Frobenius Norm Regularization for the Multivariate Von Mises Distribution

Luis Rodriguez-Lujan,* Pedro Larrañaga,† Concha Bielza‡
Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain

Penalizing the model complexity is necessary to avoid overfitting when the number of data samples is low with respect to the number of model parameters. In this paper, we introduce a penalization term that places an independent prior distribution for each parameter of the multivariate von Mises distribution. We also propose a circular distance that can be used to estimate the Kullback–Leibler divergence between any two circular distributions as goodness-of-fit measure. We compare the resulting regularized von Mises models on synthetic data and real neuroanatomical data to show that the distribution fitted using the penalized estimator generally achieves better results than nonpenalized multivariate von Mises estimator.

1. INTRODUCTION

Directional data appears in many science domains in the form of directions or angles, but also as any kind of periodical data like the hours of the day. Indeed, this periodicity is the main characteristic which differentiates directional data from linear data. Recent examples of directional data in the literature cover a plethora of topics, including wind direction,[1,2] handwriting recognition,[3] people orientation in computer vision,[4] animal orientation,[5,6] marine currents,[6] protein backbone angles,[7–9] and text mining similarity measures[10] among others.

Directional statistics[11,12] provides specific tools for modeling directional data. In the past years, new circular distributions and techniques have emerged in the literature for univariate circular data[13,14] but also for multivariate circular data.[6,9,15,16] The von Mises distribution, the circular analogue of the normal distribution, is still the distribution of choice in the directional statistics field.

The extension of the von Mises distribution to a multivariate distribution[7] presents the problem that no closed formulation is known for the normalization term when the number of variables is greater than two,[17] and therefore it cannot be

*Author to whom all correspondence should be addressed; e-mail: luis.rodriguezl@upm.es.
†e-mail: pedro.larranaga@fi.upm.es.
‡e-mail: mcbielza@fi.upm.es.

easily fitted nor compared to other distributions. In this article, we reformulate the log pseudo-likelihood expression for the multivariate von Mises distribution in such a way that it becomes easier to compute.

In some application areas such as neuroanatomy, quality data are scarce and the process to obtain new data, a three-dimensional (3D) reconstruction of a neuron, could take up to several days. In this situation when we deal with a small-sample learning problem, penalizing the model complexity is needed to prevent overfitting or, as shown in this paper, to compensate the estimator bias. Although the usual approach is to use a uniform L_1 penalization over the parameters of the model,[8] in the present paper we propose a more general penalization term that could, in some way, be aware of previous knowledge about the structure of dependencies between the variables in the data, for example, taking into account spatial relationships between variables. In the linear case, for the multivariate normal distribution, this structure-aware penalization paradigm has been applied to learn graphical models with hubs[18] or to penalize according to some defined distance between the variables.[19]

The results in this paper extend those in Rodriguez-Lujan et al.[20] The added contributions of the present paper are a redefinition of a penalization term for learning the parameters of the multivariate conditional distributions, a brief proof of consistence of the penalized estimator, and the study of its bias and variance properties through numerical experiments. The application to real-world data in neuronatomy is extended to include new data sets from different species rather than focus exclusively on human neurons.

This paper is organized as follows: Section 2 reviews the univariate and multivariate von Mises distributions and defines the maximum pseudo-likelihood estimator. In Section 3, we propose a penalization term for the log pseudo-likelihood based on the Frobenius norm, prove its asymptotic convergence, and compare it against the nonpenalized estimator. Then, in Section 4, we compare the von Mises distribution and the Gaussian distribution over real data from human, rat, and mouse neurons using an approximation of the Kullback-Leibler (KL) divergence, introduced for the first time for circular data in this paper, as the assessment metric. We conclude the paper in Section 5 with a final discussion and some proposals for future work.

2. THE MULTIVARIATE VON MISES DISTRIBUTION

2.1. Definition of Density Functions

The von Mises distribution is one of the most relevant probability distributions in the field of directional statistics, and it is often considered as the normal distribution in the circumference.[12] Contrary to other directional distributions, such as wrapped[11,14] or projected distributions,[15,21] the von Mises distribution is a native directional distribution. Owing to this purely directional nature, the von Mises distribution has better mathematical properties than other nonnative circular distributions like the wrapped-normal distribution. Perhaps, one of the most important properties of the von Mises distribution is that it belongs to the canonical exponential family,

which is obvious when we examine its density function:

$$f_{VM}(\theta;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta-\mu)\}; \quad \theta,\mu \in [0,2\pi),\ \kappa > 0 \tag{1}$$

where μ stands for the mean angle, κ measures the concentration around the mean, and I_0 is the modified Bessel function of the first kind of order 0. The modified Bessel function of the first kind of order n, when n is a integer, can be expressed as the following integral formula:

$$I_n(z) = \frac{1}{\pi}\int_0^{\pi} e^{z\cos\theta}\cos(n\theta)d\theta \tag{2}$$

The density in Equation (1) is the angular interpretation of the von Mises distribution. It is easy to define the equivalent geometrical formulation using similar parameters over the unit circumference $S_1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$:

$$f_{VM}(\vec{x};\vec{\mu},\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa\ \vec{x}\cdot\vec{\mu}\}; \quad \vec{x},\vec{\mu} \in S_1,\ \kappa > 0 \tag{3}$$

Here $\cdot$ is the canonical dot product in $\mathbb{R}^2$ and $\vec{\mu}$ is the mean direction. From this definition, we can observe that the probability density for each direction (unitary vector) is proportional to its projection onto the real line determined by the mean direction. For example, the von Mises–Fisher distribution,[22] the analogue of the von Mises distribution in higher dimensions, is usually defined from this geometric point of view. In this paper, however, we will follow the angular interpretation of the von Mises distribution in both univariate and multivariate cases.

Continuing the analogy between the normal distribution and the von Mises distribution, we can define a multivariate von Mises distribution (MVM)[7] as the directional equivalent to the multivariate normal distribution. The density function of the p-variate von Mises distribution is

$$\begin{aligned} f_{MVM}(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\Lambda}) = \frac{1}{C(\boldsymbol{\kappa},\boldsymbol{\Lambda})} &\exp\{\boldsymbol{\kappa}\cdot\cos(\boldsymbol{\theta}-\boldsymbol{\mu}) \\ &+ \frac{1}{2}\sin(\boldsymbol{\theta}-\boldsymbol{\mu})\boldsymbol{\Lambda}\sin(\boldsymbol{\theta}-\boldsymbol{\mu})^T\}; \\ \boldsymbol{\theta},\boldsymbol{\mu} \in [0,2\pi)^p,\ \boldsymbol{\kappa} = (\kappa_1,\ldots,\kappa_p),\ \kappa_j > 0 \quad \forall j,\ \boldsymbol{\Lambda} \in S^p(\mathbb{R}) \end{aligned} \tag{4}$$

where $S^p(\mathbb{R})$ is the set of real symmetric matrices of size p, $\boldsymbol{\mu}$ is a p-dimensional vector, the multivariate equivalent of the mean angle in Equation (1), $\boldsymbol{\kappa}$ is the concentration vector, $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a square symmetric matrix of size p whose diagonal elements λ_{ii} are zero, and $C(\boldsymbol{\kappa},\boldsymbol{\Lambda})$ is the normalization term. Here cos and sin are applied entrywise to the p-dimensional vector $\boldsymbol{\theta}-\boldsymbol{\mu}$. The matrix $\boldsymbol{\Lambda}$ can be seen as a dependency matrix, where the element λ_{ij} measures the conditional probabilistic dependency between the ith and the jth variables. Actually, if the

distribution is highly concentrated, that is, the fluctuations in each component are sufficiently small, we can express the correlation coefficient between the ith and the jth component as a function of λ_{ij}:[7]

$$\rho_{ij} = \frac{\lambda_{ij}}{\sqrt{\kappa_i \kappa_j}}$$

The normalization term $C(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ in Equation (4) is the source of many problems that arise with the multivariate von Mises distribution. It does not have[17] a known closed-form formula for $p > 2$; therefore, its exact value has to be computed through numerical approximation, which can be troublesome in high-dimensional settings. However, as we will see in the following subsections, we can overcome some of the difficulties derived from the lack of a closed formula for the normalization term using the univariate conditional distributions of a multivariate von Mises distribution, which are indeed univariate von Mises distributions with known parameters:

$$f_{MVM}(\theta_j|\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p;\ \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = f_{VM}(\theta_j;\ \mu_{\setminus j}, \kappa_{\setminus j})$$

$$\mu_{\setminus j}(\boldsymbol{\theta};\ \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = \mu_j + \arctan\left(\frac{1}{\kappa_j}\sum_{l\neq j}\lambda_{jl}\sin(\theta_l - \mu_l)\right)$$

$$\kappa_{\setminus j}(\boldsymbol{\theta};\ \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = \sqrt{\kappa_j^2 + \left(\sum_{l\neq j}\lambda_{jl}\sin(\theta_l - \mu_l)\right)^2} \tag{5}$$

It is also important to note that the multivariate von Mises distribution is symmetrical and rotationally equivariant with respect to the mean $\boldsymbol{\mu}$. This last property allows us to assume in the following sections that $\mu_j = 0, \forall j = 1, \ldots, p$ without loss of generality. We will also restrict the values of $\boldsymbol{\kappa}$ and $\boldsymbol{\Lambda}$ such that matrix

$$\mathbf{P} = \mathrm{diag}(\kappa_1, \ldots, \kappa_p) - \boldsymbol{\Lambda} \tag{6}$$

is positive definite, where $\mathrm{diag}(\kappa_1, \ldots, \kappa_p)$ denotes the square diagonal matrix whose diagonal entries are $\kappa_1, \ldots, \kappa_p$. The positive-definiteness of $\mathbf{P}$ is a sufficient condition to ensure that the unique maximum of the multivariate von Mises distribution is attained at $\boldsymbol{\mu}$.[23]

2.2. Sampling

In this subsection, we present the two methods provided in the literature to generate samples from a multivariate von Mises distribution. Both methods rely on efficient sample generation from a von Mises distribution.[24] The first method

relies on the Gibbs sampling (GS) technique.[7,8] Our implementation of the GS is described in Algorithm 1, where $k_{thinning}$ is introduced to break the dependence between consecutive draws. In practice, a certain number of the first draws (N_{burnin}) is discarded to reduce the dependency of the generated samples on the starting point, so the selected samples are closer to the stationary distribution.

ALGORITHM 1.
Input: Parameters $\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}$
Output: $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$. That is N random samples from $MVM(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})$
Steps:

1. Initialize $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$ to some arbitrary value and set $i = 1$
2. Repeat N times:
   (a) Update $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$
   (b) Select one component $\theta_j^{(i)}$ at random and update it as:
   $\theta_j^{(i)} \sim f_{MVM}(\theta_j^{(i)} | \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i)}, \ldots, \theta_p^{(i)})$[24]
   (c) Repeat the previous step a minimum of $k_{thinning}$ times until all components have been updated at least once
   (d) Update $i = i + 1$
3. Return $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$

The second method is based on the rejection sampling algorithm using an independent product of one-dimensional von Mises distributions.[23] This method is only applicable when the matrix $\mathbf{P}$ is positive-definite. Additionally, the acceptance probability of the sampler decreases exponentially on the number of variables p, but also depends on the eigenvalues of $\mathbf{P}$. In our experiments, we will use the rejection sampling whenever the number of variables is low ($p \leq 10$), preferring the Gibbs sampler for medium and high dimensionality settings ($p > 10$).

2.3. Parameter Learning: Maximum Pseudo-Likelihood Estimator

Using the maximum likelihood estimator for the multivariate von Mises distribution is an often complicated and costly process due to the lack of a known closed formula for the normalization term $C(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ in Equation (4). Although with high concentration values and moderate correlation between variables $C(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ can be approximated by a Taylor series expansion,[9] this procedure cannot be extended to the general case. As a consequence, for each possible value of $\boldsymbol{\kappa}$ and $\boldsymbol{\Lambda}$ the value of the $C(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ has to be approximated numerically. Since we select the parametric configuration comparing the likelihood (or pseudo-likelihood) value, even relatively small errors in the estimation of the normalization term could lead to a completely wrong parameter estimation. Although there are methods that can provide an estimate avoiding the curse of dimensionality, e.g., Monte Carlo integration with uniform sampling, if we require an extremely precise estimation of the normalization term, the computational cost does not justify the precision gain compared to the use of the pseudo-likelihood. This cost just grows bigger as we need to recalculate the normalization term in each step of the minimization algorithm. For this reason, some authors have proposed the use of other estimators different from the

maximum likelihood estimator, like the pseudo-likelihood, to learn the parameters of a multivariate von Mises distribution given a set of data samples.[6–8]

The pseudo-likelihood of a multivariate density function[25] is defined as the product of all its univariate conditional densities $f(\theta_j^{(i)}|\theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i)}, \ldots, \theta_p^{(i)})$. Fortunately, as we have mentioned before, the univariate conditional densities of a multivariate von Mises distribution are von Mises distributions with known parameters. This makes the maximum pseudo-likelihood estimator computationally tractable, at the expense of providing less efficient estimations than the maximum likelihood.[26] The full pseudo-likelihood for N independent p-dimensional samples $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}\}$ is expressed as

$$\mathcal{PL}(\boldsymbol{\Theta}|\, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = (2\pi)^{-Np} \prod_{i=1}^{N} \prod_{j=1}^{p} \frac{1}{I_0(\kappa_{\backslash j}^{(i)})} \exp\left\{\kappa_{\backslash j}^{(i)} \cos\left(\theta_j^{(i)} - \mu_{\backslash j}^{(i)}\right)\right\} \tag{7}$$

where $\mu_{\backslash j}^{(i)}$ and $\kappa_{\backslash j}^{(i)}$ are the univariate conditional parameters in Equation (5) given the ith data sample $\boldsymbol{\theta}^{(i)}$. The usual approach is to maximize the natural logarithm of the pseudo-likelihood expression, which is more tractable and does not change the location of the maximum. In our case, the expression of the log pseudo-likelihood is

$$\begin{aligned} \log \mathcal{PL}(\boldsymbol{\Theta}|\, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = &-Np \log(2\pi) \\ &+ \sum_{i=1}^{N} \sum_{j=1}^{p} \left(-\log I_0(\kappa_{\backslash j}^{(i)}) + \kappa_{\backslash j}^{(i)} \cos\left(\theta_j^{(i)} - \mu_{\backslash j}^{(i)}\right)\right) \end{aligned} \tag{8}$$

Similarly to the multivariate normal distribution, the maximum likelihood estimator for the mean parameter $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_p)$, given a data sample, is the principal argument of the sample (circular) mean:

$$\hat{\boldsymbol{\mu}}_j = \arg \sum_{k=1}^{N} e^{i\theta_j^{(k)}} \tag{9}$$

We can compute $\hat{\boldsymbol{\mu}}$ and then rotate (center) the data sample $\boldsymbol{\Theta}$, so that we can assume $\boldsymbol{\mu} = 0$ for the rest of this section.

To find the maximum of function (8), we use the low memory extension of the widespread Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method with simple box constraints (L-BFGS-B).[27] The concentration values are restricted to be positive, that is, $\kappa_j > 0, \forall\ j = 1, \ldots, p$. An advantage of the quasi-Newton methods is that it suffices to compute first-order derivatives of the function to maximize (8), i.e., we need to compute the partial derivatives of the log pseudo-likelihood

function with respect to κ_j and λ_{jk}. The resulting equations are

$$\frac{\partial \log \mathcal{PL}}{\partial \kappa_j} = \sum_{i=1}^{N} \left[\frac{\partial \kappa_{\setminus j}^{(i)}}{\partial \kappa_j} \left(\cos\left(\theta_j^{(i)} - \mu_{\setminus j}^{(i)}\right) - A_1\left(\kappa_{\setminus j}^{(i)}\right) \right) + \frac{\partial \mu_{\setminus j}^{(i)}}{\partial \kappa_j} \kappa_{\setminus j}^{(i)} \sin\left(\theta_j^{(i)} - \mu_{\setminus j}^{(i)}\right) \right]$$

$$\frac{\partial \log \mathcal{PL}}{\partial \lambda_{jk}} = \sum_{i=1}^{N} \left[\frac{\partial \kappa_{\setminus j}^{(i)}}{\partial \lambda_{jk}} \left(\cos\left(\theta_j^{(i)} - \mu_{\setminus j}^{(i)}\right) - A_1\left(\kappa_{\setminus j}^{(i)}\right) \right) + \frac{\partial \mu_{\setminus j}^{(i)}}{\partial \lambda_{jk}} \kappa_{\setminus j}^{(i)} \sin\left(\theta_j^{(i)} - \mu_{\setminus j}^{(i)}\right) \right]$$

$$\begin{aligned}
\frac{\partial \kappa_{\setminus j}^{(i)}}{\partial \kappa_R} &= \delta(R, j) \frac{\kappa_j}{\kappa_{\setminus j}^{(i)}} \\
\frac{\partial \kappa_{\setminus j}^{(i)}}{\partial \lambda_{R,S}} &= \delta(R, j) \frac{\sin(\theta_S) \sum_{l \neq j} \lambda_{R,l} \sin \theta_l^{(i)}}{\kappa_{\setminus j}^{(i)}} \\
\frac{\partial \mu_{\setminus j}^{(i)}}{\partial \kappa_R} &= \delta(R, j) \frac{-\sum_{l \neq j} \lambda_{R,l} \sin \theta_l^{(i)}}{\left(\kappa_{\setminus j}^{(i)}\right)^2} \\
\frac{\partial \mu_{\setminus j}^{(i)}}{\partial \lambda_{R,S}} &= \delta(R, j) \frac{\sum_{l \neq j} \lambda_{R,l} \sin \theta_l^{(i)}}{\left(\kappa_{\setminus j}^{(i)}\right)^2}
\end{aligned} \tag{10}$$

where $A_1 = \frac{I_1}{I_0}$ is the ratio of the modified Bessel functions of order one and zero, and $\delta(x, y) = 1$ if $x = y$ and zero otherwise. In these expressions, R and S can be substituted for any valid index from 1 to p.

2.3.1. *Computational Complexity Reduction*

Our goal in this section is to simplify function (8) and specially (10) from a computational point of view. Our first step is to express certain sums as a matrix product to take advantage of highly efficient implementations of linear algebra computations such as BLAS or LAPACK.[28] We define the $N \times p$ auxiliary matrix $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi} = \sin(\boldsymbol{\Theta})\boldsymbol{\Lambda}, \text{ that is, } \quad \psi_{ij} = \sum_{l \neq j} \lambda_{jl} \sin\left(\theta_l^{(i)}\right) \tag{11}$$

Then we focus on the right-hand term of (8). By applying some basic trigonometric identities, we obtain

$$\cos\left(\theta_j^{(i)} - \mu_{\setminus j}^{(i)}\right) = \cos\left(\theta_j^{(i)} - \arctan\left(\frac{1}{\kappa_j} \sum_{l \neq j} \lambda_{j,l} \sin\left(\theta_l^{(i)}\right)\right)\right)$$

$$
\begin{aligned}
&= \cos\left(\theta_j^{(i)}\right) \cos\left(\arctan\left(\frac{1}{\kappa_j}\sum_{l\neq j}\lambda_{j,l}\sin\left(\theta_l^{(i)}\right)\right)\right) \\
&\quad + \sin\left(\theta_j^{(i)}\right) \sin\left(\arctan\left(\frac{1}{\kappa_j}\sum_{l\neq j}\lambda_{j,l}\sin\left(\theta_l^{(i)}\right)\right)\right) \\
&= \cos\left(\theta_j^{(i)}\right)\frac{\kappa_j}{\kappa_{\setminus j}^{(i)}} - \sin\left(\theta_j^{(i)}\right)\frac{\sum_{l\neq j}\lambda_{j,l}\sin\left(\theta_l^{(i)}\right)}{\kappa_{\setminus j}^{(i)}} \qquad (12)
\end{aligned}
$$

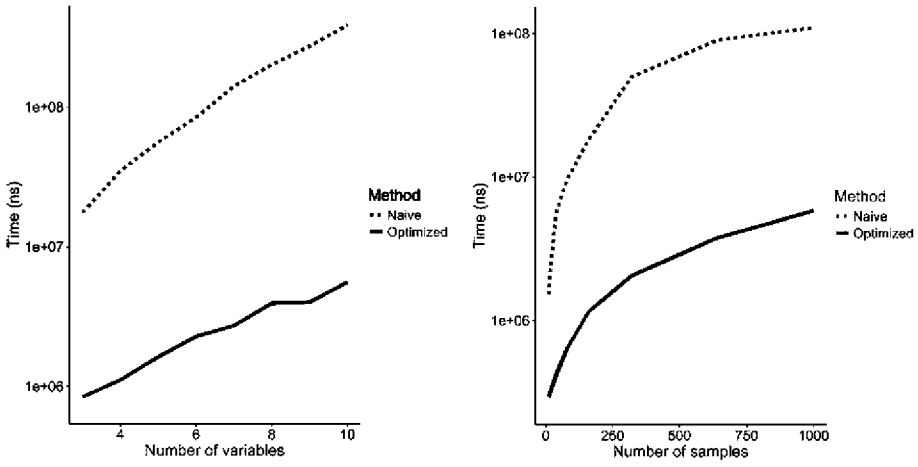
If we substitute this expression in (8) and express the sums as elements of the $\boldsymbol{\Psi}$ matrix defined in Equation (11), we obtain a more compact version of the log pseudo-likelihood function that does not require to compute the inverse of the tangent:

$$
\begin{aligned}
\log \mathcal{PL}_{op}(\boldsymbol{\Theta}|\,\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\Lambda}) = -Np\log(2\pi) + \sum_{i=1}^{N}\sum_{j=1}^{p}\Big(&-\log I_0\big(\kappa_{\setminus j}^{(i)}\big) \\
&+ \cos\left(\theta_j^{(i)}\right)\kappa_j - \sin\left(\theta_j^{(i)}\right)\psi_{ij}\Big) \qquad (13)
\end{aligned}
$$

Additionally, the partial derivatives obtained from Equation (13) are simpler partial derivatives than those in (10):

$$
\begin{aligned}
\frac{\partial \log \mathcal{PL}_{op}}{\partial \kappa_j} &= \sum_{i=1}^{N}\left[\cos\left(\theta_j^{(i)}\right) - A_1\big(\kappa_{\setminus j}^{(i)}\big)\frac{\kappa_j}{\kappa_{\setminus j}^{(i)}}\right] \\
\frac{\partial \log \mathcal{PL}_{op}}{\partial \lambda_{jk}} &= \sum_{i=1}^{N}\left[\sin\left(\theta_k^{(i)}\right)\left(\sin\left(\theta_j^{(i)}\right) - A_1\big(\kappa_{\setminus j}^{(i)}\big)\frac{\psi_{ij}}{\kappa_{\setminus j}^{(i)}}\right)\right] \qquad (14)
\end{aligned}
$$

To evaluate the performance gain between expressions (8) and (13), we compare the average execution time to compute the maximum pseudo-likelihood estimation using each one. To maximize the likelihood function, we use the L-BFGS-B algorithm, a quasi-Newton method with reduced memory footprint. In both cases, the routines that perform most of the calculations are implemented in *ANSI C*. We propose two experiments: First, the number of variables p varies for a fixed number of samples N (Experiment A); and second, the other way around (Experiment B). We repeat each experiment 100 times for every one of the 10 randomly generated $\mathbf{P}$ positive definite matrices to compute the average time as well as the standard deviation. The experimental configuration is summarized in Table I. The experiments were executed in a common desktop computer with an i7-4790k processor at 4 GHz and 8 GB of RAM at 1600 MHz.



(a) $N = 100$, p varying (b) $p = 3$, N varying

Figure 1. Average fitting time for the naive implementation (8) and the optimized version (13) of the log pseudo-likelihood function. Each execution has been repeated 100 times with two additional warm-up iterations. We have plotted an envelope of width σ around the average time line, but, due to its small size, it can be hardly seen.

Table I. Experimental configuration summary for time comparison.

Experiment	p	N	Number of repetitions
A	$\{3, 4, \ldots, 10\}$	100	10 random $\mathbf{P}$ matrices $\times$ 100 times
B	3	$\{10, 20, 40, \ldots, 1000\}$	10 random $\mathbf{P}$ matrices $\times$ 100 times

Figure 1 shows that the optimized version is significantly faster in every case, and it scales better when the number of samples (N) or variables (p) increases, although both versions have the same computational complexity $\mathcal{O}(p^2N)$. Based on these results, we will use this optimized version of the log pseudo-likelihood function in the next experiments.

2.3.2. *Estimator Properties*

Asymptotic consistency of an estimator is perhaps the most important property of any admissible estimator as it guarantees that the estimator is arbitrarily close to the estimated value provided sufficient number of samples. In other words, the estimation improves as the number of samples increases. The consistency of the maximum pseudo-likelihood estimator has been already proved by other authors.[8]

Although there are some related works on the analytical study of mean and variance properties on implicitly defined biased estimators,[29] we follow an empirical approach to evaluate the bias and the variance of the maximum pseudo-likelihood estimator when estimating $\boldsymbol{\kappa}$ and $\boldsymbol{\Lambda}$. Specifically, we aim to determine the

Table II. Experiments to estimate bias and variance properties.

Experiment	p	N	$\boldsymbol{\kappa}$	$\boldsymbol{\Lambda}$	Number of repetitions
A	3	10	(3, 3, 3)	$\boldsymbol{\Lambda} = 0$	5000
B	3	10	(3, 3, 3)	$\boldsymbol{\Lambda} \neq 0$	5000
C	3	10	(0.01, 0.01, 0.01)	$\boldsymbol{\Lambda} = 0$	5000
D	3	10	(0.01, 0.01, 0.01)	$\boldsymbol{\Lambda} \neq 0$	5000
E	10	50	(3, . . . , 3)	$\boldsymbol{\Lambda} = 0$	5000
F	10	50	(3, . . . , 3)	$\boldsymbol{\Lambda} \neq 0$	5000
G	50	100	(3, . . . , 3)	$\boldsymbol{\Lambda} = 0$	10000
H	50	100	(3, . . . , 3)	$\boldsymbol{\Lambda} \neq 0$	10000

properties of the estimator for low and medium concentration values ($\boldsymbol{\kappa}$), as well as for configurations when the variables are independent ($\boldsymbol{\Lambda} = \mathbf{0}$) and dependent ($\boldsymbol{\Lambda} \neq \mathbf{0}$). Based on the previous goals, we define the experimental configurations listed in Table II.

For each parametric configuration, we repeat the same process a high number of times: First, N independent samples are generated from a MVM($\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\kappa}$, $\boldsymbol{\Lambda}$). Then, the maximum pseudo-likelihood estimator is applied to get the estimated values $\hat{\boldsymbol{\xi}}_{\mathcal{PL}} = (\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\Lambda}})$. Finally, after all repetitions have been completed, we compute the bias and the variance of the estimator. Note that in the cases where $\boldsymbol{\Lambda} = \mathbf{0}$ the pseudo-likelihood estimator is in fact the maximum likelihood estimator.

For each experiment, we apply the Hotelling's T-square test,[30] a multivariate hypothesis test, over $\hat{\boldsymbol{\kappa}}$ and $\hat{\boldsymbol{\Lambda}}$ independently to verify whether the estimator is unbiased for that configuration. For those cases where the unbiasedness is rejected, we perform a one-sided t-test for each individual parameter ξ_j to check if its estimator is overestimating the real value ($\mathbb{E}[\frac{\hat{\xi}_j}{\xi_j}] > 1$) or either underestimating it. The $\hat{\boldsymbol{\kappa}}$ unbiasedness hypothesis is rejected with a significance level $\alpha = 0.05$ in every case, whereas it is not rejected for $\hat{\boldsymbol{\Lambda}}$ in experiments A, C, and E and rejected in the rest with the same α. Defying our belief that the $\hat{\boldsymbol{\Lambda}}$ estimator is unbiased when $\boldsymbol{\Lambda} = \mathbf{0}$, the hypothesis is rejected in experiment G, even when the average bias value is low ($\approx 10^{-2}$) for each $\hat{\lambda}_{ij}$. Therefore, due to this last result, we cannot draw any conclusion about the unbiasedness hypothesis. Finally, the overestimation hypothesis $H_0 : \mathbb{E}[\frac{\hat{\xi}_j}{\xi_j}] > 1$ is not rejected for every single parameter where the unbiasedness hypothesis has been rejected, which means that the maximum pseudo-likelihood statistic overestimates the real value of the parameters.

Finally, we perform a last experiment. We want to check for a relation between the distance of $\boldsymbol{\Lambda}$ to the zero matrix, measured as the Frobenius norm of $\boldsymbol{\Lambda}$ ($\|\boldsymbol{\Lambda}\|_F$), and $Bias(\hat{\boldsymbol{\Lambda}})$. To do so, we generate $N = 10$ samples from a three-variate MVM with fixed $\boldsymbol{\kappa} = (3, 3, 3)$ and compute the Euclidean norm of the estimator bias while varying $\|\boldsymbol{\Lambda}\|_F$ from 0 to 4, verifying in each case that the matrix $\mathbf{P}$ is positive definite. For every parameter configuration, the process is repeated 10,000 times. As seen in Figure 2, there is a clear linear relationship between $\|\boldsymbol{\Lambda}\|_F$ and $\|Bias(\hat{\boldsymbol{\Lambda}})\|_2$.

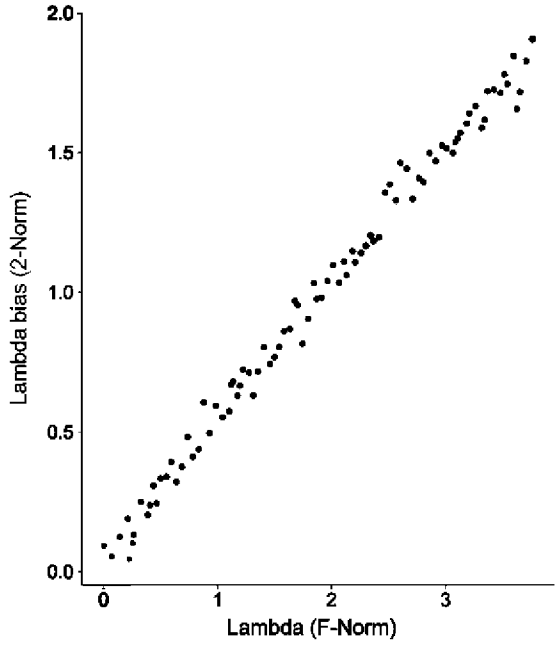


Figure 2. Relation between the distance to the zero matrix $\|\mathbf{\Lambda}\|_F$ and $\|Bias(\hat{\mathbf{\Lambda}})\|_2$.

3. PENALIZED PARAMETER LEARNING

Learning the parameters of a multivariate von Mises distribution using a penalized maximum pseudo-likelihood estimator has been already proposed in the literature by other authors.[8] Their approach is to add a term in the pseudo-likelihood expression that penalizes the absolute value of all λ_{ij} elements equally. Since the pseudo-likelihood overestimates the parameters, specially when learning from small samples, it is broadly justified to apply a penalization on the parameter absolute value. The actual problem is that, if we apply the exact same penalty on each λ_{ij} we are implicitly assuming that all λ_{ij} are similar in size.

From the Bayesian data analysis perspective,[31] the parameters $\boldsymbol{\kappa}$ and $\mathbf{\Lambda}$ are real-valued random variables with a specific distribution given the data $\mathbf{\Theta}$, the posterior probability distribution $p(\boldsymbol{\kappa}, \mathbf{\Lambda} \,|\mathbf{\Theta})$. This function is obtained from the prior probability distribution $p(\boldsymbol{\kappa}, \mathbf{\Lambda})$, a distribution that reflects the prior knowledge about the value of the parameters, and the aforementioned likelihood function through Bayes' theorem:

$$p(\boldsymbol{\kappa}, \mathbf{\Lambda} \,|\mathbf{\Theta}) \propto p(\boldsymbol{\kappa}, \mathbf{\Lambda}) p(\mathbf{\Theta}|\, \boldsymbol{\kappa}, \mathbf{\Lambda}) \tag{15}$$

Hence, the maximum likelihood estimator corresponds to a scenario with an uninformative prior where maximizing the posterior probability actually corresponds to

maximizing the likelihood. Under this Bayesian framework, a uniform L_1 regularization corresponds to setting the exact same prior distribution, centered at zero, for each parameter and to assuming that all λ_{ij} are similar in size. This issue has been already studied in the literature, but focusing on multivariate normal distributions.[18,19] Here we propose a penalized estimator based on the maximum pseudo-likelihood estimator where each parameter is penalized individually; in other words, the prior distribution for each parameter is independent from the rest.

3.1. F-Norm Penalized Pseudo-Likelihood

Let $\mathbf{P}$ be the positive definite matrix defined in Equation (6), $\mathbf{\Phi}$ a $p \times p$ real symmetric matrix, and $\boldsymbol{H}$ a triangular real $p \times p$ matrix, whose elements h_{ij} are nonnegative. Then, we add the following penalization term to the log pseudo-likelihood introduced in Equation (13):

$$pen(\boldsymbol{\kappa}, \mathbf{\Lambda}) = -\|(\mathbf{P} - \mathbf{\Phi}) \circ \boldsymbol{H}\|_F \tag{16}$$

where $\|A\|_F = \sqrt{tr(AA^T)}$ is the Frobenious matrix norm (F-norm) and $\circ$ is the entrywise Hadamard product $A \circ B = (a_{ij}b_{ij})$. In the penalization term given by Equation (16), the matrix $\mathbf{\Phi}$ is our prior guess about the elements p_{ij} of $\mathbf{P}$; whereas h_{ij}, the elements of the confidence matrix $\boldsymbol{H}$,[32] measures our degree of confidence in the correctness of the value ϕ_{ij}. Note that the triangular structure imposed to $\boldsymbol{H}$ is not arbitrary since it prevents the same $\lambda_{ij} = \lambda_{ji}$ parameter from being penalized twice.

The last requirement to use the penalized pseudo-likelihood with the term given by Equation (16) is to update the partial derivatives in Equation (14) by adding the following terms:

$$\begin{aligned} \frac{\partial pen(\boldsymbol{\kappa}, \mathbf{\Lambda})}{\partial \kappa_j} &= -\frac{h_{jj}^2(\kappa_j - \phi_{jj})}{\|(\mathbf{P} - \mathbf{\Phi}) \circ \boldsymbol{H}\|_F} \\ \frac{\partial pen(\boldsymbol{\kappa}, \mathbf{\Lambda})}{\partial \lambda_{jk}} &= -\frac{h_{jk}^2(\lambda_{jk} - \phi_{jk})}{\|(\mathbf{P} - \mathbf{\Phi}) \circ \boldsymbol{H}\|_F} \end{aligned} \tag{17}$$

Now, we face the problem of selecting values for $\mathbf{\Phi}$ and $\boldsymbol{H}$ such that the estimation given by the penalized version is at least as good as the unpenalized one. Based on the results in the preceding section that show how the maximum pseudo-likelihood estimator overestimates the scale of $\boldsymbol{\kappa}$ and $\mathbf{\Lambda}$, we can mimic the L_1 penalization and set $\mathbf{\Phi} = \mathbf{0}$ if no prior information is available, but we still need to define the values of the confidence matrix for each parameter. A conservative approach, given the absence of reliable prior information, is to set the elements h_{ij} to some medium-low value close to 1; otherwise, there is a risk of underestimating the parameters. Obviously, the exact best value for each h_{ij} depends on the number of variables p and the real value of p_{ij}. In Section 4, we show a real example where the structure of this confidence matrix is clearly induced by the problem itself. Another interesting

approach for highly concentrated data, could be to use the inverse of the sample circular covariance matrix[7] as our $\mathbf{\Phi}$ matrix.

Going back to the Bayesian analysis framework, the penalization term in Equation (16) in the log pseudo-likelihood could be intuitively seen as placing an independent normal prior $\mathcal{N}(\phi_{ij}, \sigma(h_{ij}))$ over each p_{ij}.

3.2. Properties

As we did with the maximum pseudo-likelihood estimator in Section (2.3.2), in this section we analyze the properties of the penalized estimator and compare the experimental results between the unpenalized and the penalized maximum pseudo-likelihood estimators. The following proposition addresses the consistency of the penalized estimator.

PROPOSITION 1. *The penalized maximum pseudo-likelihood estimator*

$$\hat{\boldsymbol{\xi}}_{\mathcal{PPL}} := \underset{\boldsymbol{\xi}=(\boldsymbol{\kappa},\boldsymbol{\Lambda})}{\operatorname{argmax}} \{\log \mathcal{PL}(\boldsymbol{\Theta}|\ \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})_{op} + pen(\boldsymbol{\kappa}, \boldsymbol{\Lambda})\}$$

is consistent.

Proof. Let $\hat{\boldsymbol{\xi}}^n_{\mathcal{PL}}$ and $\hat{\boldsymbol{\xi}}^n_{\mathcal{PPL}}$ be the unpenalized and penalized maximum pseudo-likelihood estimators, respectively, and $\boldsymbol{\xi}_0$ the real parameter that exists is unique and well-defined. Then, we can define that the maximum pseudo-likelihood estimator $\hat{\boldsymbol{\xi}}^n_{\mathcal{PL}}$ as an M-estimator that maximizes the function:

$$\mathbb{M}_n(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}(\log \mathcal{PL}(\boldsymbol{\theta}^i|\ \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}))$$

If we restrict ourselves to a compact neighborhood of $\boldsymbol{\xi}_0$, $\mathcal{B}(\boldsymbol{\xi}_0)$, and having in mind that the log-pseudo-likelihood function is continuous and square-integrable, then the conditions to apply the uniform strong law of large numbers are met on that neighborhood, as a result:

$$\sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \|\mathbb{M}_n(\boldsymbol{\xi}) - \mathbb{E}[\log \mathcal{PL}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})]\| \underset{a.s.}{\longrightarrow} 0$$

For convenience, in the rest of the proof we will refer to the function $\mathbb{E}[\log \mathcal{PL}(\boldsymbol{\theta}|\ \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})]$ simply as $M(\boldsymbol{\xi})$. Please recall that $\boldsymbol{\xi} = (\boldsymbol{\kappa}, \boldsymbol{\Lambda})$.

On the other hand, let $h_n(\boldsymbol{\xi}) = \frac{\|(\mathbf{P}-\mathbf{\Phi})\circ \boldsymbol{H}\|_F}{N}$ be our complexity penalization function. It is clear that h_n is continuous, positive and the sequence decreases monotonically (i.e., $h_n < h_m,\ \forall\ m > n$). Since the numerator is obviously finite in any compact set of the parameter space, $h_n \longrightarrow 0$ pointwise. Furthermore, by applying Dini's theorem, we can state that h_n converges to 0 uniformely in any compact subspace of the parameter space.

As we did before, the penalized maximum pseudo-likelihood estimator $\hat{\boldsymbol{\xi}}^n_{\mathcal{PPL}}$ can be also defined as an M-estimator:

$$\mathbb{M}_n^{pen}(\boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}(\log \mathcal{PL}(\boldsymbol{\theta}^i \mid \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})) + h_n(\boldsymbol{\xi})$$

Now, due to the fact that $\boldsymbol{\xi}_0$ exists, is unique and well-defined, $\exists R$ such that $\forall\, r < R\ S_r : \{\boldsymbol{\xi} \mid \|\boldsymbol{\xi}_0 - \boldsymbol{\xi}\| \leq \frac{r}{2}\}$ verifies that

$$\forall \boldsymbol{\xi} \notin S_r\ M(\boldsymbol{\xi}) \leq M(\boldsymbol{\xi}_0) - \epsilon_r$$

Therefore, $\forall \delta > 0\ \exists N_1 \mid \forall n_1 > N_1$ such that

$$P\left\{\sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \|\mathbb{M}_n(\boldsymbol{\xi}) - M\| > \frac{\epsilon_r}{2}\right\} < \delta$$

In parallel, $\sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \|\mathbb{M}_n - \mathbb{M}_n^{pen}\| = \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} h_n$, as h_n converges uniformly to 0 in any compact, then $\exists N_2 \mid \forall n_2 > N_2 \quad \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} h_n < \frac{\epsilon_r}{2}$. It follows that

$$\sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \left\|M - \mathbb{M}_n^{pen}\right\| \leq \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \|M - \mathbb{M}_n\| + \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \left\|\mathbb{M}_n - \mathbb{M}_n^{pen}\right\|$$

$$\Rightarrow P\left\{\sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}_0)} \left\|M - \mathbb{M}_n^{pen}\right\| > \epsilon_r\right\} < \delta$$

which implies that $P\{\|\boldsymbol{\xi}_0 - \hat{\boldsymbol{\xi}}^n_{\mathcal{PPL}}\| > \frac{r}{2}\} < \delta \quad \forall\, n > \max(N_1, N_2)$. Then when can take $r \longrightarrow 0$ and $\delta \longrightarrow 0$ so the argument holds when $n \longrightarrow \infty$. □

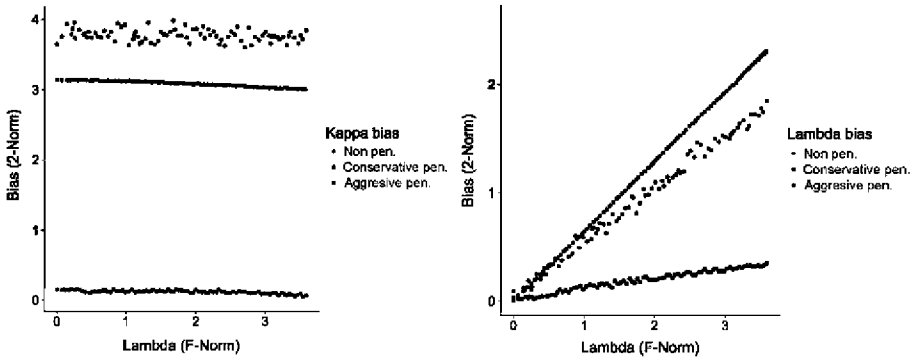
To experimentally study the properties of the penalized estimator and compare it with the unpenalized version, we repeat the experiments of Table II. Following our own recommendations, we will test the penalized estimator with prior matrix $\boldsymbol{\Phi} = 0$ and compare two penalization strategies: a conservative approach where the elements $h_{i\geq j} = 0.5$, and a more aggressive penalization with mid-high confidence values $h_{i\geq j} = 3$. Our expectation is that while the conservative approach will reduce both bias and variance with respect to the nonpenalized estimator, the aggressive approach will increase the bias due to underestimation in low-dimensional settings where the real parameters values are far from zero, but with a much lower variance. On the other hand, in the cases where the number of parameters is high we expect a sharp increase in the parameter overestimation produced by the pseudo-likelihood estimator. It is reasonable to believe that in those cases a higher penalization would perform better. Results in Table III agree with our expectations.

Finally, we also repeat the low-dimensional experiment shown in Figure 2 including the two penalized approaches. In Figure 3, we can see how the conservative approach clearly reduces the bias in both $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\kappa}}$ estimations and outperforms the

Table III. Euclidean norm of the bias, trace of the variance matrix, and mean square error (MSE) comparison between unpenalized, conservative penalization, and aggressive penalization for the experiments defined in Table (II).

	Nonpenalized			Penalized $h_{ij} = 0.5$			Penalized $h_{ij} = 3$		
Experiment	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
A	3.75	61.69	75.80	**0.14**	**5.55**	**5.57**	3.13	0.13	9.90
B	3.94	80.57	96.70	**1.38**	**5.31**	**7.22**	4.57	0.18	21.07
C	1.17	3.99	5.37	1.08	1.90	3.07	**1.00**	**0.64**	**1.65**
D	1.31	4.04	5.77	1.24	2.19	3.74	**1.15**	**0.70**	**2.04**
E	1.88	31.15	34.70	1.27	24.08	25.69	**1.10**	**7.16**	**8.38**
F	2.24	61.15	66.18	**1.71**	43.79	49.72	2.04	**6.47**	**10.65**
G	14.34	1717.14	1922.76	14.10	1712.23	1910.97	**10.76**	**852.14**	**967.99**
H	14.23	1676.50	1879.01	13.64	1469.97	1655.95	**10.16**	**798.11**	**901.24**

Best results are in bold characters.



(a) $\hat{\kappa}$ bias comparison (b) $\hat{\mathbf{\Lambda}}$ bias comparison

Figure 3. Bias comparison between unpenalized, conservative penalization, and aggressive penalization for a three-variate von Mises distribution with $\boldsymbol{\kappa} = (3, 3, 3)$ from $N = 10$ samples and $\|\mathbf{\Lambda}\|_F$ varying from 0 to 4.

aggressive approach. As we anticipated before, Figure 3b shows how the most aggressive penalization actually increases the bias of $\hat{\mathbf{\Lambda}}$ due to underestimation, which it is not the case in Figure 3a.

4. A REAL-WORLD APPLICATION IN NEUROANATOMY

As an application of the penalized estimator on real-world data, we use the von Mises multivariate distribution to model the angles between basal dendrites of pyramidal cells. A neuron can be divided into three main parts: the cell body or soma, dendrites, and axon. Among all the different types of neurons in the nervous system, pyramidal neurons stand out as one of the most important types. They play a key role in the connectivity of cortical columns which are the functional blocks in

the neocortex. Since the morphology of a neuron determines its connectivity pattern to a large extent,[33] we are interested in studying the pyramidal neurons from an anatomical point of view. A unique characteristic of the pyramidal neurons is that they have basal dendrites, so called because they grow from the base of the soma and spread out horizontally. Our goal is to characterize the angles between these basal dendrites and to find differences or similarities between species, brain regions, etc. Ultimately, this knowledge will lead to better single-cell simulations and will improve our understanding on how the brain works.

4.1. Evaluation: Approximated KL Divergence

First of all, we need to determine how to assess the goodness-of-fit of our models. Here we face the problem of dealing with multivariate directional data, for which we have very few available tools compared to the well-known field of multivariate linear data. Other authors use a variety of techniques to evaluate the goodness of fit, such as using a multivariate normal approximation[7] that cannot be applied in the general case or they assess the log-likelihood value,[8] which can be hard to compute accurately for the multivariate von Mises distribution, and it is not appropriate to compare distributions from different families.

Here, we propose to use an approximation of the KL divergence for multivariate distributions that does not make any assumption on the underlying true distribution of the samples[34] as evaluation metric. Suppose that we have two sets of samples $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^{M}$ from two p-dimensional unknown probability distributions, P and Q, respectively. Then the estimated KL divergence is

$$\hat{D}_k(P||Q) = \frac{p}{N}\sum_{i=1}^{N}\left[\log\left(\frac{s_k(\boldsymbol{x}_i)}{r_k(\boldsymbol{x}_i)}\right)\right] + \log\frac{M}{N-1} \tag{18}$$

where $r_k(\boldsymbol{x}_i)$ and $s_k(\boldsymbol{x}_i)$ are the distance to the kth nearest neighbor of $\boldsymbol{x}_i$ in $\boldsymbol{X} \setminus \{\boldsymbol{x}_i\}$ and $\boldsymbol{Y}$, respectively.

With this approximation of the KL divergence between the unknown underlying distribution of the data and the fitted model, we can measure how far is the fitted distribution from the data just by taking one set of samples from the data and the other generated by sampling the fitted distribution using any method defined in Section 2.2 for the multivariate von Mises distribution. However, there is still a missing piece to compute the estimation: We need to define which distance are we going to use. Any linear norm such as the Euclidean norm is not adequate since they do not take into account the periodicity of the data. Another option would be to use the geodesic distance in the hyper-torus $\mathbb{T}^p$, but it is rather complicated to compute. To overcome these drawbacks, we define a distance between two points $\boldsymbol{a},\boldsymbol{b} \in [0, 2\pi)^p$ in Equation (19) that is easy to compute and takes into account the periodicity of circular data.

$$d(\boldsymbol{a}, \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}^*||_2 \tag{19}$$

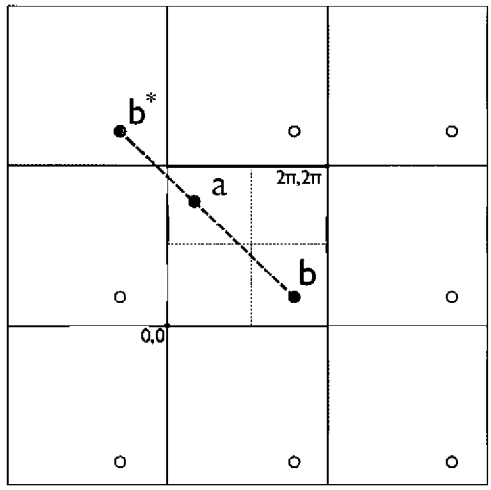


Figure 4. Circular distance computation for $p = 2$ defined by Equation (19).

where

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

Figure 4 provides a graphical explanation on how this distance works for $p = 2$.

4.2. Data Extraction and Modeling

Two sets of 3D reconstructions of pyramidal neurons were downloaded from NeuroMorpho.Org,[35] a public repository of neural reconstructions. One contains 1915 reconstructions of human pyramidal cells from different brain regions,[36] whereas the other has 90 reconstructions of mouse pyramidal cells from the neocortex.[37] A third set of 38 reconstructions of rat pyramidal neurons from the somatosensory cortex was downloaded from a different source.[38,39] Table IV contains the count of reconstructions per data set aggregated by the number of basal dendrites. We only pick those cases where the number of samples is at least six; otherwise, we are not able to estimate the KL divergence properly as with smaller samples the test set would be too small to provide a reliable KL estimation. If a neuron has p basal dendrites, then we have $p - 1$ variables between dendrites since the last angle is fully determined by the rest. Note that there is a linear restriction over the angles as they must total less than 2π, although it is not considered in the model. The restriction should be taken into account when computing the conditional distribution of one angle given the others, as the range of accepted values is restricted. However, the model is still valid as a first approach, and these issues will be considered as future improvements.

Table IV. Number of reconstructions per data set aggregated by the number of basal dendrites.

	Data set		
Basal count	Human	Mouse	Rat
1	8	0	0
2	11	0	0
3	133	1	1
4	475	11	4
5	598	31	13
6	420	32	6
7	143	9	4
8	87	5	7
9	18	1	1
10	11	0	2
11	8	0	0
12	2	0	0
13	1	0	0
Total	1915	90	38

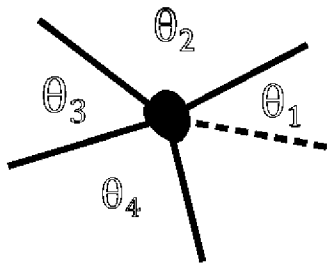


Figure 5. Angles between dendrites.

Since basal dendritic roots do not exactly lie in a plane, we need to project them onto the basal plane, defined as the plane that minimizes the distance of the basal roots to their projection onto the plane, which is exactly what principal component analysis solves. Then, we need to establish a consistent criterion to select the basal tree from which we start numbering in counter-clockwise order as shown in Figure 5. The data extraction procedure is detailed in Algorithm 2, which is applied to measure the angles from each reconstruction.

ALGORITHM 2.
Input: 3D reconstruction
Output: $\theta_1, \ldots, \theta_p$ angles between basal dendrites
Steps:

1. Extract Cartesian coordinates of the basal roots
2. Find the basal plane using principal component analysis on the coordinates of the basal roots
3. Project the basal roots onto the basal plane
4. Measure the total length of each basal tree and designate the longest as the first
5. Measure planar angles between the projected basal roots in counter-clockwise order

Table V. Mean KL divergence results for angles between basal dendrites ± standard deviation (lower is better).

		Multivariate von Mises				MV Normal	
p	Samples	Nonpenalized	$h_{j>i} = 0.5$	$\boldsymbol{H} = \boldsymbol{H}_{dist}$	$L_1(1)$	Linear	Wrapped
				Rat			
4	12	1.56 ± 1.26	0.62 ± 0.42	0.68 ± 0.45	**0.51 ± 0.36**	0.79 ± 0.76	1.09 ± 0.90
7	8	9.31 ± 1.86	2.72 ± 0.85	**1.96 ± 0.64**	2.60 ± 0.85	2.69 ± 0.85	2.69 ± 0.94
				Mouse			
3	11	2.04 ± 1.16	1.01 ± 0.48	**0.56 ± 0.44**	0.75 ± 0.48	1.33 ± 0.85	1.78 ± 1.28
4	31	0.37 ± 0.35	0.72 ± 0.40	**0.34 ± 0.26**	0.55 ± 0.31	**0.34 ± 0.27**	0.35 ± 0.30
5	32	0.44 ± 0.35	0.61 ± 0.40	0.39 ± 0.28	0.42 ± 0.28	**0.35 ± 0.29**	**0.35 ± 0.26**
6	9	8.95 ± 1.99	3.07 ± 0.64	**0.94 ± 0.63**	1.78 ± 1.11	2.29 ± 1.12	2.65 ± 1.30
				Human			
3	475	**0.51 ± 0.10**	0.53 ± 0.10	0.52 ± 0.09	0.52 ± 0.09	0.55 ± 0.09	1.16 ± 0.11
4	598	**0.74 ± 0.10**	**0.74 ± 0.10**	0.76 ± 0.10	0.76 ± 0.11	0.85 ± 0.11	0.94 ± 0.11
5	420	0.87 ± 0.13	0.88 ± 0.11	0.85 ± 0.12	0.85 ± 0.11	0.80 ± 0.13	**0.68 ± 0.12**
6	143	0.67 ± 0.21	0.70 ± 0.21	0.63 ± 0.23	0.69 ± 0.24	0.51 ± 0.17	**0.47 ± 0.22**
7	87	0.33 ± 0.22	0.56 ± 0.26	0.35 ± 0.22	0.45 ± 0.28	**0.27 ± 0.24**	0.29 ± 0.23
8	18	4.69 ± 1.96	1.71 ± 0.70	0.80 ± 0.69	**0.63 ± 0.43**	0.84 ± 0.68	0.84 ± 0.68
9	11	8.96 ± 1.60	3.10 ± 0.93	**1.41 ± 0.93**	2.88 ± 1.04	2.26 ± 1.19	2.25 ± 1.09
Average rank		4.62	4.31	2.23	3.38	3.00	3.46

For the F-norm penalized multivariate von Mises estimation, the prior matrix is $\boldsymbol{\Phi} = \mathbf{0}$ in both cases, whereas the penalization parameter is 1 in the L_1 case. Best results are in bold characters.

6. Return $\theta_1, \ldots, \theta_p$

Probably, the weakest point in the procedure described in Algorithm 2 is the determination of the order of the angles, i.e., selecting the first angle based on the longest basal tree. In the absence of a common coordinate system for all reconstructions in a data set, any local criterion would be (up to some point) arbitrary. Indeed, we have verified that the null hypothesis that all marginal distributions are equal is not rejected by the q-sample uniform-scores test[12] in all cases included in Table V. This result suggests that either the criterion is totally arbitrary, the reconstruction angles are truly equidistributed, or both are true. Despite these shortcomings, this method is valuable as, to the best of our knowledge, there are no other multivariate approaches in the literature in this regard.

In our experiments, we compare the KL divergence estimate between the underlying distribution and the multivariate von Mises distributions with parameters estimated with the unpenalized maximum pseudo-likelihood, penalized with $\boldsymbol{\Phi} = \mathbf{0}$ and $h_{j>i} = 0.5$ and penalized with $\boldsymbol{\Phi} = \mathbf{0}$ and $\boldsymbol{H} = \boldsymbol{H}_{dist}$. Here, $\boldsymbol{H}_{dist}$ is defined in such a way that h_{ij} is linearly proportional to the distance between the ith and jth angles with $\max h_{j>i} = 1$ and diagonal elements $h_{ii} = \min h_{j>i}$. For example,

in the case depicted in Figure 5 the $\boldsymbol{H}_{dist}$ matrix is

$$\begin{pmatrix} 0.5 & 0.5 & 1 & 1 \\ 0 & 0.5 & 0.5 & 1 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}$$

In addition, we include in the comparison the L_1 penalized multivariate von Mises distribution,[8] the multivariate normal and the wrapped-normal multivariate distributions with parameters estimated using the method of moments but shrinking the covariance matrix if it is not positive definite. To get an honest estimation of the KL divergence, we performed repeated train and test validation with 100 repetitions. In each trial, the 50% of the original samples are selected at random without replacement as training set, leaving the remaining portion of samples as test set to compute the KL divergence estimation.

4.3. Results

Results are summarized in Table V. The penalized von Mises distribution obtains similar or better results than the unregularized estimation, specially in the cases where the number of samples is low with respect to the number of variables. Also, we can see how the penalization with the structure-aware confidence matrix $\boldsymbol{H}_{dist}$ provides, in general terms, better estimations that the ones using a uniform confidence matrix.

To assess the statistical significance of the results, we first apply a variation of the Friedman test, a nonparametric equivalent of the repeated ANOVA. The test compares the ranks of each method in different data sets under the null hypothesis that all methods are equivalent.[40] The statistic

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{20}$$

is distributed as an F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom under the null hypothesis. In Equation (20), k is the number of methods that we are comparing (six in our case), N is the number of data sets on which we evaluate our algorithms (13), and χ_F^2 is the Friedman statistic. The test rejects the null hypothesis with a significance level $\alpha = 0.05$, which means that there are significant performance differences between the methods. Then, we compare all methods with the F-norm penalized von Mises estimator with structure-aware confidence matrix $\boldsymbol{H}$ using the Bonferroni–Dunn test. The test finds significant differences ($\alpha = 0.05$) with the nonpenalized von Mises estimation and with the F-norm penalized estimator with uniform confidence matrix. Figure 6 summarizes these results.

The plots in Figure 7 summarize the von Mises distribution fitted for pyramidal neurons with five basal dendrites from the three species: human, rat, and mouse. The boxes in the diagonal show the marginal distribution of the original data for each variable. Each diagonal box contains a rose plot (i.e., a circular histogram) and

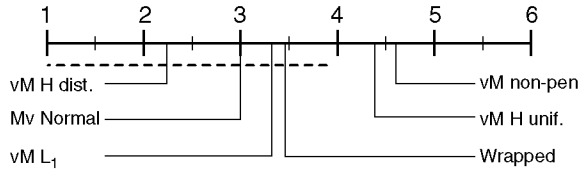


Figure 6. Comparison of F-norm penalized von Mises estimator with structure-aware confidence matrix $\boldsymbol{H}$ against the others with the Bonferroni–Dunn test. All methods with ranks outside the dashed interval are significantly different ($\alpha = 0.05$) from the control.

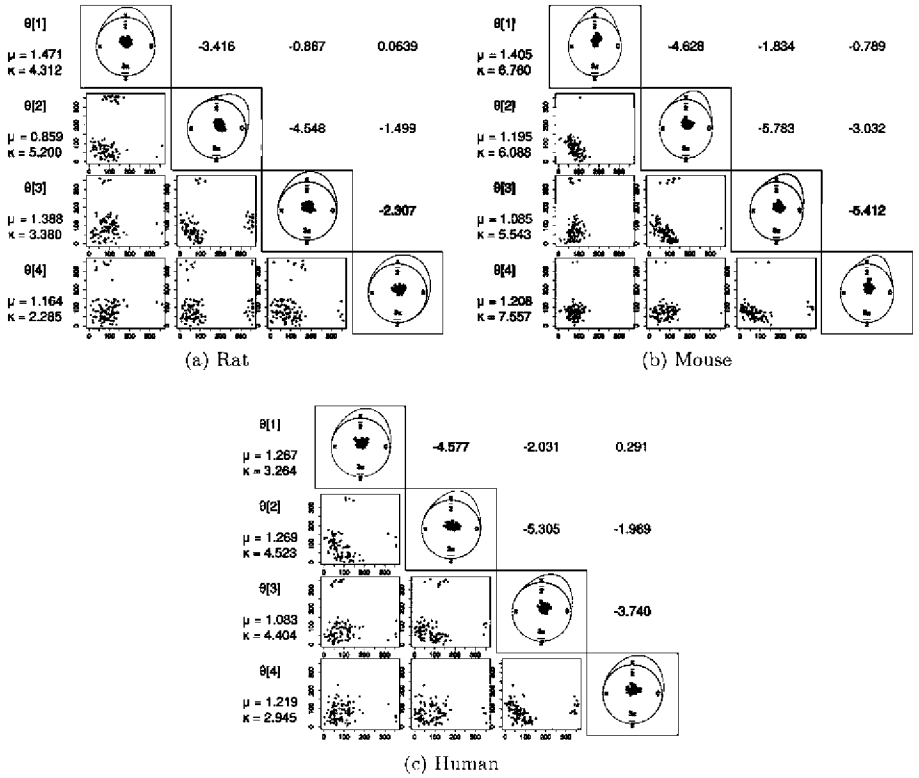


Figure 7. Fitted distributions for the angles between basal dendrites from rat, mouse and human with five basal dendrites ($p = 4$).

the circular density approximated with the von Mises circular kernel. The boxes in the upper triangle are the λ_{ij} parameters of the fitted distributions, whereas the boxes in the lower contain a scatterplot of the variables in that row/column. Finally, the κ_i and μ_i parameters are represented in the first column. We do not observe significant differences between species, except for the anomalous low mean for the second angle for the rat in Figure 7a.

5. CONCLUSION AND FUTURE WORK

In this paper, we have introduced a simplified formulation of the log pseudo-likelihood for the multivariate von Mises distribution that it is also faster to compute, reducing significantly the time needed to obtain the maximum pseudo-likelihood estimator. We have also proposed a penalization term based on the Frobenius norm of the $\mathbf{P}$ matrix that allows to penalize each parameter independently. We have proved the consistency of the penalized estimator, and we have also provided some recommendations to set the penalization parameters. Additionally, we have defined a multivariate circular distance that is easy to compute and takes into account the periodicity of circular data and then used it to compute a nonparametric approximation of the KL divergence between two circular samples. The penalized estimator has been tested in both synthetic and real-world experiments using the KL approximation with the circular distance in the latter case. The results prove that the penalized estimator provides similar or better estimations, specially when the number of samples is low, and we have a reasonable prior knowledge of the dependencies between the variables.

An analytical study of the properties of the maximum pseudo-likelihood estimator could provide consistent rules to set the penalization parameters $\mathbf{\Phi}$ and $\boldsymbol{H}$, whereas the study on efficient approximations of the normalization term of the multivariate von Mises distribution will help to apply this same penalization to the maximum likelihood instead. Another major work line in the future could be to analyze the convergence rate of the penalized estimator compared to the nonregularized estimator. Additionally, it would be interesting to work on a more efficient sampling method based on rejection sampling in high-dimensional settings, for example, by using a mixture of multivariate normal distributions as reference.

All methods described in this paper have been included in the R package *mvCircular*[a] that implements sampling and fitting of the multivariate von Mises distribution as well as multivariate circular plots and statistics.

Acknowledgments

We thank the referees for their review and comments, which significantly contributed to improving the quality of this paper.

This work was partially supported by the Spanish Ministry of Science and Innovation (MCINN) via a doctoral grant to the first author (FPU014/04818), the Spanish Ministry of Economy and Competitiveness (MINECO) 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).

[a] https://github.com/lrodriguezlujan/mvcircular.

References

1. Zhang J, Chowdhury S, Messac A, Castillo L. A multivariate and multimodal wind distribution model. Renew Energy 2013;51:436–447.
2. Masseran N, Razali AM, Ibrahim K, Latif MT. Fitting a mixture of von Mises distributions in order to model data on wind direction in peninsular Malaysia. Energy Convers Manage 2013;72:94–102.
3. Bahlmann C. Directional features in online handwriting recognition. Pattern Recognit 2006;39(1):115–125.
4. Baltieri D, Vezzani R, Cucchiara R. People orientation recognition by mixtures of wrapped distributions on random trees. In: Lecture Notes in Computer Science, Vol 7576. Berlin:Springer;2012. pp 270–283.
5. Masden EA, Reeve R, Desholm M, Fox AD, Furness RW, Haydon DT. Assessing the impact of marine wind farms on birds through movement modelling. J R Soc Interface 2012;9(74):2120–2130.
6. Lagona F. Regression analysis of correlated circular data based on the multivariate von Mises distribution. Environ Ecol Stat 2015;23:1–25.
7. Mardia KV, Hughes G, Taylor CC, Singh H. A multivariate von Mises distribution with applications to bioinformatics. Can J Stat 2008;36(1):99–109.
8. Razavian N, Kamisetty H, Langmead CJ. The von Mises graphical model: regularized structure and parameter learning. Tech. Rep. CMU-CS-11-108, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA; 2011.
9. Mardia KV, Kent JT, Zhang Z, Taylor CC, Hamelryck T. Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. J Appl Stat 2012;39(11):2475–2492.
10. Banerjee A, Dhillon IS, Ghosh J, Sra S. Clustering on the unit hypersphere using von Mises–Fisher distributions. J Mach Learn Res 2005;6:1345–1382.
11. Fisher NI. Statistical analysis of circular data. Cambridge, UK: Cambridge University Press; 1993.
12. Mardia KV, Jupp PE. Directional statistics. Hoboken, NJ: Wiley; 2009.
13. Gatto R, Jammalamadaka SR. The generalized von Mises distribution. Stat Methodol 2007;4(3):341–353.
14. Kato S, Jones M. An extended family of circular distributions related to wrapped Cauchy distributions via Brownian motion. Bernoulli 2013;19(1):154–171.
15. Wang F, Gelfand AE. Directional data analysis under the general projected normal distribution. Stat Methodol 2013;10(1):113–127.
16. Hernandez-Stumpfhauser D, Breidt FJ, van der Woerd MJ. The general projected normal distribution of arbitrary dimension: modeling and Bayesian inference. Bayesian Anal Advance Publication, 19 January 2016.
17. Singh H, Hnizdo V, Demchuk E. Probabilistic model for two dependent circular variables. Biometrika 2002;89(3):719–723.
18. Tan KM, London P, Mohan K, Lee SI, Fazel M, Witten D. Learning graphical models with hubs. J Mach Learn Res 2014;15(1):3297–3331.
19. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. Annals Stat 2008;36(1):199–227.
20. Rodriguez-Lujan L, Bielza C, Larra ñaga P. Regularized multivariate von Mises distribution. In: Lecture Notes in Computer Science, Vol 9422. Berlin: Springer; 2015. pp 25–35.
21. Presnell B, Morrison SP, Littell RC. Projected multivariate linear models for directional data. J Am Stat Assoc 1998;93(443):1068–1077.
22. Fisher R. Dispersion on a sphere. Proc R S London A: Math Phys Eng Sci 1953;217:295–305.
23. Mardia KV, Voss J. Some fundamental properties of a multivariate von Mises distribution. Commun Stat Theory Methods 2014;43(6):1132–1144.

24. Best DJ, Fisher NI. Efficient simulation of the von Mises distribution. Appl Stat 1979;28:152–157.
25. Besag J. Statistical analysis of non-lattice data. Statistician 1975;24(3):179–195.
26. Liang P, Jordan MI. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In: Proc 25th Int Conf on Machine Learning, Helsinki, FI, 5–9 July. New York: ACM; 2008. pp 584–591.
27. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans Math Softw 1997;23(4):550–560.
28. Anderson E, Bai Z, Bischof C, Blackford S, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen DC. LAPACK users' guide, 3rd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics; 1999.
29. Fessler JA. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography. IEEE Trans Image Process 1996;5(3):493–506.
30. Anderson TW. An introduction to multivariate statistical analysis, 3rd ed. Hoboken, NJ: Wiley; 2003.
31. Gelman A, Carlin JB, Stern HS. Bayesian data analysis, 2nd ed. Chapman & Hall/CRC: Boca Raton, FL, USA; 2004.
32. Qi H, Sun D. An augmented Lagrangian dual approach for the h-weighted nearest correlation matrix problem. IMA J Num Anal 2011;31(2):491–511.
33. Rall W, Segev I, Rinzel J, Shepherd GM. The theoretical foundation of dendritic function. Cambridge, MA: MIT Press; 1995.
34. Pérez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions. In: IEEE Int Symp on Information Theory, Toronto, ON, 6–11 July, IEEE; 2008. pp 1666–1670.
35. Ascoli GA, Donohue DE, Halavi M. Neuromorpho.org: a central resource for neuronal morphologies. J Neurosci 2007;27(35):9247–9251.
36. 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. Cerebral Cortex 2001;11(6):558–571.
37. Ballesteros-Yánez I, Valverde O, Ledent C, Maldonado R, DeFelipe J. Chronic cocaine treatment alters dendritic arborization in the adult motor cortex through a CB1 cannabinoid receptor–dependent mechanism. Neuroscience 2007;146(4):1536–1545.
38. Markram H, Muller E, Ramaswamy S, Reimann MW, Abdellah M, Sanchez CA, Ailamaki A, Alonso-Nanclares L, Antille N, Arsever S, Kahou GA, Berger TK, Bilgili A, Buncic N, Chalimourda A, Chindemi G, Courcol JD, Delalondre F, Delattre V, Druckmann S, Dumusc R, Dynes J, Eilemann S, Gal E, Gevaert ME, Ghobril JP, Gidon A, Graham JW, Gupta A, Haenel V, Hay E, Heinis T, Hernando JB, Hines M, Kanari L, Keller D, Kenyon J, Khazen G, Kim Y, King JG, Kisvarday Z, Kumbhar P, Lasserre S, Le Bé JV, Magalhães BR, Merchán-Pérez A, Meystre J, Morrice BR, Muller J, Muñoz-Céspedes A, Muralidhar S, Muthurasa K, Nachbaur D, Newton TH, Nolte M, Ovcharenko A, Palacios J, Pastor L, Perin R, Ranjan R, Riachi I, Rodríguez JR, Riquelme JL, Rössert C, Sfyrakis K, Shi Y, Shillcock JC, Silberberg G, Silva R, Tauheed F, Telefont M, Toledo-Rodriguez M, Tränkler T, Van Geit W, Díaz JV, Walker R, Wang Y, Zaninetta SM, DeFelipe J, Hill SL, Segev I, Schürmann F. Reconstruction and simulation of neocortical microcircuitry. Cell 2015;163(2):456–492.
39. Ramaswamy S, Courcol JD, Abdellah M, Adaszewski SR, Antille N, Arsever S, Atenekeng G, Bilgili A, Brukau Y, Chalimourda A, Chindemi G, Delalondre F, Dumusc R, Eilemann S, Gevaert ME, Gleeson P, Graham JW, Hernando JB, Kanari L, Katkov Y, Keller D, King JG, Ranjan R, Reimann MW, Rössert C, Shi Y, Shillcock JC, Telefont M, Van Geit W, Diaz JV, Walker R, Wang Y, Zaninetta SM, DeFelipe J, Hill SL, Muller J, Segev I, Schürmann F, Muller EB, Markram H. The neocortical microcircuit collaboration portal: a resource for rat somatosensory cortex. Front Neural Circuits 2015;9(44).
40. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.