von Mises Quasi-Processes for Bayesian Circular Regression

Yarden Cohen [1] **Alexandre Khae Wu Navarro** [2] **Jes Frellsen** [3] **Richard E. Turner** [4] **Raziel Riemer** [1] **Ari Pakman** [1]

Abstract

The need for regression models to predict circular values arises in many scientific fields. In this work we explore a family of expressive and interpretable distributions over circle-valued random functions related to Gaussian processes targeting two Euclidean dimensions conditioned on the unit circle. The resulting probability model has connections with continuous spin models in statistical physics. Moreover, its density is very simple and has maximum-entropy, unlike previous Gaussian process-based approaches, which use wrapping or radial marginalization. For posterior inference, we introduce a new Stratonovich-like augmentation that lends itself to fast Markov Chain Monte Carlo sampling. We argue that transductive learning in these models favors a Bayesian approach to the parameters. We present experiments applying this model to the prediction of (i) wind directions and (ii) the percentage of the running gait cycle as a function of joint angles.

1. Introduction

Directional or circular data arises in many areas of science, including astronomy, biology, physics, earth science and meteorology. Since such data is supported on compact manifolds such as tori or (hyper)spheres, it is generally inappropriate to apply to them standard statistical methods, designed for observations with more familiar supports like $\mathbb{R}^d$. For general article-length reviews of probability models for directional data see (Jupp & Mardia, 1989; Lee, 2010; Pewsey & García-Portugués, 2021) and for book-length treatments see (Mardia & Jupp, 1999; Jammalamadaka & Sengupta, 2001; Pewsey et al., 2013; Ley & Verdebout, 2017; 2018).

[1]Ben-Gurion University of the Negev [2]Unilever [3]Technical University of Denmark [4]University of Cambridge. Correspondence to: Ari Pakman <pakman@bgu.ac.il>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

In this work we explore a model to perform non-parametric Bayesian regression over circular data. The model is introduced in Section 2, inference and learning are studied in Section 3 and Section 4, related works are reviewed in Section 5, followed by some examples in Section 6 and conclusions in Section 7.

2. The von Mises Quasi-Process

Consider a regression problem where a set of circular variables $\{\theta_i\}_{i=1}^n$ are observed at input locations $\{x_i\}_{i=1}^n$. Our goal is to predict new values $\{\phi_i\}_{i=1}^m$ at unseen locations $\{x_i^*\}_{i=1}^m$. A typical application is in geostatistics, where one wants to predict the direction of wind or ocean waves.

A prior for both the unobserved and observed variables, $\boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\theta}]$ can be obtained from a Gaussian Process (GP) over $d = m + n$ 2D Euclidean random functions $(f_{1,i}, f_{2,i}) \in \mathbb{R}^2$, where $f_{1,i}$ and $f_{2,j}$ are uncorrelated for all i, j and share the same covariance matrix $K_{ij} = k(x_i, x_j)$ with inverse $M_{ij} = (K^{-1})_{ij}$ with $i, j = 1 \dots d$. The density is

$$p(\mathbf{f}_1, \mathbf{f}_2|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\mathbf{f}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{M}(\mathbf{f}_1 - \boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{f}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{M}(\mathbf{f}_2 - \boldsymbol{\mu}_2)\right\}, \tag{1}$$

where $\boldsymbol{\mu}_i = \mu_i \mathbf{1}_d$ for $i = 1, 2$ with $\mathbf{1}_d$ a vector of 1's, i.e., a constant mean for each target coordinate. Using polar coordinates $(f_{1,i}, f_{2,i}) = (r_i \cos(\varphi_i), r_i \sin(\varphi_i))$ and conditioning on $r_i = 1$ yields

$$p(\boldsymbol{\varphi}|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{d} M_{ij}\cos(\varphi_i - \varphi_j) + \kappa \sum_{i=1}^{d} \cos(\varphi_i - \nu)\right\}, \quad \kappa > 0, \tag{2}$$

where κ and ν in (2) are simple functions of μ_1, μ_2 and $\boldsymbol{M}$, but we treat them in (2) as independent parameters. In Euclidean space GPs it is common to set $\mu_1 = \mu_2 = 0$, subtract from the data the empirical average and let the covariance kernel fully specify the process. But doing so would lead to $\kappa = 0$, which implies a uniform distribution

for the marginal of each variable φ_i in (2). Unfortunately there is no simple transformation of circular data that renders it uniformly distributed on the circle, and therefore in general we must keep $\kappa > 0$. We call the model (2) a *von Mises Quasi-Process (vMQP)* to stress the connection with Gaussian processes, although this is not a stochastic process, as explained in Section 4.

Let us split now the variables in (2) into $\boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\theta}]$, where $\boldsymbol{\phi} \in [0, 2\pi]^m$ are the unobserved variables and $\boldsymbol{\theta} \in [0, 2\pi]^n$ are observed. The posterior distribution of $\boldsymbol{\phi}$, obtained from (2) (up to a constant factor) by setting constant $\boldsymbol{\theta}$, is

$$p(\boldsymbol{\phi}|\boldsymbol{\theta}, \mathbf{x}) \propto \exp\left\{\boldsymbol{\rho}_c \cdot \cos(\boldsymbol{\phi}) + \boldsymbol{\rho}_s \cdot \sin(\boldsymbol{\phi}) - \frac{1}{2}\cos(\boldsymbol{\phi})^\top \boldsymbol{Q} \cos(\boldsymbol{\phi}) - \frac{1}{2}\sin(\boldsymbol{\phi})^\top \boldsymbol{Q} \sin(\boldsymbol{\phi})\right\} \tag{3}$$

where we defined

$$\boldsymbol{\rho}_c = -\boldsymbol{M}_{\boldsymbol{\phi}\boldsymbol{\theta}} \cos(\boldsymbol{\theta}) + \kappa \cos(\nu)\mathbf{1}_m \tag{4}$$

$$\boldsymbol{\rho}_s = -\boldsymbol{M}_{\boldsymbol{\phi}\boldsymbol{\theta}} \sin(\boldsymbol{\theta}) + \kappa \sin(\nu)\mathbf{1}_m \tag{5}$$

$$\boldsymbol{Q} = \boldsymbol{M}_{\boldsymbol{\phi}\boldsymbol{\phi}} \tag{6}$$

with $\mathbf{1}_m$ is an m dimensional vector of 1's, and

$$\boldsymbol{M} \equiv \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{\phi}\boldsymbol{\phi}} & \boldsymbol{M}_{\boldsymbol{\phi}\boldsymbol{\theta}} \\ \boldsymbol{M}_{\boldsymbol{\theta}\boldsymbol{\phi}} & \boldsymbol{M}_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{bmatrix} \in \mathbb{R}^{2d \times 2d}. \tag{7}$$

In the machine learning literature, the distribution (3) was studied previously in (Navarro et al., 2017), which showed that this is a maximum entropy distribution with fixed first circular moments and proposed its use for circular regression. But the practical exploitation of this model has been hindered by the lack of effective methods to (i) sample efficiently, and (ii) learn the parameters (κ, ν and those of the kernel). In this work, we address both of these challenges in Sections 3 and 4, respectively.

3. Sampling circular variables

Unlike Gaussian processes, closed-form expressions for the posterior mean and variance of the distribution (3) are not available, thus the need for an efficient sampling approach. Note that this is not a big limitation, since even in standard GPs, many quantities of interest, such as expectations of nonlinear functionals, require posterior samples for their estimation (Wilson et al., 2020; 2021).

Our method starts by noting that $\boldsymbol{Q} \in \mathbb{R}^{m \times m}$ in (3) is positive definite, since it is a submatrix of $\mathbf{M}$. Let $\lambda \in \mathbb{R}$ be a number bigger than all the eigenvalues of $\boldsymbol{Q}$. Define now a diagonal matrix $\boldsymbol{D} = \lambda \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the identity matrix, such that $\boldsymbol{D} - \boldsymbol{Q}$ is positive definite and can thus be Cholesky decomposed as

$$\mathbf{A}^\top \mathbf{A} = \boldsymbol{D} - \boldsymbol{Q}. \tag{8}$$

The key idea is to use $\mathbf{A}$ to augment (3) with a pair of Gaussian random variables $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^m$ with densities

$$p(\mathbf{z}_1|\boldsymbol{\phi}) = \mathcal{N}\left(\mathbf{z}_1; \mathbf{A}\cos(\boldsymbol{\phi}), \boldsymbol{I}_m\right) \tag{9}$$

$$p(\mathbf{z}_2|\boldsymbol{\phi}) = \mathcal{N}\left(\mathbf{z}_2; \mathbf{A}\sin(\boldsymbol{\phi}), \boldsymbol{I}_m\right). \tag{10}$$

Multiplying (3) by both densities in (9)-(10) we get the augmented distribution

$$p(\mathbf{z}, \boldsymbol{\phi}) = p(\boldsymbol{\phi})p(\mathbf{z}_1|\boldsymbol{\phi})p(\mathbf{z}_2|\boldsymbol{\phi}) \propto \exp\left\{(\boldsymbol{\rho}_c^\top + \mathbf{z}_1^\top \mathbf{A})\cos(\boldsymbol{\phi}) + (\boldsymbol{\rho}_s^\top + \mathbf{z}_2^\top \mathbf{A})\sin(\boldsymbol{\phi}) - \frac{1}{2}\mathbf{z}^\top \mathbf{z}\right\}, \tag{11}$$

where we denoted $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]$ and omitted the conditioning variables $(\boldsymbol{\theta}, \mathbf{x})$ to simplify the notation. Importantly, the terms quadratic in $[\cos(\boldsymbol{\phi}), \sin(\boldsymbol{\phi})]$ in the exponent of (3) have canceled in (11). This linearization of the trigonometric dependence in the exponent is similar to the Hubbard–Stratonovich transformation in field theory (Altland & Simons, 2010), and in machine learning has been similarly applied to interacting log-quadratic binary variables in (Martens & Sutskever, 2010; Zhang et al., 2012; Ostmeyer et al., 2021).

Since integrating out $\mathbf{z}$ in (11) gives back the original distribution (3), sampling from (11) and keeping only the $\boldsymbol{\phi}$ samples yields samples from (3). The advantage of the augmented form is twofold. First, we have found that Hamiltonian Monte Carlo (Neal et al., 2011) applied to (11) has better mixing properties than applied to (3). Secondly, the augmented distribution (11) lends itself to using a simple Gibbs sampler that alternates between

1. Sample $\mathbf{z}|\boldsymbol{\phi}$ from (9)-(10) by sampling $\varepsilon \sim \mathcal{N}\left(0, \boldsymbol{I}_{2m}\right)$ and setting

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}\cos(\boldsymbol{\phi}) \\ \mathbf{A}\sin(\boldsymbol{\phi}) \end{bmatrix} + \varepsilon \tag{12}$$

2. Sample $\boldsymbol{\phi}|\mathbf{z}$ from

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \prod_{i=1}^{m} \exp(a_i \cos(\phi_i - \gamma_i)), \tag{13}$$

where we defined

$$a_i = \sqrt{b_{c,i}^2 + b_{s,i}^2} \qquad \tan(\gamma_i) = \frac{b_{s,i}}{b_{c,i}}, \tag{14}$$

and $b_{c,i}$ and $b_{s,i}$ are the components of

$$\boldsymbol{b}_c = \boldsymbol{\rho}_c^\top + \mathbf{z}_1^\top \mathbf{A} \qquad \boldsymbol{b}_s = \boldsymbol{\rho}_s^\top + \mathbf{z}_2^\top \mathbf{A}. \tag{15}$$

The distribution (13) is a product of independent one-dimensional von Mises distributions, and can be sampled from efficiently using rejection-sampling (Best & Fisher, 1979), as implemented in standard packages.

4. Learning the parameters

Our learning objective is determined by the fact that, unlike standard Gaussian (Williams & Rasmussen, 2006) or t-processes (Shah et al., 2014), von Mises Quasi-Processes are not consistent under marginalization. Recall that we observe $\boldsymbol{\theta}$ at locations $\mathbf{x}$ and are interested in predictions $\boldsymbol{\phi}$ at $\mathbf{x}^*$. Let us express the normalized distribution (2) as

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}, \mathbf{x}^*, w) = \frac{e^{-U(\boldsymbol{\varphi}|w)}}{Z[w]} \qquad \boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\theta}] \tag{16}$$

where

$$\begin{aligned} U(\boldsymbol{\varphi}|w) =& \frac{1}{2}\sum_{i,j=1}^{d} M_{w,ij}\cos(\varphi_i - \varphi_j) \\ &- \kappa\sum_{i=1}^{d}\cos(\varphi_i - \nu), \end{aligned} \tag{17}$$

$$Z[w] = \int d^d\boldsymbol{\varphi} e^{-U(\boldsymbol{\varphi}|w)}, \tag{18}$$

and w indicates the parameters we want to learn. Since $\boldsymbol{\phi}$ are latent variables, maximum likelihood corresponds to

$$\hat{w} = \underset{w}{\operatorname{argmax}} \log \int d\boldsymbol{\phi}\, p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}, \mathbf{x}^*, w). \tag{19}$$

In standard Gaussian or t-processes, the above expression would reduce to

$$\hat{w} = \underset{w}{\operatorname{argmax}} \log p(\boldsymbol{\theta}|\mathbf{x}, w), \tag{20}$$

i.e., the test locations $\mathbf{x}^*$ disappear upon marginalization of $\boldsymbol{\phi}$. But this is not the case in our model,[1] and therefore, we must fix the test locations $\mathbf{x}^*$ at training time, a setting known as *transductive learning* (Vapnik, 2006).

The gradient of the learning objective (19) is

$$\begin{aligned} &\nabla_w \log \int d\boldsymbol{\phi}\, p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}, \mathbf{x}^*, w) = \\ &- \mathbb{E}_{p(\boldsymbol{\phi}|\boldsymbol{\theta}, \mathbf{x}, \mathbf{x}^*, w)}[U(\boldsymbol{\phi}, \boldsymbol{\theta})] + \mathbb{E}_{p(\boldsymbol{\phi}', \boldsymbol{\theta}'|\mathbf{x}, \mathbf{x}^*, w)}[U(\boldsymbol{\phi}', \boldsymbol{\theta}')], \end{aligned} \tag{21}$$

where we used the notation $U(\boldsymbol{\varphi}|w) = U(\boldsymbol{\phi}, \boldsymbol{\theta})$. This is the standard contrastive divergence gradient in the presence of latent variables, similar to that used to train Restricted Boltzmann Machines (RBMs) (Carreira-Perpinan & Hinton, 2005). But, unlike the latter case, we have found in experiments, using MCMC estimates for the expectations in (21), that this approach is utterly ineffective, arguably due to two differences. First, unlike RBMs, we only learn a small number of parameters (e.g. just three or four parameters in the examples in Section 6), and thus small amounts of noise or bias in the gradient estimation have potentially deleterious effects. Second, again unlike RBMs, the number of latent variables m in our case can be much higher than the number of observed variables n, and thus capturing the difference between the two terms in (21) might require impractically large numbers of Monte Carlo samples.

[1] Note that since models with different number of prediction locations are not related via marginalization, the Kolmogorov extension theorem (Durrett, 2019) does not apply and this prevents us from calling these models *processes*.

4.1. A fully Bayesian approach

To avoid the problems mentioned above, we resort instead to a fully Bayesian approach. We are interested in sampling from the joint posterior of parameters and unobserved variables,

$$p(\boldsymbol{\phi}, w|\boldsymbol{\theta}) \propto p(\boldsymbol{\phi}, \boldsymbol{\theta}|w)p(w) \tag{22}$$

where $p(w)$ is a prior distribution and we omit from now on the conditioning locations $(\mathbf{x}, \mathbf{x}^*)$. Using block Gibbs sampling, the conditional $p(\boldsymbol{\phi}|\boldsymbol{\theta}, w)$ is sampled with the method of Section 3. On the other hand, sampling from the conditional

$$p(w|\boldsymbol{\varphi}) \propto p(w)\frac{f(\boldsymbol{\varphi}|w)}{Z[w]}, \qquad \boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\theta}], \tag{23}$$

where we defined

$$f(\boldsymbol{\varphi}|w) = \exp\left[-U(\boldsymbol{\varphi}|w)\right], \tag{24}$$

is challenging, because we lack a closed-form expression for the normalization constant $Z[w]$. Several algorithms have been developed to tackle this problem (see Park & Haran (2018) for a review). In the following, we review the Exchange and Double Metropolis-Hastings algorithms (Murray et al., 2006; Liang, 2010).

The Exchange and Doubly Metropolis-Hastings algorithms. The Exchange algorithm (Murray et al., 2006), inspired by (Møller et al., 2006), starts by augmenting (23) with a freely-chosen proposal distribution $q(w'|w)$ over new parameters w' and a fictitious data point $\boldsymbol{\xi} \in [0, 2\pi]^d$ generated by w' on the full space (observed and unobserved),

$$p(w, w', \boldsymbol{\xi}|\boldsymbol{\varphi}) = p(w|\boldsymbol{\varphi})q(w'|w)\frac{f(\boldsymbol{\xi}|w')}{Z[w']}, \tag{25}$$

$$\propto p(w)q(w'|w)\frac{f(\boldsymbol{\varphi}|w)}{Z[w]}\frac{f(\boldsymbol{\xi}|w')}{Z[w']}. \tag{26}$$

One now samples from this joint distribution by alternating between two Monte Carlo moves:

1. Sample $w' \sim q(w'|w)$ and then $\boldsymbol{\xi} \sim f(\boldsymbol{\xi}|w')/Z[w']$.
2. Propose to exchange $w \leftrightarrow w'$ and accept with

Metropolis-Hastings (MH) probability

$$\max\left(1, \frac{p(w', w, \boldsymbol{\xi}|\boldsymbol{\varphi})}{p(w, w', \boldsymbol{\xi}|\boldsymbol{\varphi})}\right) = \max\left(1, \frac{p(w')q(w|w')f(\boldsymbol{\varphi}|w')f(\boldsymbol{\xi}|w)}{p(w)q(w'|w)f(\boldsymbol{\varphi}|w)f(\boldsymbol{\xi}|w')}\right), \quad (27)$$

where, remarkably, $Z[w]$ and $Z[w']$ cancel mutually.

If exact sampling of $\boldsymbol{\xi}$ in Step 1 is difficult, as in our model, it was suggested in (Liang, 2010) to approximate an exact sample by running long enough a Markov chain targeting $p(\boldsymbol{\xi}) \propto f(\boldsymbol{\xi}|w')$, an approach dubbed Double Metropolis-Hastings. Note that if we knew the normalization constant $Z[w]$ in (23), we could just sample a proposal from $q(w'|w)$ and accept it with MH probability

$$\max\left(1, \frac{p(w')q(w|w')f(\boldsymbol{\varphi}|w')Z[w]}{p(w)q(w'|w)f(\boldsymbol{\varphi}|w)Z[w']}\right). \quad (28)$$

Equation (27) in the Exchange algorithm corresponds instead to replacing the ratio of normalizing constants in (28) with a one-sample importance sampling approximation

$$\frac{Z[w]}{Z[w']} \simeq \frac{f(\boldsymbol{\xi}|w)}{f(\boldsymbol{\xi}|w')} \qquad \boldsymbol{\xi} \sim f(\boldsymbol{\xi}|w')/Z[w']. \quad (29)$$

5. Related works

Relation to statistical physics

Distributions of the type (3) are known in statistical physics as XY- or $O(2)$-models, with the angles ϕ_i being a continuous generalization of the $\{\pm 1\}$ spins of Ising models. The case when the ϕ_i's are located in a d-dimensional regular lattice and $\boldsymbol{Q}$ has a sparse structure, with non-zero entries only between nearest-neighbours, has been intensely researched since the 1960s (Friedli & Velenik, 2017), in particular for $d = 2$. Although physicists have developed several specialized algorithms to sample from XY-models, their efficiency depends on sign or sparsity properties of the $\boldsymbol{Q}$ matrix absent in our Bayesian regression setting, characterized by unsigned, dense $\boldsymbol{Q}$ matrices. For example, cluster flipping algorithms perform well in the 2D lattice ferromagnetic regime (Wolff, 1989) (when non-zero entries of $\boldsymbol{Q}$ are negative), but fail for spin-glasses (Kessler & Bretz, 1990) (when non-zero entries of $\boldsymbol{Q}$ have both signs), the relevant case for us. Worm algorithms (Prokof'ev & Svistunov, 2001; Wang, 2005) rely on the lattice nearest-neighbour topology, absent in our case with dense $\boldsymbol{Q}$s. Finally, a piecewise-deterministic Monte Carlo sampler for the XY-model (Michel et al., 2015; Michel, 2016), with exactly-solvable event times, is inefficient for non-sparse $\boldsymbol{Q}$ since the times between consecutive events tend to zero as the number of non-zero elements in $\boldsymbol{Q}$ grows.

Other circular models from Gaussian processes.

Gaussian process-like models for this task have been obtained in the past by starting from a GP in an Euclidean target space and applying a transformation to yield a distribution in circular space. The most popular approaches are wrapping and projecting (Jona-Lasinio et al., 2018). The wrapping approach consists of imposing an equivalence structure on the target space, namely $y_i \simeq y_i + 2\pi k$ for $k \in \mathbb{Z}$ for all i. The projecting approach considers a $2d$-dimensional space similar to (1), and after expressing the target in polar coordinates, $(f_{1,i}, f_{2,i}) = (r_i \cos(\varphi_i), r_i \sin(\varphi_i))$ it marginalizes the radial components r_i. Both approaches have been extensively explored in the literature. In both cases, the resulting probability densities lack a simple form such as (2), but the normalization constant is known since it is inherited from the original GP.

Wrapped distributions necessarily require approximations due to infinite sums implicit in their definition and are often truncated at the third harmonic (Mardia & Jupp, 1999). More recently, adaptive truncation schemes (Jona-Lasinio et al., 2012) and modelling the truncation point as a latent variable (Jona-Lasinio et al., 2014) were suggested to alleviate the approximation errors caused by truncation. Fully Bayesian approaches, using variations of Gibbs samplers for learning and inference have been developed both for the wrapped approach in (Ferrari, 2009; Jona-Lasinio et al., 2018; Jona Lasinio et al., 2020; Marques et al., 2022), and for the projecting approach in (Nuñez-Antonio & Gutiérrez-Peña, 2005; Nuñez-Antonio et al., 2011; Nuñez-Antonio & Gutiérrez-Peña, 2014; Hernandez-Stumpfhauser et al., 2017; Jona-Lasinio et al., 2018; Jona Lasinio et al., 2020; Zito & Kowal, 2023). On the other hand, the conditioning approach of the present paper, despite the attractive simplicity of its density, seems to have remained unexplored since its proposal in (Navarro et al., 2017), where a variational approach to inference and learning was suggested.

6. Experiments

6.1. Wind directions in Germany

In this experiment, we considered the problem of predicting wind directions at selected locations based on spatial proximity. We used data publicly available on the website of the German weather service (Deutcher Wetterdienst (DWD)), which consists of measurements collected at 260 weather stations every 10 minutes. We considered the same stormy weather period studied in (Marques et al., 2022), from September 23 to October 2, 2019. As in (Marques et al., 2022), the training and test data were circularly averaged over the entire time period. We only considered the spatial location as a the covariate, and used an exponential

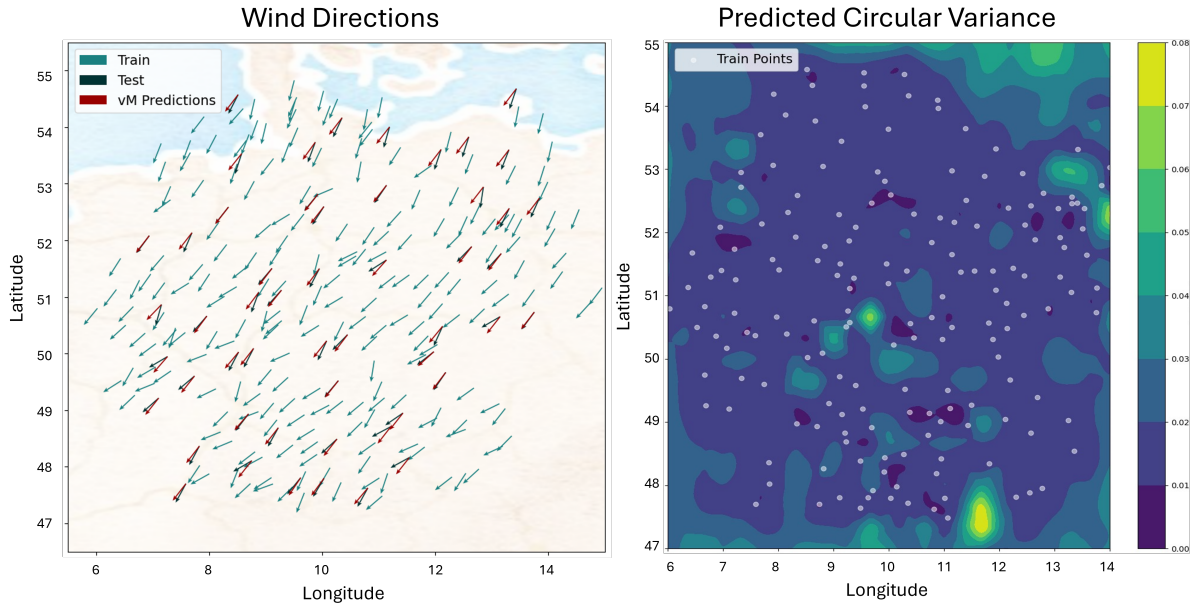

Figure 1: **Left:** Circular mean of the wind directions in 260 weather stations in Germany, randomly split between 208 train and 52 test locations. The predictions from the von Mises model are indicated on the test locations. **Right:** Predicted circular variance over a uniform grid of 25×25 points. Both figures best seen in color.

Table 1: Prediction of wind directions on test locations

METHOD	CRPS↓
WRAPPED GP	0.1363
PROJECTED GP	0.0205
VMQP	0.0254

covariance kernel of the form

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2l^2} - \right), \qquad (30)$$

with $i, j = 1 \dots 260$, with two parameters (σ^2, l^2). Here $\mathbf{x}_i$ is the 2D location of each weather station (longitude and latitude). We note that our model is admittedly too simple, and other variables variables such as humidity, temperature, altitude, etc. should be used in more realistic models.

We randomly split the data between 208/52 train/test locations. Figure 1 (Left) shows the circular mean of the training, test and predicted wind directions, while Figure 1 (Right) shows the predicted circular variance. We compared the predictions from our model with both wrapped and projected GPs, using the implementations of the `CircSpaceTime` R package (Jona Lasinio et al., 2020). All the predictions were evaluated against the test data using the circular continuous ranked probability score (CRPS) (Grimit et al., 2006). The results, presented in Table 1, show that the von Mises Quasi-Process yields a score very similar to the projected GP, and both are better than the wrapped GP.

6.2. Percentage of running gait cycle from joints angles

When a human runs, the positions of lower-limb joints go through a recurrent trajectory known as the *gait cycle*, illustrated in Figure 2 (Left). The joints' positions along the cycle are described by three joint angles, which are critical for understanding the neuromechanics and energetics of human locomotion (Winter, 1983). From an applied science perspective, joint angles and knowledge of phases in the running cycle are used to guide control software in both exoskeletons (Gad et al., 2022) and controllers for commercial devices, such as ATLAS (Sanz-Merodio et al., 2014), ReWalk (Esquenazi et al., 2012), as well as prostheses (Markowitz et al., 2011; Sup et al., 2008).

In this experiment, we consider the task of predicting the phase in the running gait cycle, measured as a percentage $t \in [0, 100]$, as a function of the joint angles. Since t is a circular variable, the von Mises Quasi-Process provides an appropriate prediction model. We used data from (Shkedy Rabani et al., 2022), who collected data from 16 healthy adults, while running on treadmill at several surface gradients.[2] See (Shkedy Rabani et al., 2022) for more details about the experimental protocol. The data consists of values of the

[2]The surface gradient is $100 \times$ tangent of inclination angle, and is indicated with a % symbol.

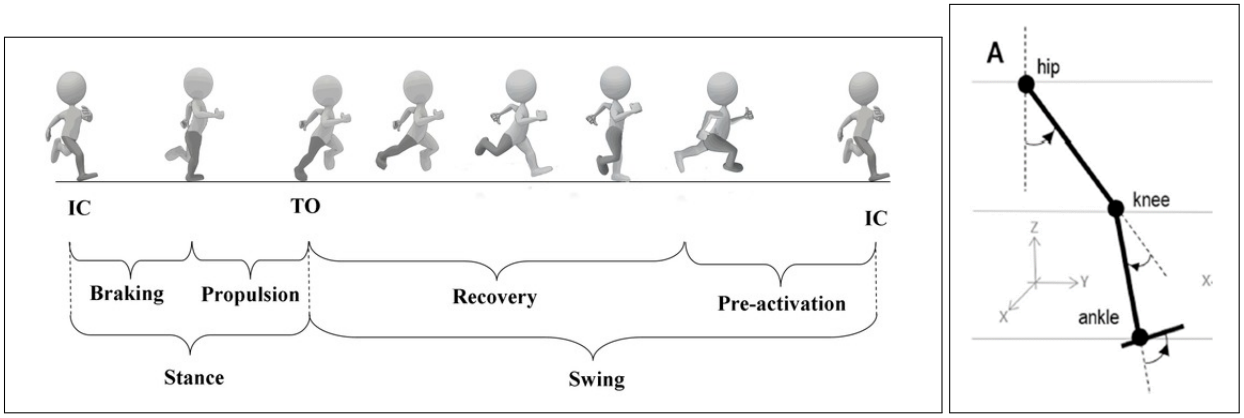

Figure 2: **Left:** Phases of the running gait cycle, defined w.r.t. the right leg of the figure. The cycle begins and ends with the initial contact (IC) of the right leg. The toe-off point is indicated as TO. Image from (Howard, 2016). **Right:** Definition of the three joint angles used to predict the location in the gait cycle. The angles are defined over the sagittal plane.

three angles defined in Figure 2 (Right), estimated at 100 values of the percentage $t \in \{1, 2 \dots 100\}$ from individuals running at five surface gradients $s \in \{0\%, \pm 5\%, \pm 10\%\}$. We used four gradients $s \in \{\pm 5\%, \pm 10\%\}$ for training, yielding 400 training points. Testing was performed on triplets of angles measured at surface gradient $s = 0\%$ at 20 points uniformly selected along the gait cycle. Training data for $s = \pm 10\%$ are shown in Figure 3. The vertical line at 40% of the cycle indicates approximately the toe-off point, see Figure 2 (Left).

In this experiment we assumed $\kappa = 0$ (see eq.(2)), because the training set contains points located uniformly along the full cycle of t, and thus there is no concentrated direction. We used an anisotropic exponential kernel of the form

$$K((\boldsymbol{a}_i, s_i), (\boldsymbol{a}_j, s_j)) \qquad i, j = 1 \dots 420, \tag{31}$$
$$= \sigma^2 \exp\left(-\frac{||\boldsymbol{a}_i - \boldsymbol{a}_j||^2}{2l^2} - \frac{(s_i - s_j)^2}{2g^2}\right),$$

Here $\boldsymbol{a}_i$ is a three-dimensional vector containing the three joint angles and $s_i \in \{0, \pm 5, \pm 10\}$ indicates the surface gradient.

At test time, we are given 20 measurements of the three joint angles, with each measurement performed at a different point in the gait cycle of humans running at zero surface gradients. Figure 4 shows the results, showing a high accuracy in the prediction of the gait cycle percentage. Note that the predictions are better after the toe-off point. The error bars are proportional to the circular variance of each point.

7. Conclusions and outlook

In this work we explored a general model for Bayesian non-parametric regression of circular variables, for which we explored efficient MCMC techniques for learning and

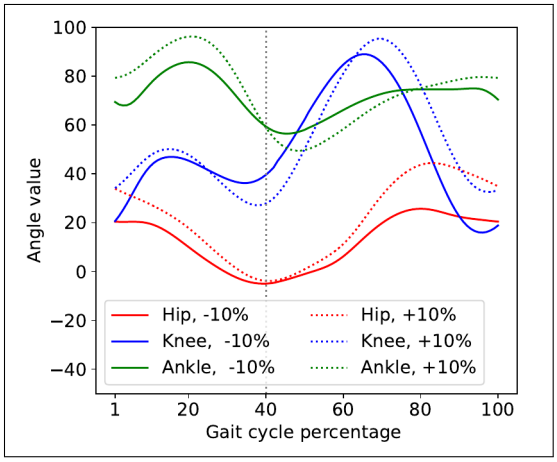

Figure 3: Data examples for the gait percentage experiment.

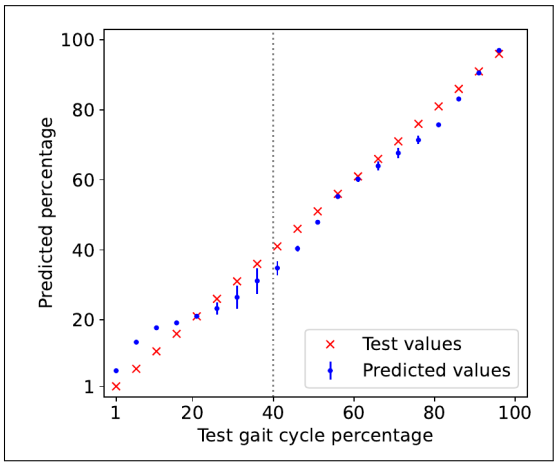

Figure 4: Predictions vs. test values of cycle percentage, for 20 measurements of the joint angles at zero gradient.

inference. Note that both the theoretical and experimental results presented assume noiseless data. Relevant future work includes incorporating noise models, exploring the augmentation from Section 3 in statistical physics XY-models for which current approaches are not efficient (Michel, 2016) (e.g. glassy regimes or 3D lattices) and exploring parameter learning using score matching (Vértes & Sahani, 2016).

References

Altland, A. and Simons, B. D. *Condensed matter field theory*. Cambridge University Press, 2010.

Best, D. and Fisher, N. I. Efficient simulation of the von Mises distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):152–157, 1979.

Carreira-Perpinan, M. A. and Hinton, G. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pp. 33–40. PMLR, 2005.

Deutcher Wetterdienst (DWD). https://dwd.de.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Esquenazi, A., Talaty, M., Packel, A., and Saulino, M. The rewalk powered exoskeleton to restore ambulatory function to individuals with thoracic-level motor-complete spinal cord injury. *American journal of physical medicine & rehabilitation*, 91(11):911–921, 2012.

Ferrari, C. *The Wrapping Approach for Circular Data Bayesian Modelling*. PhD thesis, Università di Bologna, Bologna, 2009.

Friedli, S. and Velenik, Y. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, 2017. ISBN 978-1-107-18482-4. doi: 10.1017/9781316882603.

Gad, M., Lev-Ari, B., Shapiro, A., Ben-David, C., and Riemer, R. Biomechanical knee energy harvester: Design optimization and testing. *Frontiers in Robotics and AI*, 9: 998248, 2022.

Grimit, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2925–2942, 2006.

Hernandez-Stumpfhauser, D., Breidt, F. J., and van der Woerd, M. J. The General Projected Normal Distribution of Arbitrary Dimension: Modeling and Bayesian Inference. *Bayesian Analysis*, 12(1):113–133, 03 2017. doi: 10.1214/15-BA989. URL http://dx.doi.org/10.1214/15-BA989.

Howard, R. M. *The application of data analysis methods for surface electromyography in shot putting and sprinting*. PhD thesis, University of Limerick, 2016.

Jammalamadaka, S. R. and Sengupta, A. *Topics in circular statistics*, volume 5. World Scientific, 2001.

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

Jona-Lasinio, G., Mastrantonio, G., and Gelfand, A. E. Models for space-time directional data using Wrapped Gaussian processes. In Cabras, S., Battista, T. D., and Racugno, W. (eds.), *Proceedings of the 47th Scientific Meeting of the Italian Statistical Society*, pp. 1–10, Italy, 2014. ISBN 9788884678744. URL http://www.sis2014.it/proceedings/.

Jona-Lasinio, G., Gelfand, A. E., and Mastrantonio, G. Spatial and spatio-temporal circular processes with application to wave directions. *Applied Directional Statistics: Modern Methods and Case Studies*, pp. 35, 2018.

Jona Lasinio, G., Santoro, M., and Mastrantonio, G. CircSpaceTime: an R package for spatial and spatio-temporal modelling of circular data. *Journal of Statistical Computation and Simulation*, 90(7):1315–1345, 2020. doi: 10.1080/00949655.2020.1725008. URL https://doi.org/10.1080/00949655.2020.1725008.

Jupp, P. and Mardia, K. A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review/Revue Internationale de Statistique*, pp. 261–294, 1989.

Kessler, D. A. and Bretz, M. Unbridled growth of spin-glass clusters. *Physical Review B*, 41(7):4778, 1990.

Lee, A. Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):477–486, 2010.

Ley, C. and Verdebout, T. *Modern directional statistics*. CRC Press, 2017.

Ley, C. and Verdebout, T. *Applied directional statistics: modern methods and case studies*. CRC Press, 2018.

Liang, F. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

Mardia, K. V. and Jupp, P. E. *Directional statistics*. John Wiley & Sons, 1999.

Markowitz, J., Krishnaswamy, P., Eilenberg, M. F., Endo, K., Barnhart, C., and Herr, H. Speed adaptation in a powered transtibial prosthesis controlled with a neuromuscular model. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1570):1621–1631, 2011.

Marques, I., Kneib, T., and Klein, N. A non-stationary model for spatially dependent circular response data

based on wrapped gaussian processes. *Statistics and Computing*, 32(5):73, 9 2022.

Martens, J. and Sutskever, I. Parallelizable sampling of Markov random fields. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 517–524. JMLR Workshop and Conference Proceedings, 2010.

Michel, M. *Irreversible Markov chains by the factorized Metropolis filter: Algorithms and applications in particle systems and spin models.* PhD thesis, École Normale Supérieure, Paris, France, 2016.

Michel, M., Mayer, J., and Krauth, W. Event-chain Monte Carlo for classical continuous spin models. *Europhysics Letters*, 112(2):20003, 2015.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

Murray, I., Ghahramani, Z., and MacKay, D. J. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 359–366, 2006.

Navarro, A., Frellsen, J., and Turner, R. The multivariate generalised von Mises distribution: inference and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Neal, R. M. et al. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Nuñez-Antonio, G. and Gutiérrez-Peña, E. A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, 32(10):995–1001, 2005.

Nuñez-Antonio, G. and Gutiérrez-Peña, E. A Bayesian model for longitudinal circular data based on the projected normal distribution. *Computational Statistics & Data Analysis*, 71:506–519, 2014.

Nuñez-Antonio, G., Gutiérrez-Peña, E., and Escarela, G. A Bayesian regression model for circular data based on the projected normal distribution. *Statistical Modelling*, 11 (3):185–201, 2011.

Ostmeyer, J., Berkowitz, E., Luu, T., Petschlies, M., and Pittler, F. The Ising model with Hybrid Monte Carlo. *Computer Physics Communications*, 265:107978, 2021.

Park, J. and Haran, M. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390, 2018. doi: 10.1080/01621459.2018.1448824. URL https://doi.org/10.1080/01621459.2018.1448824.

Pewsey, A. and García-Portugués, E. Recent advances in directional statistics. *Test*, 30(1):1–58, 2021.

Pewsey, A., Neuhäuser, M., and Ruxton, G. D. *Circular statistics in R*. OUP Oxford, 2013.

Prokof'ev, N. and Svistunov, B. Worm algorithms for classical statistical models. *Physical review letters*, 87(16): 160601, 2001.

Sanz-Merodio, D., Cestari, M., Arevalo, J. C., Carrillo, X. A., and Garcia, E. Generation and control of adaptive gaits in lower-limb exoskeletons for motion assistance. *Advanced Robotics*, 28(5):329–338, 2014.

Shah, A., Wilson, A., and Ghahramani, Z. Student-t processes as alternatives to Gaussian processes. In *Artificial intelligence and statistics*, pp. 877–885. PMLR, 2014.

Shkedy Rabani, A., Mizrachi, S., Sawicki, G. S., and Riemer, R. Parametric equations to study and predict lower-limb joint kinematics and kinetics during human walking and slow running on slopes. *PloS one*, 17(8): e0269061, 2022.

Sup, F., Bohara, A., and Goldfarb, M. Design and control of a powered transfemoral prosthesis. *The International journal of robotics research*, 27(2):263–273, 2008.

Vapnik, V. Transductive inference and semi-supervised learning. In Chapelle, O., Schölkopf, B., and Zien, A. (eds.), *Semi-Supervised Learning*, chapter 24, pp. 453–472. MIT Press, 2006.

Vértes, E. and Sahani, M. Learning doubly intractable latent variable models via score matching. *Advances in Approximate Bayesian Inference NIPS 2016 Workshop*, 2016.

Wang, J.-S. Worm algorithm for two-dimensional spin glasses. *Physical Review E*, 72(3):036706, 2005.

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

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning*, pp. 10292–10302. PMLR, 2020.

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. P. Pathwise conditioning of Gaussian processes. *The Journal of Machine Learning Research*, 22(1):4741–4787, 2021.

Winter, D. A. Energy generation and absorption at the ankle and knee during fast, natural, and slow cadences. *Clinical Orthopaedics and Related Research (1976-2007)*, 175: 147–154, 1983.

Wolff, U. Collective Monte Carlo updating for spin systems. *Physical Review Letters*, 62(4):361, 1989.

Zhang, Y., Ghahramani, Z., Storkey, A. J., and Sutton, C. Continuous relaxations for discrete Hamiltonian Monte Carlo. *Advances in Neural Information Processing Systems*, 25, 2012.

Zito, J. and Kowal, D. The projected dynamic linear model for time series on the sphere. *arXiv preprint arXiv:2308.14996*, 2023.