose a multivariate circular distance that can be used to compute an empirical approximation of the Kullback-Leibler divergence. We have studied the behaviour of normal and von Mises distributions using this approximated measure as reference.

Also, a generalized L_1-penalization for the multivariate von Mises distribution has been proposed and tested in cases where the number of samples is low. We applied the regularized model to the angles between basal dendrites of human pyramidal cells. A thorough study of the penalization matrix needs to be done in order to clarify the parameter scale and the impact in the final result.

All methods described in this paper will be published in an R package that will support sampling and fitting of the multivariate von Mises distribution as well as multivariate circular plots and statistics.

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. Ascoli, G.A., Donohue, D.E., Halavi, M.: Neuromorpho.org: a central resource for neuronal morphologies. J. Neurosci. **27**(35), 9247–9251 (2007)

2. Jacobs, B., Schall, M., Prather, M., Kapler, E., Driscoll, L., Baca, S., Jacobs, J., Ford, K., Wainwright, M., Treml, M.: Regional dendritic and spine variation in human cerebral cortex: a quantitative Golgi study. Cereb. Cortex **11**(6), 558–571 (2001)
3. Mardia, K.V., Hughes, G., Taylor, C.C., Singh, H.: A multivariate von Mises distribution with applications to bioinformatics. Can. J. Stat. **36**(1), 99–109 (2008)
4. Mardia, K.V., Jupp, P.E.: Directional Statistics, vol. 494. Wiley, New York (2009)
5. Mardia, K.V., Voss, J.: Some fundamental properties of a multivariate von Mises distribution. Commun. Stat-Theor. M. **43**(6), 1132–1144 (2014)
6. Mardia, K., Zemroch, P.: Algorithm as 86: the von Mises distribution function. Appl. Stat. **24**, 268–272 (1975)
7. Pérez-Cruz, F.: Kullback-Leibler divergence estimation of continuous distributions. In: IEEE International Symposium on Information Theory, pp. 1666–1670. IEEE (2008)
8. Razavian, N., Kamisetty, H., Langmead, C.J.: The von Mises graphical model: Regularized structure and parameter learning. Technical report CMU-CS-11-108, Carnegie Mellon University, Department of Computer Science (2011)
9. Tan, K.M., London, P., Mohan, K., Lee, S.I., Fazel, M., Witten, D.: Learning graphical models with hubs. J. Mach. Learn. Res. **15**(1), 3297–3331 (2014)
10. Zhu, C., Byrd, R.H., Lu, P., Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM T. Math. Softw. **23**(4), 550–560 (1997)