Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

The Multivariate Generalised von Mises Distribution: Inference and Applications

Alexandre K. W. Navarro
Department of Engineering
University of Cambridge
Cambridge, UK
akwn2@cam.ac.uk

Jes Frellsen
Department of Engineering
University of Cambridge
Cambridge, UK
jf519@cam.ac.uk

Richard E. Turner
Department of Engineering
University of Cambridge
Cambridge, UK
ret26@cam.ac.uk

Abstract

Circular variables arise in a multitude of data-modelling contexts ranging from robotics to the social sciences, but they have been largely overlooked by the machine learning community. This paper partially redresses this imbalance by extending some standard probabilistic modelling tools to the circular domain. First we introduce a new multivariate distribution over circular variables, called the multivariate Generalised von Mises (mGvM) distribution. This distribution can be constructed by restricting and renormalising a general multivariate Gaussian distribution to the unit hyper-torus. Previously proposed multivariate circular distributions are shown to be special cases of this construction. Second, we introduce a new probabilistic model for circular regression inspired by Gaussian Processes, and a method for probabilistic Principal Component Analysis with circular hidden variables. These models can leverage standard modelling tools (e.g. kernel functions and automatic relevance determination). Third, we show that the posterior distribution in these models is a mGvM distribution which enables development of an efficient variational free-energy scheme for performing approximate inference and approximate maximum-likelihood learning.

1 Introduction

Many data modelling problems in science and engineering involve circular variables. For example, the spatial configuration of a molecule (Boomsma et al. 2008; Frellsen et al. 2009), robot, or the human body (Chirikjian and Kyatkin 2000) can be naturally described using a set of angles. Phase variables arise in image and audio modelling scenarios (Wadhwa et al. 2013), while directional fields are also present in fluid dynamics (Jona-Lasinio, Gelfand, and Jona-Lasinio 2012), and neuroscience (Ben-Yishai, Bar-Or, and Sompolinsky 1995). Phase-locking to periodic signals occurs in a multitude of fields ranging from biology (Gao et al. 2010) to the social sciences (Brunsdon and Corcoran 2006).

It is possible, at least in principle, to model circular variables using distributional assumptions that are appropriate for variables that live in a standard Euclidean space. For example, a naïve application might represent a circular variable in terms of its angle $\phi \in [0, 2\pi)$ and use a standard distribution over this variable (presumably restricted to the valid domain). Such an approach would, however, ignore the topology of the space e.g. that $\phi = 0$ and $\phi = 2\pi$ are equivalent. Alternatively, the circular variable can be represented as a unit vector in $\mathbb{R}^2$, $\boldsymbol{x} = [\cos(\phi), \sin(\phi)]^\top$, and a standard bivariate distribution used instead. This partially alleviates the aforementioned topological problem, but standard distributions place probability mass off the unit circle which adversely affects learning, prediction and analysis.

In order to predict and analyse circular data it is therefore key that machine learning practitioners have at their disposal a suite of bespoke modelling, inference and learning methods that are specifically designed for circular data (Lebanon 2005). The fields of circular and directional statistics have provided a toolbox of this sort (Mardia and Jupp 2000). However, the focus has been on fairly simple and small models that are applied to small datasets enabling MCMC to be tractably deployed for approximate inference.

The goal of this paper is to extend the existing toolbox provided by statistics, by leveraging modelling and approximate inference methods from the probabilistic machine learning field. Specifically, the paper makes three technical contributions. First, in Section 3 it introduces a central multivariate distribution for circular data—called the multivariate Generalised von Mises distribution—that has elegant theoretical properties and which can be combined in a plug-and-play manner with existing probabilistic models. Second, in Section 4 it shows that this distribution arises in two novel models that are circular versions of Gaussian Process regression and probabilistic Principal Component Analysis with circular hidden variables. Third, it develops efficient approximate inference and learning techniques based on variational free-energy methods as demonstrated on four datasets in Section 6.

2 Circular distributions primer

In order to explain the context and rationale behind the contributions made in this paper, it is necessary to know a little background on circular distributions. Since *multidimensional* circular distributions are not generally well-known in the machine learning community, we present a brief review of the main concepts related to these distributions in this section. The expert reader can jump to Section 3 where the multivariate Generalised von Mises distribution is introduced.

A univariate circular distribution is a probability distribution defined over the unit circle. Such distributions can

be constructed by wrapping, marginalising or conditioning standard distributions defined in Euclidean spaces and are classified as *wrapped*, *projected* or *intrinsic* according to the geometric interpretation of their construction.

More precisely, the *wrapped* approach consists of taking a univariate distribution $p(x)$ defined on the real line, parametrising any point $x \in \mathbb{R}$ as $x = \phi + 2\pi k$ with $k \in \mathbb{Z}$ and summing over all k so that $p(x)$ is wrapped around the unit circle. The most commonly used wrapped distribution is the Wrapped Gaussian distribution (Ferrari 2009; Jona-Lasinio, Gelfand, and Jona-Lasinio 2012).

An alternative approach takes a standard bivariate distribution $p(x, y)$ that places probability mass over $\mathbb{R}^2$, transforms it to polar coordinates $[x, y]^\top \rightarrow [r\cos\phi, r\sin\phi]^\top$ and marginalises out the radial component $\int_0^\infty p(r\cos\phi, r\sin\phi) r \mathrm{d}r$. This approach can be interpreted as projecting all the probability mass that lies along a ray from the origin onto the point where it crosses the unit circle. The most commonly used projected distribution is the Projected Gaussian (Wang and Gelfand 2013).

Instead of marginalising the radial component, circular distributions can be constructed by conditioning it to unity, $p(x, y|x^2+y^2 = 1)$. This can be interpreted as restricting the original bivariate density to the unit circle and renormalising. A distribution constructed in this way is called "intrinsic" (to the unit circle). The construction has several elegant properties. First, the resulting distribution inherits desirable characteristics of the base distribution, such as membership of the exponential family. Second, the form of the resulting density often affords more analytical tractability than those produced by wrapping or projection. The most important intrinsic distribution is the von Mises (vM), $p(\phi|\mu, \kappa) \propto \exp(\kappa\cos(\phi-\mu))$, which is obtained by conditioning an isotropic bivariate Gaussian to the unit circle. The vM has two parameters, the mean $\mu \in [0, 2\pi)$ and the concentration $\kappa \in \mathbb{R}^+$. If the covariance matrix of the bivariate Gaussian is a general real positive definite matrix, we obtain the Generalised von Mises (GvM) distribution (Gatto and Jammalamadaka 2007)[1]

$$p(\phi) \propto \exp(\kappa_1 \cos(\phi - \mu_1) + \kappa_2 \cos(2(\phi - \mu_2))) . \quad (1)$$

The GvM has four parameters, two mean-like parameters $\mu_i \in [0, 2\pi)$ and two concentration-like parameters $\kappa_i \in \mathbb{R}^+$ and is an exponential family distribution. The GvM is generally asymmetric. It has two modes when $4\kappa_2 \geq \kappa_1$, otherwise it has one mode except when it is a uniform distribution $\kappa_2 = \kappa_1 = 0$.

The GvM is arguably more tractable than the distributions obtained by wrapping or projection as its unnormalised density takes a simple form. In comparison, the unnormalised density of the wrapped normal involves an infinite sum and that of the projected normal is complex and requires special functions. However, the normalising constant of the GvM (and its higher moments) are still complicated, containing infinite sums of modified Bessel functions (Gatto 2008).

[1]To be precise, Gatto and Jammalamadaka define this to be a Generalised von Mises of order 2, but since higher-order Generalised von Mises distributions are more intractable and consequently have found fewer applications, we use the shorthand throughout.

In this paper the focus will be on the extensions to vectors of dependent circular variables that lie on a (hyper-) torus (although similar methods can be applied to multivariate hyper-spherical models). An example of a multivariate distribution on the hyper-torus is the multivariate von Mises (mvM) by Mardia et al. (2008)

$$mv\mathcal{M}(\boldsymbol{\phi}) \propto \exp\left\{\boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi}) + \sin(\boldsymbol{\phi})^\top \boldsymbol{G} \sin(\boldsymbol{\phi})\right\} . \quad (2)$$

The terms $\cos(\boldsymbol{\phi})$ and $\sin(\boldsymbol{\phi})$ denote element-wise application of sine and cosine functions to the vector $\boldsymbol{\phi}$, $\boldsymbol{\kappa}$ is a element-wise positive D-dimensional real vector, $\boldsymbol{\nu}$ is a D-dimensional vector whose entries take values on $[0, 2\pi)$, and $\boldsymbol{G}$ is a matrix whose diagonal entries are all zeros.

The mvM distribution draws its name from the its property that the one dimensional conditionals, $p(\phi_d|\boldsymbol{\phi}_{\neq d})$, are von Mises distributed. As shown in the Supplementary Material, this distribution can be obtained by applying the *intrinsic* construction to a $2D$-dimensional Gaussian, mapping $\boldsymbol{x} \rightarrow (r\cos\boldsymbol{\phi}^\top, r\sin\boldsymbol{\phi}^\top)^\top$ and assuming its precision matrix has the form

$$\boldsymbol{W} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \Lambda & \boldsymbol{A} \\ \boldsymbol{A}^\top & \Lambda \end{bmatrix} \quad (3)$$

where Λ is a diagonal D by D matrix and $\boldsymbol{A}$ is an antisymmetric matrix. Other important facts about the mvM are that it bears no simple closed analytic form for its normalising constant, it has $D + (D-1)D/2$ degrees of freedom in its parameters and it is not closed under marginalisation.

We will now consider multi-dimensional extensions of the GvM distribution.

3 The multivariate Generalised von Mises

In this section, we present the multivariate Generalised von Mises (mGvM) distribution as an *intrinsic* circular distribution on the hyper-torus and relate it to existing distributions in the literature. Following the construction of *intrinsic* distributions, the multivariate Generalised von Mises arises by constraining a $2D$-dimensional multivariate Gaussian with arbitrary mean and covariance matrix to the D-dimensional torus. This procedure yields the distribution

$$m\mathcal{G}v\mathcal{M}(\boldsymbol{\phi}; \boldsymbol{\nu}, \boldsymbol{\kappa}, \boldsymbol{W}) \propto \exp\left\{\boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi} - \boldsymbol{\nu}) - \frac{1}{2}\begin{bmatrix}\cos(\boldsymbol{\phi}) \\ \sin(\boldsymbol{\phi})\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{W}^{cc} & \boldsymbol{W}^{cs} \\ (\boldsymbol{W}^{cs})^\top & \boldsymbol{W}^{ss}\end{bmatrix} \begin{bmatrix}\cos(\boldsymbol{\phi}) \\ \sin(\boldsymbol{\phi})\end{bmatrix}\right\} \quad (4)$$

where $\boldsymbol{W}^{cc}, \boldsymbol{W}^{cs}, \boldsymbol{W}^{ss}$ are the blocks of the underlying Gaussian precision matrix $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\nu}$ is a D-dimensional angle vector and $\boldsymbol{\kappa}$ is a D-dimensional concentration vector. Equation (4) is over-parametrised with $2D + 3(D-1)D/2$ parameters, D more than the degrees of freedom of the most succinct form of the mGvM given in Supplemental Material.

The mGvM distribution generalises the multivariate von Mises by Mardia et al. (2008); it collapses to the mvM when $\boldsymbol{W}$ has the form of Equation (3). Whereas the one-dimensional conditionals of the mvM are von Mises and therefore unimodal and symmetric, those of the mGvM are generalised von Mises and therefore can be bimodal and

asymmetric. The mGvM also captures a richer set of dependencies between the variables than the mvM, notice that the mvM is not the most general form of mGvM that has vM conditionals. The tractability of the one-dimensional conditionals of the mGvM can be leveraged for approximate inference using variational mean-field approximations and Gibbs sampling (see Section 5). The mGvM is a member of the exponential family and a maximum entropy distribution subject to multidimensional first and second order circular moments constraints. We will now show that the mGvM can be used to build rich probabilistic models for circular data.

4 Some applications of the mGvM

In this section, we outline two novel and important probabilistic models in which inference produces a posterior distribution that is a mGvM. The first model is a circular analogue of Gaussian Process regression and the second is a version of Principal Component Analysis for circular latent variables.

4.1 Regression of circular data

Consider a regression problem in which a set of noisy output circular variables $\{\psi_n\}_{n=1}^N$ have been collected at a number of input locations $\{\boldsymbol{s}_n\}_{n=1}^N$. The treatment will apply to inputs that can be multi-dimensional and lie in any space (e.g. they could be circular themselves). The goal is to predict circular variables $\{\psi_m^*\}_{m=1}^M$ at unseen input points $\{\boldsymbol{s}_m^*\}_{m=1}^M$. Here we leverage the connection between the mGvM distribution and the multivariate Gaussian in order to produce a powerful class of probabilistic models for this purpose based upon Gaussian Processes. In what follows the outputs and inputs will be represented as vectors and matrices respectively, that is $\boldsymbol{\psi}$, $\mathcal{S}$, $\boldsymbol{\psi}^*$ and $\mathcal{S}^*$.

In standard Gaussian Process regression (Rasmussen and Williams 2006) a multivariate Gaussian prior is placed over the underlying unknown function values at the input points $p(\boldsymbol{f}|\mathcal{S}) = \mathcal{GP}(\boldsymbol{f}; 0, \boldsymbol{K}(\boldsymbol{s}, \boldsymbol{s}'))$, and a Gaussian noise model is assumed to produce the observations at each input location, $p(y_n|f_n, \boldsymbol{s}_n) = \mathcal{N}(y_n; f_n, \sigma_y^2)$. The prior over the function values is specified using the Gaussian Process's covariance function $K(\boldsymbol{s}, \boldsymbol{s}')$ that encapsulates prior assumptions about the properties of the underlying function, such as smoothness, periodicity, stationarity etc. Prediction then involves forming the posterior predictive distribution, $p(\boldsymbol{f}^*|\boldsymbol{y}, \mathcal{S}, \mathcal{S}^*)$, which also takes a Gaussian form due to conjugacy.

Here an analogous approach is taken. The circular underlying function values and observations are denoted $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. The prior over the underlying function is given by a mGvM in overparametrised form $p(\boldsymbol{\phi}|\mathcal{S}) = m\mathcal{G}\upsilon\mathcal{M}(\boldsymbol{\phi}; 0, 0, \boldsymbol{K}(\boldsymbol{s}, \boldsymbol{s}')^{-1})$ and the observations are assumed to be von Mises noise corrupted versions of this function $p(\psi_n|\phi_n, \boldsymbol{s}_n) = \upsilon\mathcal{M}(\psi_n; \phi_n, \kappa)$. In order to construct a sensible prior over circular function values we use a construction that is inspired by a multi-output GP to produce bivariate variables at each input location. We then leverage the intrinsic construction of the mGvM to constrain each regressed point to the unit circle to allow the mGvM to inherit the properties from the GP covariance function it was built from. This is central to creating a flexible and powerful mGvM regression framework, as GP covariance functions that can handle exotic input variables such as circular variables, strings or graphs (Gärtner, Flach, and Wrobel 2003; Duvenaud, Nickisch, and Rasmussen 2011).

Inference proceeds subtly differently to that in a GP due to an important difference between multivariate Gaussian and multivariate Generalised von Mises distributions. That is, the former are consistent under marginalisation whilst the latter are not: if a subset of mGvM variables are marginalised out, the remaining variables are not distributed according to a mGvM. Technically, this means that for analytic tractability of inference we have to handle the joint posterior predictive distribution $p(\boldsymbol{\phi}, \boldsymbol{\phi}^*|\boldsymbol{\psi}, \mathcal{S}, \mathcal{S}^*)$, which is a mGvM due to conjugacy, rather than $p(\boldsymbol{\phi}^*|\boldsymbol{\psi}, \mathcal{S}, \mathcal{S}^*)$, which is not. Whilst this is somewhat less elegant than GP regression as it requires the prediction locations to be known up front, in many applications this is not a great restriction. This model type is termed transductive (Quiñonero-Candela and Rasmussen 2005).

4.2 Latent angles: dimensionality reduction and representation learning

Next consider the task of learning the motion of an articulated rigid body from noisy measurements on a Euclidean space. Articulated rigid bodies can represent a large class of physical problems including mechanical systems, human motion and molecular interactions. The dynamics of rigid bodies can also be fully described by rotations around a fixed point plus a translation and, therefore, can be succinctly represented using angles see (Chirikjian and Kyatkin 2000). For simplicity, we will restrict our treatment to a rigid body with D articulations on a 2-dimensional Euclidean space and rotations only, as the discussion trivially generalises to higher dimensional spaces and translations can be incorporated through an extra linear term. Extensions for 3-dimensional models follow directly from the 2-dimensional case, which can be seen as a first step towards these more complex models.

The Euclidean components of any point on an articulated rigid body can be described using the angles between each articulation and their distances. More precisely, for an upright, counter-clockwise coordinate system, the horizontal and vertical components of a point in the d-th articulator can be written as $x_d = \sum_{j=1}^d l_j \sin(\boldsymbol{\varphi}_j)$ and $y_d = -\sum_{j=1}^d l_j \cos(\boldsymbol{\varphi}_j)$, where l_j is the length of a link j to the next link or the marker. Without loss of generality, we can model only the variation around the mean angle for each joint, i.e. $\varphi_d = \phi_d - \nu_d$ which results in the general model for noisy measurements

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix} = \begin{bmatrix}-\boldsymbol{L}\\ \boldsymbol{L}\end{bmatrix}\begin{bmatrix}\cos(\boldsymbol{\varphi})\\ \sin(\boldsymbol{\varphi})\end{bmatrix} + \boldsymbol{\epsilon} = \begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix}\begin{bmatrix}\cos(\boldsymbol{\phi})\\ \sin(\boldsymbol{\phi})\end{bmatrix} + \boldsymbol{\epsilon} \quad (5)$$

where $\boldsymbol{L}$ is the matrix that encodes the distances between joints, $\boldsymbol{A}$ and $\boldsymbol{B}$ are the distance matrix rotated by the vector $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$. The prior over the joint angles can be modelled by a multivariate Generalised von Mises. Here we take inspiration from Principal Component Analysis, and use independent von Mises distributions

$$p(\boldsymbol{\phi}_{1,\ldots,N}) = \prod_{n=1}^N \prod_{d=1}^D \upsilon\mathcal{M}\left(\phi_{d,n}; 0, \kappa_d\right). \quad (6)$$

Due to conjugacy, the posterior distribution over the latent angles is a mGvM distribution. This can be informally verified by noting that the priors on the latent angles $\boldsymbol{\phi}$ are exponentials of linear functions of sines and cosines, while the likelihood is the exponential of a quadratic function in sine and cosines. This leads to the posterior being an exponential quadratic function of sines and cosines and, hence, mGvM.

The model can be extended to treat the parameters in Bayesian way by including sparse priors over the coefficient matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and the observation noise. A standard choice for this task is to define Automatic Relevance Detection priors (MacKay 1994) over the columns of these matrices defined as $\mathcal{N}(\boldsymbol{A}_{m,d}; 0, \sigma^2_{\boldsymbol{A},d})$ and $\mathcal{N}(\boldsymbol{B}_{m,d}; 0, \sigma^2_{\boldsymbol{B},d})$ in order to perform automatic structure learning. Additional Inverse Gamma priors over $\sigma^2_{\boldsymbol{A},d}$, $\sigma^2_{\boldsymbol{B},d}$ and σ^2 are also employed.

The dimensionality of the latent angle space can be lower than the dimensionality of the observed space, in which case learning and inference perform dimensionality reduction that maps real-valued data to a lower-dimensional torus. Besides motion capture, toroidal manifolds can also prove useful when modelling other relevant applications, such as electroencephalogram (EEG) and audio signals (Turner and Sahani 2011a). Further connections between dimensionality reduction with the mGvM and Probabilistic Principal Component Analysis (PPCA) proposed by Tipping and Bishop (1999) (including limiting behaviour and geometrical relations between these models) are explored in the Supplementary Material. As a consequence of these similarities, we denote this model as circular Principal Component Analysis (cPCA).

5 Approximate inference for the mGvM

The multivariate Generalised von Mises does not admit an analytic expression for its normalizing constant, therefore we need to resort to approximate inference techniques. This section presents two approaches that exploit the tractable univariate conditionals of the mGvM: Gibbs sampling and mean-field variational inference.

5.1 Gibbs sampling

A Gibbs sampling procedure for sampling the mGvM of Equation (4) can be derived leveraging the GvM form of the one-dimensional mGvM conditionals. In particular, the Gibbs sampler updates for the d-th conditional of the mGvM will have the form

$$p(\phi_d|\boldsymbol{\phi}_{\neq d}) = \mathcal{GvM}(\phi_d; \tilde{\kappa}_{1,d}, \kappa_{2,d}, \tilde{\nu}_{1,d}, \nu_{2,d}) \qquad (7)$$

where $\tilde{\kappa}_{1,d}$ and $\tilde{\nu}_{1,d}$ are functions of $\boldsymbol{\kappa}$, $\boldsymbol{\nu}$ and $\boldsymbol{\phi}_{\neq d}$ given in the Supplementary Material.

The Gibbs sampler can be used to support approximate maximum-likelihood learning by using it to compute the expectations required by the EM algorithm (Wei and Tanner 1990). However, it is well-known that Gibbs sampling becomes less effective as the joint distribution becomes more correlated and the dimensionality grows. This is particularly significant when using the distribution in high-dimensional cases with rich correlational structure, such as those considered later in the paper.

5.2 Mean-field Variational inference

As a consequence of the problems encountered when using Gibbs sampling, the variational inference framework emerges as an attractive, scalable approach to handling inference in when the posterior distribution is a mGvM.

The variational inference framework (Jordan et al. 1999) aims to approximate an intractable posterior $p(\boldsymbol{\phi}|\boldsymbol{\psi}, \theta)$ with a distribution $q(\boldsymbol{\phi}|\rho)$ by minimising the Kullback-Leiber divergence from the distribution q to p. If the approximating distribution is chosen to be fully factored, i.e. $q(\boldsymbol{\phi}) = \prod_{d=1}^{d} q_d(\phi_d)$, the optimal functional form for $q_d(\phi_d)$ can be obtained analytically using calculus of variations. The functional form of each mean-field factor is inherited from the one-dimensional conditionals and consequently is a Generalised von Mises of the form

$$q_d(\phi_d) = \mathcal{GvM}(\phi_d; \bar{\kappa}_{1,d}, \kappa_{2,d}, \bar{\nu}_{1,d}, \nu_{2,d})$$

where the formulas for the parameters $\bar{\kappa}_{1,d}$ and $\bar{\nu}_{1,d}$ are similar in nature to the Gibbs sampling update and given in the Supplementary Material.

Furthermore, since the moments of the Generalised von Mises can be computed through series approximations (Gatto 2008), the errors from series truncation are negligible if a sufficiently large number of terms is considered. It is possible to obtain gradients of the variational free energy and optimise it with standard optimisation methods such as Conjugate Gradient or Quasi-Newton methods instead of resorting to coordinate-ascent under the variational Expectation-Maximization algorithm which often is slow to converge.

Despite these improvements, we found empirically that accurate calculations of the moments of a Generalised von Mises distribution can become costly when the magnitude of the concentration parameters exceeds $\approx$100 and the posterior concentrates. This numerical instability occurs when the infinite expansion for computing the moments contains a large number of significant terms that have alternating signs leading to accumulation of numerical errors. It is possible to use other approximate integrations schemes if these cases arise during inference. An alternative way to alleviate this problem is to consider a sub-optimal form of factorised approximating distribution. An obvious choice is to use von Mises factors as this results in tractable updates and requires simpler moment calculations. A von Mises field can also be motivated as a first order approximation to a GvM field by requiring that the log approximating distribution is linear in sine and cosine terms, as shown in the Supplementary Material.

In addition to inference, we can use the same variational framework for learning in cases where the mGvM we wish to approximate is a posterior of tractable likelihoods and priors, as in the cPCA model. To achieve this, we form the variational free-energy lower bound on the log-marginal likelihood as

$$\log p(\boldsymbol{\psi}|\theta) \geq \mathcal{F}(q, \theta) = \langle \log p(\boldsymbol{\psi}, \boldsymbol{\phi}|\theta) \rangle_{q(\boldsymbol{\phi}|\rho)} + \mathcal{H}(q),$$

where $\mathcal{H}(q)$ is the entropy of the approximating distribution and $q(\boldsymbol{\phi}|\rho)$, $p(\boldsymbol{\phi}, \boldsymbol{\psi}|\theta)$ is the model log-joint distribution, $\mathcal{F}(q, \theta, \rho)$ is the variational free-energy, θ are the model parameters and ρ represents the parameters of the approximating distribution. The same bound cannot be used directly

for doubly-intractable mGvM models, such as the circular regression model, and it constitutes an area for further work.

6 Experimental results

To demonstrate approximate inference on the applications outlined in Section 4 we present experiments on synthetic and real datasets. A comprehensive description and the data sets used in the all experiments conducted are available at http://tinyurl.com/mgvm-release. Further experimental details are also provided in the Supplementary Material.

6.1 Comparison to other circular distributions

For illustrative purposes we qualitatively compared multivariate Wrapped Gaussian and mvM approximations to a base mGvM and mGvM approximations to these two distributions. The approximations were obtained by numerically minimising the KL divergence between the approximating distribution and the base distributions. These experiments were conducted on a two-dimensional setting in order to render the computation of the normalising constant of the mGvM and the mvM tractable by numerical integration. The resulting distributions are shown in Figure 1.

In Figure 1, the mvM and the multivariate wrapped Gaussian cannot capture the multimodality and asymmetry of the mGvM. Moreover, these distributions approximate the multiple modes by increasing their variance and assigning high probability to the region of low-probability between the modes of the mGvM. On the other hand, when the mvM and the multivariate wrapped Gaussian are approximated by the mGvM, the mGvM is able to approximate well the high-probability zones of the wrapped Gaussian and its unimodality and fully recover the mvM.

We also compared the performance of Gibbs sampling and variational inference for a bivariate GvM. To compare the approximate inference procedures, we analysed the run time for each method and the error it produced in terms of the KL divergence between the true distribution and the approximations on a discretized grid. The Gibbs sampling procedure required a total of 3466 samples and $3.1\,\mathrm{s}$ to achieve the same level of error as the variational approximation achieved after $0.02\,\mathrm{s}$. The variational approach was considerably more efficient than Gibbs sampling, and theory suggests this behaviour holds for higher dimensions, see David MacKay (2003).

6.2 Regression with the mGvM

In this section, we investigate the advantages of employing the mGvM regression model discussed in Section 4.1 over two common approaches to handling circular data in machine learning contexts.

The first approach is to ignore the circular nature of the data and fit a non-circular model. This approach is not infrequent as it is reasonable in contexts where angles are constrained to a subset of the unit circle and there is no wrappping. A typical example of the motivation for such models is the use of a first-order Taylor approximation to the rate of change of an angle as can be found in classical aircraft control applications. To represent this approach to modeling, we will fit a one-dimensional GP (1D-GP) to the data sets.

Table 1: Log-likelihood score for regression with the mGvM, 1D-GP and 2D-GP on validation data.

Data set	mGvM	1D-GP	2D-GP
Toy	$\mathbf{2.02 \cdot 10^4}$	$-1.62 \cdot 10^3$	$8.28 \cdot 10^2$
Uber	$\mathbf{3.29 \cdot 10^4}$	$-1.49 \cdot 10^3$	$-2.83 \cdot 10^2$
Tides	$\mathbf{1.25 \cdot 10^4}$	$-6.46 \cdot 10^4$	$-8.41 \cdot 10^1$
Protein	$\mathbf{1.42 \cdot 10^5}$	$-3.34 \cdot 10^5$	$1.28 \cdot 10^5$
Yeast	$\mathbf{1.33 \cdot 10^2}$	$-1.46 \cdot 10^2$	$-1.65 \cdot 10^1$

The second approach tries to address the circular behaviour by regressing the sine and cosine of the data. In this approach, the angle can be extracted by taking the arc tangent of the ratio between sine and cosine components. While this approach partially addresses the underlying topology of the data, the uncertainty estimates for a non-circular model can be poorly calibrated. Here, each data point is modeled by a two-dimensional vector with the sine and cosine of each data point using a two-dimensional GP (2D-GP).

Five data sets were used in this evaluation. A toy data set generated by wrapping a Mexican hat function around the unit circle, a dataset consisting Uber ride requests in NYC in April 2014[2], the tide levels predictions from the UK Hydrographic Office in 2016[3] as function of the latitude and longitude of a given port, the first side chain angle of aspartate as a function of backbone angles in proteins (Harder et al. 2010), and yeast cell cycle phase as a function of gene expression (Santos, Wernersson, and Jensen 2015).

To assess how well the fitted models approximate the distribution of the data, a subset of the data points was kept for validation and the models scored in terms of the log likelihood of the validation data set. To guarantee fairness in the comparison, the likelihood of the 2D-GP was projected back to the unit circle by marginalising the radial component of the model for each point. This converts the 2D-GP into a one-dimensional projected Gaussian distribution over angles. The results are summarised in Table 1.

The results shown in Table 1 indicate that the mGvM provides a better overall fit than the 1D-GP and the 2D-GP in all experiments. The 1D-GP approach performs poorly in every case studied as it cannot account for the wrapping behaviour of circular data. The 2D-GP performs better than the 1D-GP, however in the Uber, Tides and Yeast datasets its performance is substantially closer to the one presented by the 1D-GP case rather than the mGvM. The toy dataset is examined in Figure 2, showing the 2D-GP learns a different underlying function and cannot capture bimodality.

6.3 Dimensionality reduction

To demonstrate the dimensionality reduction application, we analysed two datasets: one motion capture dataset comprising marker positions placed on a subject's arm and captured through a low resolution camera and another set comprising of a noisy simulation of a 4-DOF robot arm under the same

[2] https://github.com/fivethirtyeight/uber-tlc-foil-response

[3] http://www.ukho.gov.uk/Easytide/easytide/SelectPort.aspx

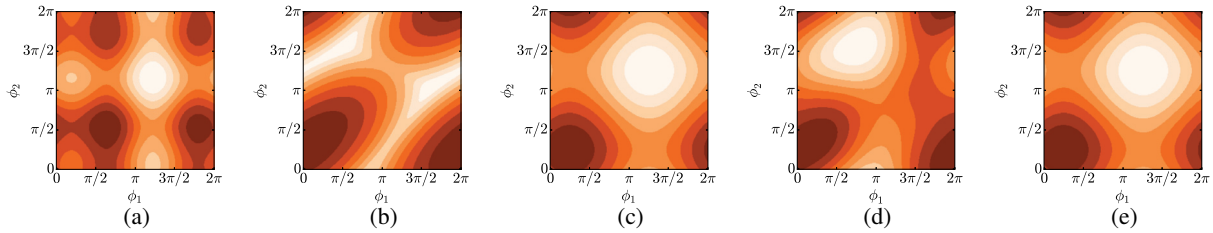

Figure 1: Circular distribution approximations: a base mGvM (a) and its optimal approximations using a multivariate wrapped Gaussian (b) and the mvM (c). The mGvM approximation to the mWG (b) is presented in (d) and the mGvM approximation to the mvM in (c) is presented in (e). Neither the multivariate wrapped Gaussian nor the mvM can accommodate for the asymmetries and the multiple modes of the mGvM, however, the mGvM is able to approximate the mWG high-probability regions and fully recover the mvM. Darker regions have higher probability than lighter regions.

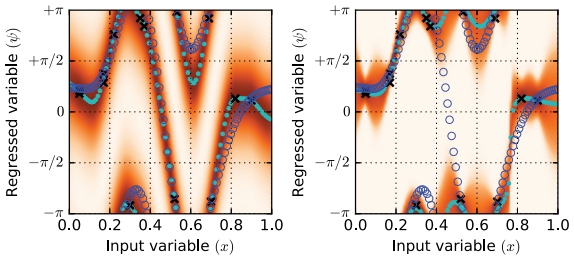

Figure 2: Regression on a toy data set using the mGvM (left) and 2D GP (right): data points are denoted by crosses, the true function by circles and predictions by solid dots.

motion capture conditions.

We compared the model using point estimates for the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, a variational Bayes approach by including ARD priors for $\boldsymbol{A}$ and $\boldsymbol{B}$, Probabilistic Principal Component Analysis (PPCA) (Tipping and Bishop 1999) and the Gaussian Process Latent Variable Model (GP-LVM) (Lawrence 2004) using a squared exponential kernel and a linear kernel. The models using the mGvM require special attention to initialisation. To initialise the test, we used a greedy clustering algorithm to estimate the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The variational Bayes model was initialised using the learned parameters for the point estimate model.

The performance of each model was assessed by denoising the original dataset corrupted by additional Gaussian noise of 2.5, 5 and 10 pixels and comparing the signal-to-noise ratio (SNR) on a test dataset. The best results after initializing the models at 3 different initial starting points are summarized in Table 2 and additional experiments for a wider range of noise levels are available in the Supplemental Material. In Table 2, the point estimate cPCA model performs best and is followed by its variational Bayes version for both datasets (the poor performance of the variational Bayes version is likely to be due to biases that can affect variational methods (Turner and Sahani 2011b)). In the motion capture dataset, the latent angles are highly concentrated. Under these circumstances, the small-angle approximation for sine and cosine provides good results and the cPCA model degenerates into the PPCA model as shown in the Supplementary Material. This behaviour is reflected in the proximity of the PPCA and

Table 2: Signal-to-noise ratio (dB) of the learned latent structure after denoising corrupted signals with by Gaussian noise.

Model	Motion Capture			Robot		
	2.5	5	10	2.5	5	10
cPCA-Point	**29.6**	**23.5**	**17.6**	**33.5**	**30.0**	**24.9**
cPCA-VB	24.6	21.9	17.6	33.2	29.8	24.8
PPCA	23.6	20.9	17.2	22.3	21.8	20.5
GPLVM-SE	8.6	8.5	8.2	21.8	15.7	15.2
GPLVM-L	11.0	7.5	8.1	24.0	16.6	15.9

cPCA signal to noise ratios in Table 2. In the robot dataset, the latent angles are less concentrated. As a result, the behaviour of the PPCA and cPCA models is different which explains the larger gap between the results obtained for these models.

7 Conclusions

In this paper we have introduced the multivariate Generalised von Mises, a new circular distribution with novel applications in circular regression and circular latent variable modelling in a first attempt to close the gap between circular statistics and the machine learning communities. We provided a brief review of the construction of circular distributions including the connections between the Gaussian distribution and the multivariate Generalised von Mises. We provided a scalable way to perform inference on the mGvM model through the variational free energy framework and demonstrated the advantages of the mGvM over GP and mvM through a series of experiments.

8 Acknowledgements

AKWN thanks CAPES grant BEX 9407-11-1. JF thanks the Danish Council for Independent Research grant 0602-02909B. RET thanks EPSRC grants EP/L000776/1 and EP/M026957/1.

References

Ben-Yishai, R.; Bar-Or, R.; and Sompolinsky, H. 1995. Theory of orientation tuning in visual cortex. *Proceedings of*

the National Academy of Sciences of the United States of America 92(9):38443848.

Boomsma, W.; Mardia, K. V.; Taylor, C. C.; Ferkinghoff-Borg, J.; Krogh, A.; and Hamelryck, T. 2008. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences* 105(26):8932–8937.

Brunsdon, C., and Corcoran, J. 2006. Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems* 30(3):300–319.

Chirikjian, G. S., and Kyatkin, A. B. 2000. *Engineering Applications of Noncommutative Harmonic Analysis: With Emphasis on Rotation and Motion Groups*. Abingdon: CRC Press.

David MacKay. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press, 2 edition.

Duvenaud, D. K.; Nickisch, H.; and Rasmussen, C. E. 2011. Additive gaussian processes. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc. 226–234.

Ferrari, C. 2009. *The Wrapping Approach for Circular Data Bayesian Modelling*. Ph.D. Dissertation, Università di Bologna, Bologna.

Frellsen, J.; Moltke, I.; Thiim, M.; Mardia, K. V.; Ferkinghoff-Borg, J.; and Hamelryck, T. 2009. A probabilistic model of RNA conformational space. *PLoS Computational Biology* 5(6):e1000406.

Gao, S.; Hartman, John L, I.; Carter, J. L.; Hessner, M. J.; and Wang, X. 2010. Global analysis of phase locking in gene expression during cell cycle: the potential in network modeling. *BMC Systems Biology* 4(1):167.

Gärtner, T.; Flach, P.; and Wrobel, S. 2003. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*. Springer Berlin Heidelberg. 129–143.

Gatto, R., and Jammalamadaka, S. R. 2007. The generalized von mises distribution. *Statistical Methodology* 4(3):341–353.

Gatto, R. 2008. Some computational aspects of the generalized von mises distribution. *Statistics and Computing* 18(3):321–331.

Harder, T.; Boomsma, W.; Paluszewski, M.; Frellsen, J.; Johansson, K. E.; and Hamelryck, T. 2010. Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics* 11(1):306.

Jona-Lasinio, G.; Gelfand, A.; and Jona-Lasinio, M. 2012. Spatial analysis of wave direction data using wrapped gaussian processes. *The Annals of Applied Statistics* 6(4):1478–1498.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine Learning* 37(2):183–233.

Lawrence, N. D. 2004. Gaussian process latent variable models for visualisation of high dimensional data. In Thrun, S.; Saul, L.; and Schölkopf, B., eds., *Advances in Neural Information Processing Systems 16*. MIT Press. 329–336.

Lebanon, G. 2005. *Riemannian Geometry and Statistical Machine Learning*. Ph.D. Dissertation, Carnegie Mellon University, Pittsburg.

MacKay, D. J. C. 1994. Bayesian non-linear modelling for the prediction competition. In *ASHRAE Transactions, V.100, Pt.2*, 1053–1062. ASHRAE.

Mardia, K. V., and Jupp, P. E. 2000. *Directional statistics*. Chichester; New York: J. Wiley.

Mardia, K. V.; Hughes, G.; Taylor, C. C.; and Singh, H. 2008. A multivariate von mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* 36(1):99–109.

Quiñonero-Candela, J., and Rasmussen, C. E. 2005. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research* 6:1939–1959.

Rasmussen, C. E., and Williams, C. K. I. 2006. *Gaussian processes for machine learning*. MIT Press.

Santos, A.; Wernersson, R.; and Jensen, L. J. 2015. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research* 43(D1):D1140–D1144.

Tipping, M. E., and Bishop, C. M. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(3):611–622.

Turner, R., and Sahani, M. 2011a. Probabilistic amplitude and frequency demodulation. In *Advances in Neural Information Processing Systems 24*, 981–989. MIT Press.

Turner, R. E., and Sahani, M. 2011b. Two problems with variational expectation maximisation for time-series models. In Barber, D.; Cemgil, T.; and Chiappa, S., eds., *Bayesian Time series models*. Cambridge University Press. chapter 5, 109–130.

Wadhwa, N.; Rubinstein, M.; Durand, F.; and Freeman, W. T. 2013. Phase-based video motion processing. *ACM Transactions on Graphics* 32(4).

Wang, F., and Gelfand, A. E. 2013. Directional data analysis under the general projected normal distribution. *Statistical methodology* 10(1):113–127.

Wei, G. C. G., and Tanner, M. A. 1990. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85(411):699–704.