Bayesian Circular Regression with von Mises Quasi-Processes

Yarden Cohen[1,2,3] **Alex Navarro**[4] **Jes Frellsen**[5] **Richard E. Turner**[6] **Raziel Riemer**[1] **Ari Pakman**[1,2,3]

[1]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev
[2]The School of Brain Sciences and Cognition, Ben-Gurion University of the Negev
[3]Data Science Research Center, Ben-Gurion University of the Negev
[4]Unilever   [5]Technical University of Denmark   [6]University of Cambridge

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 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 Gibbs sampling. We argue that transductive learning in these models favors a Bayesian approach to the parameters and apply our sampling scheme to the Double Metropolis-Hastings algorithm. 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 reviews of probability models for directional data see Jupp and Mardia (1989); Lee (2010); Pewsey and García-Portugués (2021) and for book-length treatments see Jammalamadaka and Sengupta (2001); Ley and Verdebout (2017, 2018); Mardia and Jupp (1999); Pewsey et al. (2013).

In this work we explore a model introduced in Navarro et al. (2017) to perform non-parametric Bayesian regression over circular data. The model can be obtained by starting from a Gaussian process for two-dimensional Euclidean observations and conditioning on the unit-circle. Despite its attractive features (maximum entropy and a simple distribution), the model remained practically unexploited by the lack of effective methods for posterior inference and parameter learning. Here we address both challenges by developing novel Markov Chain Monte Carlo (MCMC) techniques tailored to this setting, with emphasis on the need for transductive learning of the model parameters (see Figure 1).

The model is reviewed in Section 2, inference is studied in Section 3, parameter learning in Section 4, related works are reviewed in Section 5, followed by experiments in Section 6 and conclusions in Section 7.[1]

2 VON-MISES QUASI-PROCESSES

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. We denote the combined unobserved ϕ and observed θ variables as $\varphi = [\phi, \theta]$, define $d = m + n$ and consider a prior

---

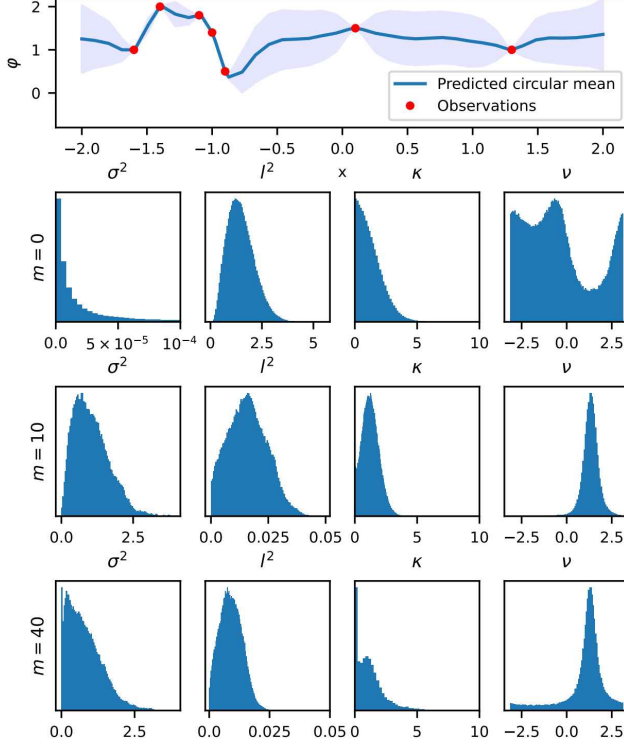[1]Our code is available at https://github.com/Yarden231/vMQP.

Figure 1: **Transductive learning in action.** Given seven angular observations on the x axis, we show histograms of posterior samples of the parameters of a vMQP model (1) with kernel $K(x_i, x_j) = \sigma^2 \exp(-(x_i - x_j)^2/2l^2)$, for different numbers m of uniformly located predictive locations. Transductive learning manifest itself in the changes of these distributions as a function of the predictive locations. Note the shrinking of l^2 as m grows and the multimodality of ν captured by the Bayesian approach. The confidence interval in the top panel is proportional to the circular variance for $m = 40$.

density for $\boldsymbol{\varphi}$ given by

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

Here $\nu \in [-\pi, +\pi]$, and we restrict to $\kappa > 0$ since the density is invariant under $(\kappa, \nu) \leftrightarrow (-\kappa, \nu - \pi)$. The matrix in (1) is the inverse $M_{ij} = (K^{-1})_{ij}$ of $K_{ij} = k(x_i, x_j)$, where $k(\cdot, \cdot)$ is a function that plays a role similar to the covariance in Gaussian processes in Euclidean target space. We call the model (1) 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.

The posterior distribution (up to a constant factor) of the unobserved $\boldsymbol{\phi}$, obtained from (1) by setting constant $\boldsymbol{\theta}$, is

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

where we defined

$$\boldsymbol{\rho}_c = -\boldsymbol{M}_{\phi\theta}\cos(\boldsymbol{\theta}) + \kappa\cos(\nu)\mathbf{1}_m, \tag{3}$$
$$\boldsymbol{\rho}_s = -\boldsymbol{M}_{\phi\theta}\sin(\boldsymbol{\theta}) + \kappa\sin(\nu)\mathbf{1}_m, \tag{4}$$
$$\boldsymbol{Q} = \boldsymbol{M}_{\phi\phi}, \tag{5}$$

with $\mathbf{1}_m \in \mathbb{R}^m$ a vector of 1's, and the matrix

$$\boldsymbol{M} \equiv \begin{bmatrix} \boldsymbol{M}_{\phi\phi} & \boldsymbol{M}_{\phi\theta} \\ \boldsymbol{M}_{\theta\phi} & \boldsymbol{M}_{\theta\theta} \end{bmatrix} \in \mathbb{R}^{d\times d} \tag{6}$$

has entries M_{ij}. In the machine learning literature, similar distributions were studied in Navarro et al. (2017), which showed that these are maximum entropy distribution with fixed first circular moments and proposed their use for circular regression. But the practical exploitation of these models has been hindered by the lack of effective methods to (i) sample efficiently, and (ii) learn the parameters $w \equiv (\kappa, \nu, \text{Kernel parameters})$. We address both of these challenges in Sections 3 and 4, respectively.

2.1 Relation to Gaussian Processes

The prior (1) can be obtained from a Gaussian Process (GP) over d 2D Euclidean random function values $(f_{1,i}, f_{2,i}) \in \mathbb{R}^2$, where $f_{1,i}$ and $f_{2,j}$ are uncorrelated for all i, j and share the same covariance matrix $K_{ij} = k(x_i, x_j)$. The density is

$$p(\mathbf{f}_1, \mathbf{f}_2|\mathbf{x}) \propto \exp\left\{ -\frac{1}{2}(\mathbf{f}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{M}(\mathbf{f}_1 - \boldsymbol{\mu}_1) \right.$$
$$\left. -\frac{1}{2}(\mathbf{f}_2 - \boldsymbol{\mu}_2)^\top \boldsymbol{M}(\mathbf{f}_2 - \boldsymbol{\mu}_2) \right\}. \tag{7}$$

Expressing $(f_{1,i}, f_{2,i}) = (r_i\cos(\varphi_i), r_i\sin(\varphi_i))$, conditioning on $r_i = 1$ and setting $\boldsymbol{\mu}_i = \mu_i\mathbf{1}_d$ for $i = 1, 2$ with $\mathbf{1}_d$ a vector of 1's, yields (1), where κ and ν are functions of μ_1, μ_2 and $\boldsymbol{M}$. In Euclidean space GPs it is common to set $\mu_1 = \mu_2 = 0$ and subtract from the data the empirical average. In our case, we keep $\kappa > 0$ in general, since $\kappa = 0$ implies a uniform distribution for the marginal of each variable φ_i, an assumption that needs to be evaluated in each case.

2.2 Including noisy observations

For noisy observations, we start with the prior (1), and instead of identifying the n last components of $\boldsymbol{\varphi}$

with the observations θ_i, we keep them as separate variables and assume that each observation θ_i has a 1D von Mises likelihood centered on φ_{m+i}, leading to a posterior

$$p(\boldsymbol{\varphi}|\boldsymbol{\theta}, \mathbf{x}) \propto \exp\{\chi \sum_{i=1}^{n} \cos(\theta_i - \varphi_{m+i})\} p(\boldsymbol{\varphi}|\mathbf{x}) \quad (8)$$

where χ is a parameter. The equations in the next Sections are presented for the noiseless case, but they are easily extended to the case (8).

3 SAMPLING CIRCULAR VARIABLES

Unlike Gaussian processes, closed-form expressions for the posterior mean and variance of the distribution (2) 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 (2) 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 $\boldsymbol{Q}$. We perform now a Cholesky decomposition,

$$\mathbf{A}^\top \mathbf{A} = \lambda \boldsymbol{I}_m - \boldsymbol{Q}, \quad (9)$$

where the r.h.s is positive definite. The key idea is to augment (2) with a pair of Gaussian random variables $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^m$ with densities

$$p(\mathbf{z}_1|\boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}_1; \mathbf{A}\cos(\boldsymbol{\phi}), \boldsymbol{I}_m) \quad (10)$$
$$p(\mathbf{z}_2|\boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}_2; \mathbf{A}\sin(\boldsymbol{\phi}), \boldsymbol{I}_m). \quad (11)$$

Multiplying (2) by both densities in (10)-(11) we get the augmented distribution

$$\begin{aligned} p(\mathbf{z}, \boldsymbol{\phi}) &= p(\boldsymbol{\phi}) p(\mathbf{z}_1|\boldsymbol{\phi}) p(\mathbf{z}_2|\boldsymbol{\phi}) \\ &\propto \exp\left\{(\boldsymbol{\rho}_c^\top + \mathbf{z}_1^\top \mathbf{A})\cos(\boldsymbol{\phi}) \right. \\ &\left. + (\boldsymbol{\rho}_s^\top + \mathbf{z}_2^\top \mathbf{A})\sin(\boldsymbol{\phi}) - \frac{1}{2}\mathbf{z}^\top \mathbf{z}\right\}, \end{aligned} \quad (12)$$

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 (2) have canceled in (12). This linearization of the trigonometric dependence in the exponent is similar to the Hubbard–Stratonovich transformation in field theory (Altland and Simons, 2010), and in machine learning it has been similarly applied to interacting log-quadratic binary variables in Martens and Sutskever (2010); Ostmeyer et al. (2021); Zhang et al. (2012).

Note that sampling from (12) and keeping only the $\boldsymbol{\phi}$ samples yields samples from (2). The advantage of the augmented form is that it lends itself to using a simple Gibbs sampler that alternates between

1. Sample $\mathbf{z}|\boldsymbol{\phi}$ from (10)-(11) by sampling $[\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2] \sim \mathcal{N}(0, \boldsymbol{I}_{2m})$ and setting

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}\cos(\boldsymbol{\phi}) \\ \mathbf{A}\sin(\boldsymbol{\phi}) \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \quad (13)$$

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

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \prod_{i=1}^{m} \exp(a_i \cos(\phi_i - \gamma_i)), \quad (14)$$

where we defined

$$a_i = \sqrt{b_{c,i}^2 + b_{s,i}^2} \qquad \tan(\gamma_i) = \frac{b_{s,i}}{b_{c,i}}, \quad (15)$$

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

$$\boldsymbol{b}_c = \boldsymbol{\rho}_c + \mathbf{A}^\top \mathbf{z}_1, \quad (16)$$
$$\boldsymbol{b}_s = \boldsymbol{\rho}_s + \mathbf{A}^\top \mathbf{z}_2. \quad (17)$$

The distribution (14) is a product of independent one-dimensional von Mises distributions, and can be sampled efficiently using rejection-sampling (Best and Fisher, 1979), as implemented in standard packages, or using exact Hamiltonian Monte Carlo (Pakman, 2025).

The role of λ. Combining (13) with (17) we get

$$\boldsymbol{b}_c = \boldsymbol{\rho}_c + \cos(\boldsymbol{\phi})[\lambda \boldsymbol{I}_m - \boldsymbol{Q}] + \mathbf{A}^\top \boldsymbol{\varepsilon}_1 \quad (18)$$

and similarly for $\boldsymbol{b}_s$. In the $\lambda \to \infty$ limit, in iteration t, we get $\boldsymbol{b}_c \simeq \lambda \cos(\boldsymbol{\phi}^{(t)})$ and the means of $p(\boldsymbol{\phi}^{(t+1)}|\mathbf{z})$ in (14) become $\gamma_i \to \phi_i^{(t)}$, hurting the exploration of the circular space. This non-rigurous heuristic argument suggest λ should not be bigger than necessary. We verify empirically this claim in Section 6.1.

4 LEARNING THE PARAMETERS

Our learning objective is determined by the fact that, unlike standard Gaussian (Williams and 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 (1) as

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}, \mathbf{x}^*, w) = \frac{e^{-U(\boldsymbol{\varphi}|w)}}{Z[w]} \qquad \boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\theta}] \quad (19)$$

where

$$U(\boldsymbol{\varphi}|w) = \frac{1}{2}\sum_{i,j=1}^{d} M_{w,ij}\cos(\varphi_i - \varphi_j) \qquad (20)$$

$$- \kappa\sum_{i=1}^{d}\cos(\varphi_i - \nu),$$

$$Z[w] = \int d\boldsymbol{\varphi}\, e^{-U(\boldsymbol{\varphi}|w)}, \qquad (21)$$

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

$$\hat{w} = \underset{w}{\operatorname{argmax}}\,\log\int d\boldsymbol{\phi}\, p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}, \mathbf{x}^*, w). \qquad (22)$$

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

$$\hat{w} = \underset{w}{\operatorname{argmax}}\,\log p(\boldsymbol{\theta}|\mathbf{x}, w), \qquad (23)$$

i.e., the test locations $\mathbf{x}^*$ disappear upon marginalization of $\boldsymbol{\phi}$. But this is not the case in our model,[2] and therefore, we must fix the test locations $\mathbf{x}^*$ at training time, a setting known as *transductive learning* (Vapnik, 2006) and illustrated in Figure 1.

4.1 Problems with point estimates

The gradient of the learning objective (22) is

$$\nabla_w \log\int d\boldsymbol{\phi}\, p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{x}, \mathbf{x}^*, w) = \qquad (24)$$
$$\mathbb{E}_{p_2(\boldsymbol{\phi}', \boldsymbol{\theta}')}[\nabla_w U(\boldsymbol{\phi}', \boldsymbol{\theta}'|w)] - \mathbb{E}_{p_1(\boldsymbol{\phi}')}[\nabla_w U(\boldsymbol{\phi}', \boldsymbol{\theta}|w)]$$

where we defined

$$p_2(\boldsymbol{\phi}', \boldsymbol{\theta}') = p(\boldsymbol{\phi}', \boldsymbol{\theta}'|\mathbf{x}, \mathbf{x}^*, w), \qquad (25)$$
$$p_1(\boldsymbol{\phi}') = p(\boldsymbol{\phi}'|\boldsymbol{\theta}, \mathbf{x}, \mathbf{x}^*, w). \qquad (26)$$

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 and Hinton, 2005). But, unlike the latter case, we have found in experiments, using MCMC estimates for the expectations in (24), 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.

[2] 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*.
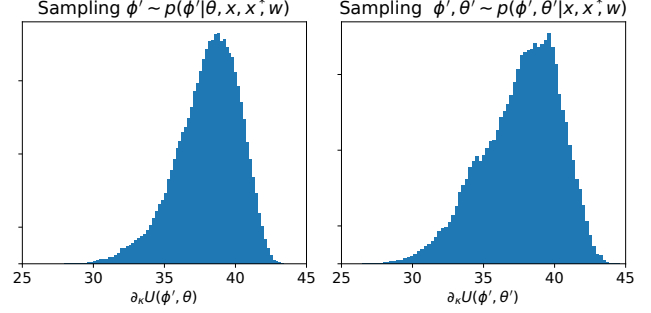

Figure 2: **Problems with maximum likelihood estimation.** Histograms of two different random evaluations of $\partial_\kappa U(\boldsymbol{\phi}, \boldsymbol{\theta})$, whose difference is required to estimate the log-likelihood gradient (24) w.r.t. κ. The samples are from the model in Figure 1 with $m = 40$. The similarity of the distributions leads to highly inaccurate gradient estimates.

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 (24) might require impractically large numbers of Monte Carlo samples. Figure 2 illustrates this problem.

4.2 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,

$$p(\boldsymbol{\phi}, w|\boldsymbol{\theta}) \propto p(\boldsymbol{\phi}, \boldsymbol{\theta}|w)p(w), \qquad (27)$$

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

$$p(w|\boldsymbol{\varphi}) \propto p(w)\frac{f(\boldsymbol{\varphi}|w)}{Z[w]}, \qquad \boldsymbol{\varphi} = [\boldsymbol{\phi}, \boldsymbol{\theta}], \qquad (28)$$

where we defined

$$f(\boldsymbol{\varphi}|w) = \exp\left[-U(\boldsymbol{\varphi}|w)\right], \qquad (29)$$

is challenging, because we lack a closed-form expression for the normalization constant $Z[w]$, defined in (21). Several algorithms have been developed to tackle this problem (see Park and Haran (2018) for a review). In the following, we review the Exchange, Double Metropolis-Hastings and Bridging algorithms.

The Exchange and Double MH algorithms. The Exchange algorithm (Murray et al., 2006), in-

spired by Møller et al. (2006), starts by augmenting (28) 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),

$$p(w, w', \boldsymbol{\xi}|\boldsymbol{\varphi}) = p(w|\boldsymbol{\varphi})q(w'|w)\frac{f(\boldsymbol{\xi}|w')}{Z[w']}\,, \qquad (30)$$

$$\propto p(w)q(w'|w)\frac{f(\boldsymbol{\varphi}|w)}{Z[w]}\frac{f(\boldsymbol{\xi}|w')}{Z[w']}\,. \qquad (31)$$

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

1. Sample

$$w' \sim q(w'|w)\,, \qquad (32)$$

$$\boldsymbol{\xi} \sim f(\boldsymbol{\xi}|w')/Z[w']\,. \qquad (33)$$

2. Propose to exchange $w \leftrightarrow w'$ and accept with Metropolis-Hastings (MH) probability

$$\min\left(1, \frac{p(w', w, \boldsymbol{\xi}|\boldsymbol{\varphi})}{p(w, w', \boldsymbol{\xi}|\boldsymbol{\varphi})}\right) =$$
$$\min\left(1, \frac{p(w')q(w|w')f(\boldsymbol{\varphi}|w')f(\boldsymbol{\xi}|w)}{p(w)q(w'|w)f(\boldsymbol{\varphi}|w)f(\boldsymbol{\xi}|w')}\right)\,, \qquad (34)$$

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

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. In our case, we run this Markov chain using an augmentation as in Section 3, based on a Cholesky decomposition

$$\mathbf{A}_{w'}^\top \mathbf{A}_{w'} = \boldsymbol{I}_d \lambda_{w'} - \boldsymbol{M}_{w'}, \qquad (35)$$

where $\lambda_{w'}$ is chosen to make the r.h.s. positive definite.

Efficient Bridging for the vMQP. Note that if we knew the normalization constant $Z[w]$ in (28), we could sample a proposal from $q(w'|w)$ and accept it with MH probability

$$\min\left(1, \frac{p(w')q(w|w')f(\boldsymbol{\varphi}|w')Z[w]}{p(w)q(w'|w)f(\boldsymbol{\varphi}|w)Z[w']}\right)\,. \qquad (36)$$

Equation (34) in the Exchange algorithm corresponds instead to replacing the ratio of normalizing constants in (36) with a one-sample importance sampling approximation

$$\frac{Z[w]}{Z[w']} \simeq \frac{f(\boldsymbol{\xi}|w)}{f(\boldsymbol{\xi}|w')}. \qquad (37)$$

where

$$\boldsymbol{\xi} \sim f(\boldsymbol{\xi}|w')/Z[w']\,. \qquad (38)$$

Now, since it is known that the acceptance rate (36) is maximal (Peskun, 1973; Tierney, 1998), one can expect to increase the acceptance rate (34) by using an estimate with lower variance than (37). The Bridging algorithm (Murray et al., 2006) achieves this by further augmenting (30)-(31) with K additional variables. In Appendix B we review the Bridging algorithm and present an efficient extension tailored to our augmented model (12).

5 RELATED WORKS

Relation to statistical physics. Distributions of the type (2) 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 and 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 and Bretz, 1990) (when non-zero entries of $\boldsymbol{Q}$ have both signs), the relevant case for us. Worm algorithms (Prokof'ev and 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, 2016; Michel et al., 2015), 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). Wrapping consists in 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 two-dimensional space similar to (7) 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 (1), but the normalization constant is known since it is inherited from the original GP.

Wrapped distributions require approximations due to infinite sums implicit in their definition and are often truncated at the third harmonic (Mardia and 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 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 Hernandez-Stumpfhauser et al. (2017); Jona-Lasinio et al. (2018); Jona Lasinio et al. (2020); Nuñez-Antonio and Gutiérrez-Peña (2005, 2014); Nuñez-Antonio et al. (2011); Zito and Kowal (2023).

Variational inference. Despite the attractive simplicity of the density (1), this model seems to have remained unexplored since the work by Navarro et al. (2017), which studied an approximation to the posterior (2) via variational inference, with a proposal given by a product of one-dimensional von Mises distributions. But such an approach is inefficient, because the resulting KL divergences lack closed form and require expensive numerical evaluations.

6 EXPERIMENTS

6.1 Selecting the λ parameter

As mentioned in Section 3, we expect better mixing of our augmented Gibbs sampler for smaller values of the parameter λ in (9). In Figure 3, we verify this claim for the the data presented in Figure 1 with $m = 10$.

6.2 Sampler comparisons

In Figure 4 we compare the efficiency of Gibbs sampling and Hamiltonian Monte Carlo (HMC) (Neal et al., 2011) applied to both the non-augmented (2) and augmented (12) distributions for the same data as in Figure 3. The Gibbs conditionals in (2) are 1D von Mises. In some cases (notably for large m), we found that it is possible to adjust the parameters of HMC (size and number of steps in each iteration) to be more efficient than the augmented Gibbs sampler,
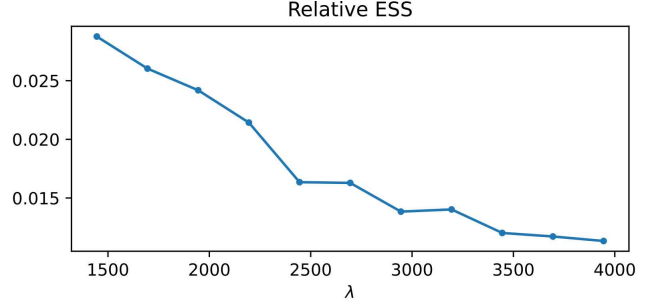

Figure 3: **Optimal λ.** Relative Effective Sample Size (ESS) (see definition in Appendix A.1) of the log of the density (2), computed with samples from the data presented in Figure 1 with $m = 10$, as a function of the parameter λ that enables the Cholesky decomposition (9). The plot confirms that smaller λ should be preferred.
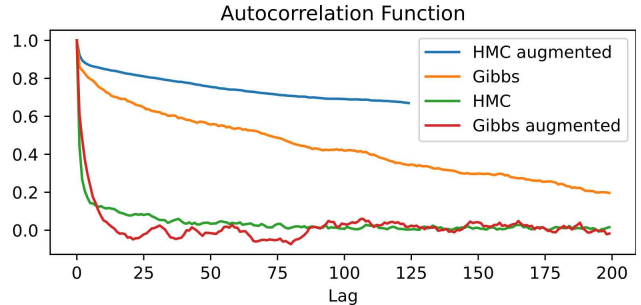

Figure 4: **Samplers comparison.** Autocorrelation function of the same log-density as in Figure 3, computed using the augmented Gibbs and Hamiltonian Monte Carlo on the non-augmented (2) and augmented (12) models. Evaluations were made between iterations such that the CPU time spent on each sample is roughly equal.

but the latter offers a simple and straightforward approach without the need to tune parameters.

6.3 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 a single observation from the final day of the calm weather period studied in Marques et al. (2022). We only considered the spatial location as a the covariate. We note that our model is admittedly too simple, and other variables such as humidity, temperature, altitude, etc. should be used in more realistic models. For our vMQP model we considered

Model	CRPS↓			
	10% Test	**20% Test**	**30% Test**	**40% Test**
Wrapped GP	0.221 ± 0.086	0.283 ± 0.093	0.305 ± 0.175	0.399 ± 0.204
Projected GP	0.067 ± 0.015	0.068 ± 0.009	0.063 ± 0.005	0.072 ± 0.009
vMQP Gaussian	$\mathbf{0.057 \pm 0.011}$	0.064 ± 0.006	0.066 ± 0.007	0.066 ± 0.010
vMQP Exponential	0.058 ± 0.015	$\mathbf{0.060 \pm 0.004}$	$\mathbf{0.057 \pm 0.006}$	$\mathbf{0.061 \pm 0.008}$

Table 1: Model performance comparison using CRPS (see Appendix A.2) across various test set sizes. Results show *mean ± standard deviation* from seven random data partitions.
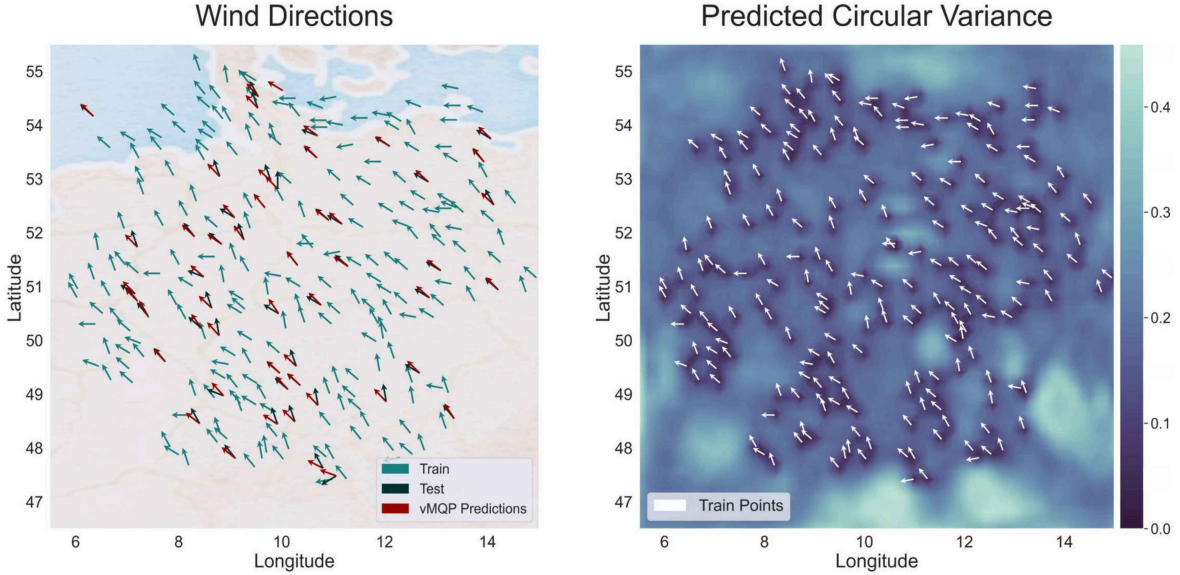

Figure 5: **Left:** Wind directions in 260 weather stations in Germany, randomly split between 208 train and 52 test locations. The predicted circular means from the vMQP model are indicated on the test locations. **Right:** Predicted circular variance over a uniform grid of 60×60 points. Note that the variance grows in regions close to train points with non-aligned directions. Both figures best seen in color.

both Gaussian and Exponential covariance kernels of the forms, respectively,

$$K_{\mathrm{G}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2l^2}\right), \qquad (39)$$

$$K_{\mathrm{E}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{l}\right), \qquad (40)$$

for $i, j = 1 \ldots 260$, with two parameters (σ^2, l). Here $\mathbf{x}_i$ is the 2D spatial coordinates (longitude and latitude) of each weather station.

We compared vMQP predictions with both wrapped and projected GPs, using the implementations of the `CircSpaceTime` R package (Jona Lasinio et al., 2020). Wrapped and Projected GP models were only tested with the Exponential kernel due to numerical instability with the Gaussian kernel. Train/test splitting involved test sizes of 10%, 20%, 30% and 40% of the

data. Figure 5 (Left) shows the training and test wind directions and the circular mean of the vMQP predictions for the 20% test size, while Figure 5 (Right) shows predicted circular variances (see Appendix A for definitions).

All the predictive distributions were evaluated against the test data using the circular continuous ranked probability score (CRPS) (Grimit et al., 2006) (see Appendix A.2 for the definition). The results, presented in Table 1, report the mean and standard deviation of CRPS across seven random train/test splits. The results confirm the advantage of the vMQP model and show that the Exponential kernel is better in most cases. See Appendix C.1 for more details.
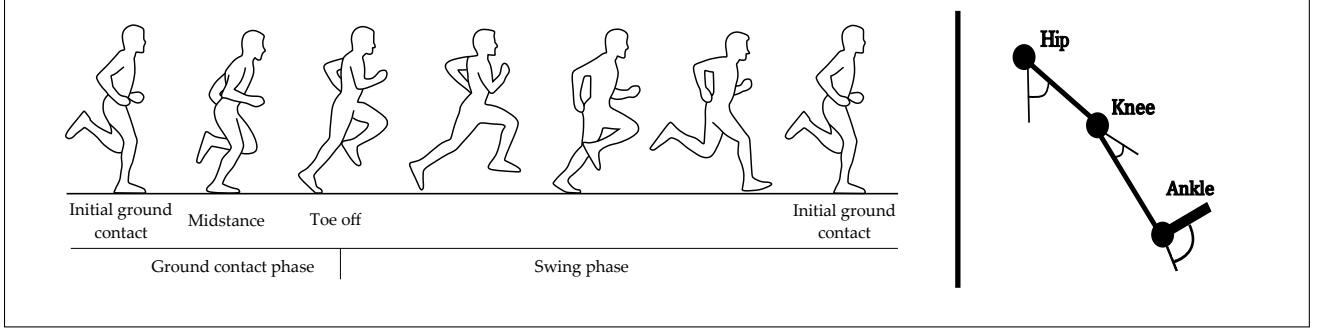
Figure 6: **Left:** Phases of the running gait cycle, defined w.r.t. the right leg of the figure. The cycle begins and ends with the initial ground contact of the right leg. Image: Zrenner (2022). **Right:** The three joint angles used to predict the location in the gait cycle, defined over the sagittal plane.

6.4 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 6. 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). In applications, joint angles and the phases in the running cycle are used to guide control software in exoskeletons (Gad et al., 2022) and controllers for commercial devices (Esquenazi et al., 2012; Sanz-Merodio et al., 2014) and 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 circular, the vMQP 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.[3] See Appendix C.2 for more details. The data consists of values of the three angles defined in Figure 6, estimated at 100 values of the percentage $t \in \{1, 2 \dots 100\}$ from individuals running at five surface gradients $s \in \{0\%, \pm5\%, \pm10\%\}$. We used four gradients $s \in \{\pm5\%, \pm10\%\}$ for training, yielding 400 training points. Training data for $s = \pm10\%$ are shown in Figure 7. Testing was performed on triplets of angles measured at surface gradient $s = 0\%$ at 20 points uniformly selected along the gait cycle.

In this experiment we assumed $\kappa = 0$ (see (1)), because the training points are located uniformly along the full cycle of t, and thus there is no concentrated direction.
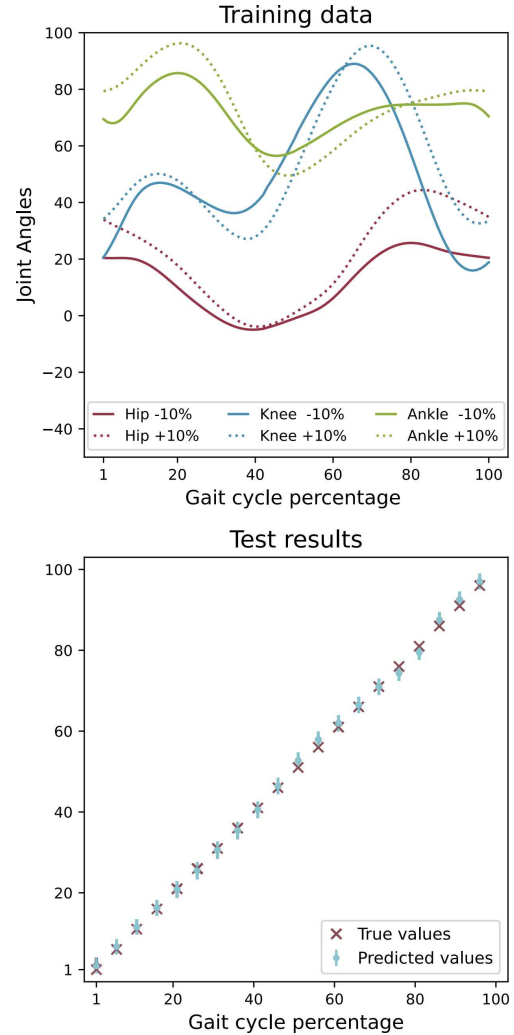

Figure 7: **Predictions of gait cycle percentage**. *Top:* training data for $\pm10\%$ gradients. *Bottom:* Predictions for 20 data points at zero gradient. The error bars are proportional to the circular variance.

---

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

We used an anisotropic exponential kernel of the form

$$K((\boldsymbol{a}_i, s_i), (\boldsymbol{a}_j, s_j)) \qquad i, j = 1 \dots 420, \qquad (41)$$
$$= \sigma^2 \exp\left(-\frac{||\boldsymbol{a}_i - \boldsymbol{a}_j||^2}{2l^2} - \frac{(s_i - s_j)^2}{2g^2}\right),$$

Here $\boldsymbol{a}_i \in \mathbb{R}^3$ contais 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 7 shows the results, showing a high accuracy in the prediction of the gait cycle percentage. For sample traces and histograms see Figure S3 in Appendix C.2.

6.5 The effect of Bridging

In the experiments presented above we did not observe a substantial increase in the MH acceptance rate when using Bridging. However, this might change when using other kernel functions or datasets. We leave further inquires on the effect of Bridging in these models for future work.

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 inference. Relevant future work includes 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 and Sahani, 2016).

Acknowledgements

References

A. Altland and B. D. Simons. *Condensed matter field theory*. Cambridge University Press, 2010. 3

D. Best and N. I. Fisher. Efficient simulation of the von Mises distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2): 152–157, 1979. 3

C-Motion Inc., Germantown, MD, USA. Visual 3d. 17

A. Cappozzo, F. Catani, U. Della Croce, and A. Leardini. Position and orientation in space of bones during movement: anatomical frame definition and determination. *Clinical biomechanics*, 10(4):171–178, 1995. 17

M. A. Carreira-Perpinan and G. Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005. 4

Deutcher Wetterdienst (DWD). https://dwd.de. 6

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. 4

A. Esquenazi, M. Talaty, A. Packel, and M. Saulino. 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. 8

C. Ferrari. *The Wrapping Approach for Circular Data Bayesian Modelling*. PhD thesis, Università di Bologna, Bologna, 2009. 6

S. Friedli and Y. Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, 2017. ISBN 978-1-107-18482-4. doi: 10.1017/9781316882603. 5

M. Gad, B. Lev-Ari, A. Shapiro, C. Ben-David, and R. Riemer. Biomechanical knee energy harvester: Design optimization and testing. *Frontiers in Robotics and AI*, 9:998248, 2022. 8

E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson. 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. 7, 13

D. Hernandez-Stumpfhauser, F. J. Breidt, and M. J. van der Woerd. 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. 6

S. R. Jammalamadaka and A. Sengupta. *Topics in circular statistics*, volume 5. World Scientific, 2001. 1

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

G. Jona-Lasinio, G. Mastrantonio, and A. E. Gelfand. Models for space-time directional data using Wrapped Gaussian processes. In S. Cabras, T. D. Battista, and W. Racugno, editors, *Proceedings of the 47th Scientific Meeting of the Italian Statistical Society*, pages 1–10, Italy, 2014. ISBN 9788884678744. URL http://www.sis2014.it/proceedings/. 6

G. Jona-Lasinio, A. E. Gelfand, and G. Mastrantonio. Spatial and spatio-temporal circular processes with application to wave directions. *Applied Directional Statistics: Modern Methods and Case Studies*, page 35, 2018. 5, 6

G. Jona Lasinio, M. Santoro, and G. Mastrantonio. 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. 6, 7

P. Jupp and K. Mardia. A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review/Revue Internationale de Statistique*, pages 261–294, 1989. 1

D. A. Kessler and M. Bretz. Unbridled growth of spin-glass clusters. *Physical Review B*, 41(7):4778, 1990. 5

R. Kumar, C. Carroll, A. Hartikainen, and O. A. Martin. Arviz a unified library for exploratory analysis of bayesian models in python. *Journal of Open Source Software*, 2019. 13

A. Lee. Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):477–486, 2010. 1

C. Ley and T. Verdebout. *Modern directional statistics*. CRC Press, 2017. 1

C. Ley and T. Verdebout. *Applied directional statistics: modern methods and case studies*. CRC Press, 2018. 1

F. Liang. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010. 5

J. S. Liu. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001. 13

K. V. Mardia and P. E. Jupp. *Directional statistics*. John Wiley & Sons, 1999. 1, 6

J. Markowitz, P. Krishnaswamy, M. F. Eilenberg, K. Endo, C. Barnhart, and H. Herr. 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. 8

I. Marques, T. Kneib, and N. Klein. A non-stationary model for spatially dependent circular response data based on wrapped gaussian processes. *Statistics and Computing*, 32(5):73, 9 2022. 6

J. Martens and I. Sutskever. Parallelizable sampling of Markov random fields. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 517–524. JMLR Workshop and Conference Proceedings, 2010. 3

M. Michel. *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. 5, 9

M. Michel, J. Mayer, and W. Krauth. Event-chain Monte Carlo for classical continuous spin models. *Europhysics Letters*, 112(2):20003, 2015. 5

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006. 5

I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006. 4, 5, 14

A. Navarro, J. Frellsen, and R. Turner. The multivariate generalised von Mises distribution: inference and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 1, 2, 6

R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001. 14

R. M. Neal et al. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011. 6

G. Nuñez-Antonio and E. Gutiérrez-Peña. A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, 32(10):995–1001, 2005. 6

G. Nuñez-Antonio and E. Gutiérrez-Peña. A Bayesian model for longitudinal circular data based on the projected normal distribution. *Computational Statistics & Data Analysis*, 71:506–519, 2014. 6

G. Nuñez-Antonio, E. Gutiérrez-Peña, and G. Escarela. A Bayesian regression model for circular data based on the projected normal distribution. *Statistical Modelling*, 11(3):185–201, 2011. 6

J. Ostmeyer, E. Berkowitz, T. Luu, M. Petschlies, and F. Pittler. The Ising model with Hybrid Monte Carlo. *Computer Physics Communications*, 265: 107978, 2021. 3

A. Pakman. Super-efficient exact Hamiltonian Monte Carlo for the von Mises distribution. *Applied Mathematics Letters*, 159:109284, 2025. 3

J. Park and M. Haran. 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. 4

P. H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973. 5, 14

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

A. Pewsey, M. Neuhäuser, and G. D. Ruxton. *Circular statistics in R*. OUP Oxford, 2013. 1

N. Prokof'ev and B. Svistunov. Worm algorithms for classical statistical models. *Physical review letters*, 87(16):160601, 2001. 5

D. Sanz-Merodio, M. Cestari, J. C. Arevalo, X. A. Carrillo, and E. Garcia. Generation and control of adaptive gaits in lower-limb exoskeletons for motion assistance. *Advanced Robotics*, 28(5):329–338, 2014. 8

A. Shah, A. Wilson, and Z. Ghahramani. Student-t processes as alternatives to Gaussian processes. In *Artificial intelligence and statistics*, pages 877–885. PMLR, 2014. 3

A. Shkedy Rabani, S. Mizrachi, G. S. Sawicki, and R. Riemer. 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. 8, 16

F. Sup, A. Bohara, and M. Goldfarb. Design and control of a powered transfemoral prosthesis. *The International journal of robotics research*, 27(2):263–273, 2008. 8

L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of applied probability*, pages 1–9, 1998. 5, 14

V. Vapnik. Transductive inference and semi-supervised learning. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, chapter 24, pages 453–472. MIT Press, 2006. 4

E. Vértes and M. Sahani. Learning doubly intractable latent variable models via score matching. *Advances in Approximate Bayesian Inference NIPS 2016 Workshop*, 2016. 9

J.-S. Wang. Worm algorithm for two-dimensional spin glasses. *Physical Review E*, 72(3):036706, 2005. 5

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. MIT Press Cambridge, MA, 2006. 3

J. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning*, pages 10292–10302. PMLR, 2020. 3

J. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Pathwise conditioning of Gaussian processes. *The Journal of Machine Learning Research*, 22(1):4741–4787, 2021. 3

D. A. Winter. 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. 8

U. Wolff. Collective Monte Carlo updating for spin systems. *Physical Review Letters*, 62(4):361, 1989. 5

Y. Zhang, Z. Ghahramani, A. J. Storkey, and C. Sutton. Continuous relaxations for discrete Hamiltonian Monte Carlo. *Advances in Neural Information Processing Systems*, 25, 2012. 3

J. Zito and D. Kowal. The projected dynamic linear model for time series on the sphere. *arXiv preprint arXiv:2308.14996*, 2023. 6

M. Zrenner. *From Raw to Big Data in Endurance Running : Application of Data Science Techniques for Knowledge Creation from Wearable Sensor Data*. PhD thesis, 2022. 8

Checklist

1. For all models and algorithms presented, check if you include:

    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

    (a) Statements of the full set of assumptions of all theoretical results. [Yes]

    (b) Complete proofs of all theoretical results. [Yes]

    (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

    (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

    (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

    (a) Citations of the creator If your work uses existing assets. [Yes]

    (b) The license information of the assets, if applicable. [Not Applicable]

    (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

    (d) Information about consent from data providers/curators. [Not Applicable]

    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

    (a) The full text of instructions given to participants and screenshots. [Not Applicable]

    (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

    (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Useful definitions

Given a set of angles ϕ_i, define

$$z = \frac{1}{n} \sum_{i=1}^{n} e^{i\phi_i} \tag{42}$$

$$\equiv R \exp(i\gamma) \quad R > 0, \gamma \in [0, 2\pi] \tag{43}$$

$$\text{Circular mean} \equiv \gamma \tag{44}$$

$$\text{Circular variance} \equiv 1 - R \tag{45}$$

A.1 Relative Effective Sample Size (RESS)

To evaluate the efficiency of the sampled Markov chains we used Relative Effective Sample Size (Liu, 2001),[4] defined as

$$\text{RESS} = \frac{1}{1 + 2 \sum_{l=1}^{\infty} \rho_l}, \tag{46}$$

where ρ_l is the autocorrelation at lag l.

A.2 Circular Continuous Ranked Probability Score (CRPS)

The circular Continuous Ranked Probability Score (CRPS) (Grimit et al., 2006) is a scoring rule used to evaluate the accuracy of probabilistic forecasts by measuring the distance between the predictive distribution and the observed outcome. For circular variables, such as wind directions, CRPS is defined as:

$$\text{CRPS}(F, \xi) = \int_0^{2\pi} \left(F(y) - \mathbf{1}(y - \xi) \right)^2 dy,$$

where:

- $F(y)$ is the cumulative distribution function (CDF) of the predictive distribution,
- ξ is the holdout observation,
- $\mathbf{1}(y - \xi)$ is the Heaviside step function.

In the case of circular variables, the CRPS can be simplified using the circular distance $d(\alpha, \beta) = 1 - \cos(\alpha - \beta)$ as:

$$\text{CRPS}(F, \xi) = \mathbb{E}\left[d(\theta, \xi) \right] - \frac{1}{2} \mathbb{E}\left[d(\theta, \theta') \right],$$

where θ and θ' are independent samples from the predictive distribution F.

This formulation compares the predictive distribution with the observed value and penalizes large spreads in the forecast distribution. Lower CRPS values indicate better predictive performance.

---

[4]For the estimation we used the `arviz` python package (Kumar et al., 2019).

B The Bridging algorithm and its vMQP extension

We first review the derivation of the Bridging algorithm of Murray et al. (2006), and then present an efficient extension to von Mises Quasi-Process models.

B.1 The Bridging algorithm

Since the acceptance rate (36) is known to be maximal (Peskun, 1973; Tierney, 1998), one can expect to increase the acceptance rate (34) by using a better approximation than (37) to the ratio of normalizing constants. The Bridging algorithm (Murray et al., 2006) achieves this by using ideas from Annealed Importance Sampling (AIS) (Neal, 2001). The idea is to further augment (31) with K additional variables $\boldsymbol{\xi}_k$. For this, let us introduce K distributions that interpolate between $f(\cdot|w)$ and $f(\cdot|w')$,

$$p_k(\boldsymbol{\xi}_k|w, w') \propto f_k(\boldsymbol{\xi}_k|w, w') \equiv f(\boldsymbol{\xi}_k|w)^{\beta_k} f(\boldsymbol{\xi}_k|w')^{1-\beta_k} \qquad k = 1, 2 \dots K \tag{47}$$

where we defined

$$\beta_k = \frac{k}{K+1} \qquad k = 1, 2 \dots K \tag{48}$$

Let us now introduce K transition operators $T_k(\boldsymbol{\xi}'_k|\boldsymbol{\xi}_k, w, w')$ satisfying detailed balance w.r.t. the distributions in (47),

$$p_k(\boldsymbol{\xi}_k|w, w')T_k(\boldsymbol{\xi}'_k|\boldsymbol{\xi}_k, w, w') = p_k(\boldsymbol{\xi}'_k|w, w')T_k(\boldsymbol{\xi}_k|\boldsymbol{\xi}'_k, w, w') . \tag{49}$$

Note now that since (47) enjoys the symmetry

$$p_k(\boldsymbol{\xi}|w, w') = p_{K+1-k}(\boldsymbol{\xi}|w', w), \tag{50}$$

we can assume that

$$T_k(\boldsymbol{\xi}'|\boldsymbol{\xi}, w, w') = T_{K+1-k}(\boldsymbol{\xi}'|\boldsymbol{\xi}, w', w) \tag{51}$$

Using these transition operators, we augment the distribution (28) as

$$p(w, w', \boldsymbol{\xi}_0, \dots, \boldsymbol{\xi}_K|\boldsymbol{\varphi}) = p(w|\boldsymbol{\varphi})q(w'|w)\frac{f(\boldsymbol{\xi}_0|w')}{Z[w']} \prod_{k=1}^{K} T_k(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1}, w, w'), \tag{52}$$

and we denoted $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$. In order to sample from (52), we now iterate over

1. Sample $w' \sim q(w'|w)$ and then $\boldsymbol{\xi}_0 \sim f(\boldsymbol{\xi}_0|w')/Z[w']$.

2. Sample

$$\boldsymbol{\xi}_k \sim T_k(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1}, w, w') \qquad k = 1 \dots K \tag{53}$$

3. Propose to exchange $w \leftrightarrow w'$ and $\{\boldsymbol{\xi}_0, \boldsymbol{\xi}_1 \dots \boldsymbol{\xi}_K\} \leftrightarrow \{\boldsymbol{\xi}_K, \boldsymbol{\xi}_{K-1} \dots \boldsymbol{\xi}_0\}$, and accept with Metropolis-Hastings probability

$$\min\left(1, \frac{p(w', w, \boldsymbol{\xi}_K \dots \boldsymbol{\xi}_0|\boldsymbol{\varphi})}{p(w, w', \boldsymbol{\xi}_0 \dots \boldsymbol{\xi}_K|\boldsymbol{\varphi})} = \frac{p(w')q(w|w')f(\boldsymbol{\varphi}|w')}{p(w)q(w'|w)f(\boldsymbol{\varphi}|w)} \prod_{k=0}^{K} \frac{f_{k+1}(\boldsymbol{\xi}_k|w, w')}{f_k(\boldsymbol{\xi}_k|w, w')}\right) . \tag{54}$$

The product expression in the rhs follows directly from (49)-(52), and we extended the definition (47) to include $f_0(\cdot|w, w') = f(\cdot|w')$ and $f_{K+1}(\cdot|w, w') = f(\cdot|w)$. Comparing (34) with (54), we see that the approximation (37) has been replaced by

$$\frac{Z[w]}{Z[w']} \simeq \prod_{k=0}^{K} \frac{f_{k+1}(\boldsymbol{\xi}_k|w, w')}{f_k(\boldsymbol{\xi}_k|w, w')}, \qquad \boldsymbol{\xi}_k \sim f_k(\boldsymbol{\xi}_k|w, w'), \tag{55}$$

which is an AIS estimate (Neal, 2001) of the ratio of normalization constants.

B.2 Efficient Bridging for von Mises Quasi-Processes via double augmentation

Back to our von Mises model, the K intermediate distributions (47) are

$$p_k(\boldsymbol{\xi}_k|w,w') \propto f_k(\boldsymbol{\xi}_k|w,w') = \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{d} M_{k,ij}\cos(\xi_{k,i}-\xi_{k,j}) + \alpha_{k,c}\sum_{i=1}^{d}\cos(\xi_i) + \alpha_{k,s}\sum_{i=1}^{d}\sin(\xi_i)\right\} \quad (56)$$

for $k = 1\ldots K$, where we defined

$$\boldsymbol{M}_k = \beta_k\boldsymbol{M}_w + \beta_{K+1-k}\boldsymbol{M}_{w'}. \tag{57}$$

$$\alpha_{k,c} = \beta_k\kappa\cos(\nu) + (1-\beta_k)\kappa'\cos(\nu') \tag{58}$$

$$\alpha_{k,s} = \beta_k\kappa\sin(\nu) + (1-\beta_k)\kappa'\sin(\nu') \tag{59}$$

Transition operators T_k satisfying detailed balance (49) w.r.t. these distributions can be obtained for each k using the method of Section 3,

$$T_k(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1},w,w') = p(\boldsymbol{\xi}_k|\mathbf{z}_k,w,w')p(\mathbf{z}_k|\boldsymbol{\xi}_{k-1},w,w') \qquad k=1\ldots K, \tag{60}$$

where $p(\mathbf{z}_k|\boldsymbol{\xi}_{k-1},w,w')$ and $p(\boldsymbol{\xi}_k|\mathbf{z}_k,w,w')$ are defined as in (13) and (14)-(17) respectively, using a matrix $\mathbf{A}_k \in \mathbb{R}^{d\times d}$ from a Cholesky decomposition of $\boldsymbol{I}_d\lambda_k - \boldsymbol{M}_k$ with an appropriate λ_k. Note that we treat the variables $\mathbf{z}_k$ as part of the randomness of the transition. However, this approach would require to perform K Cholesky decompositions for each proposal w', hurting the computational advantage we seek to gain for large K.

Instead, we show here that we can do well by reusing the two decompositions,

$$\mathbf{A}_w^\top\mathbf{A}_w = \boldsymbol{I}_d\lambda_w - \boldsymbol{M}_w, \tag{61}$$

$$\mathbf{A}_{w'}^\top\mathbf{A}_{w'} = \boldsymbol{I}_d\lambda_{w'} - \boldsymbol{M}_{w'}, \tag{62}$$

which were performed when sampling fictitious data points from $f(\boldsymbol{\xi}|w)$ and $f(\boldsymbol{\xi}|w')$ in the previous and present Doubly MH steps, respectively. Recall that $\lambda_w, \lambda_{w'}$ were chosen the make the right-hand sides positive definite. The idea is to use two Gaussian augmentations instead of one, to separately eliminate each term of $\mathbf{M}_k$ in (57). More concretely, for each k let us define

$$p_k(\mathbf{y}_1|\boldsymbol{\xi}) = \mathcal{N}\left(\mathbf{y}_1; \beta_k^{1/2}\mathbf{A}_w\cos(\boldsymbol{\xi}), \boldsymbol{I}_d\right) \qquad p_k(\mathbf{y}_2|\boldsymbol{\xi}) = \mathcal{N}\left(\mathbf{y}_2; \beta_k^{1/2}\mathbf{A}_w\sin(\boldsymbol{\xi}), \boldsymbol{I}_d\right) \tag{63}$$

$$p_k(\mathbf{y}_3|\boldsymbol{\xi}) = \mathcal{N}\left(\mathbf{y}_3; \beta_{K+1-k}^{1/2}\mathbf{A}_{w'}\cos(\boldsymbol{\xi}), \boldsymbol{I}_d\right) \qquad p_k(\mathbf{y}_4|\boldsymbol{\xi}) = \mathcal{N}\left(\mathbf{y}_4; \beta_{K+1-k}^{1/2}\mathbf{A}_{w'}\sin(\boldsymbol{\xi}), \boldsymbol{I}_d\right) \tag{64}$$

Let us augment (56) now as

$$p_k(\boldsymbol{\xi},\mathbf{y}|w,w') = p_k(\boldsymbol{\xi}|w,w')p_k(\mathbf{y}_1|\boldsymbol{\xi})p_k(\mathbf{y}_2|\boldsymbol{\xi})p_k(\mathbf{y}_3|\boldsymbol{\xi})p_k(\mathbf{y}_4|\boldsymbol{\xi}) \tag{65}$$

$$\propto \exp\left\{\boldsymbol{\kappa}_c\cdot\cos(\boldsymbol{\xi}) + \boldsymbol{\kappa}_s\cdot\sin(\boldsymbol{\xi}) - \frac{1}{2}\mathbf{y}^\top\mathbf{y}\right\}, \tag{66}$$

where we defined $\mathbf{y} = [\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3,\mathbf{y}_4]$ and

$$\boldsymbol{\kappa}_c = \beta_k^{1/2}\mathbf{y}_1^\top\mathbf{A}_w + \beta_{K+1-k}^{1/2}\mathbf{y}_3^\top\mathbf{A}_{w'} + \alpha_{k,c}\mathbf{1}_d \qquad \boldsymbol{\kappa}_s = \beta_k^{1/2}\mathbf{y}_2^\top\mathbf{A}_w + \beta_{K+1-k}^{1/2}\mathbf{y}_4^\top\mathbf{A}_{w'}. + \alpha_{k,s}\mathbf{1}_d \tag{67}$$

where $\mathbf{1}_d$ is a d-dimensional vector of 1's.

We choose as the operator T_k that satisfies detailed balance w.r.t. (56) the marginal w.r.t. $\mathbf{y}$ of the Gibbs sampling operator that leaves invariant the augmentation (65). Thus to sample (53), we perform Gibbs steps in (65),

$$\mathbf{y} \sim p_k(\mathbf{y}|\boldsymbol{\xi}_{k-1},w,w'), \tag{68}$$

$$\boldsymbol{\xi}_k \sim p_k(\boldsymbol{\xi}_k|\mathbf{y},w,w'), \quad k=1\ldots K \tag{69}$$

and discard $\mathbf{y}$, where $p_k(\mathbf{y}|\boldsymbol{\xi}_{k-1},w,w')$ is the product of the four densities in (63)-(64) and

$$p_k(\boldsymbol{\xi}_k|\mathbf{y},w,w') \propto \prod_{i=1}^{d}\exp(\alpha_i\cos(\xi_{k,i}-\gamma_i)), \tag{70}$$

where

$$\alpha_i = \sqrt{\kappa_{c,i}^2 + \kappa_{s,i}^2} \qquad \tan(\gamma_i) = \frac{\kappa_{s,i}}{\kappa_{c,i}}. \tag{71}$$
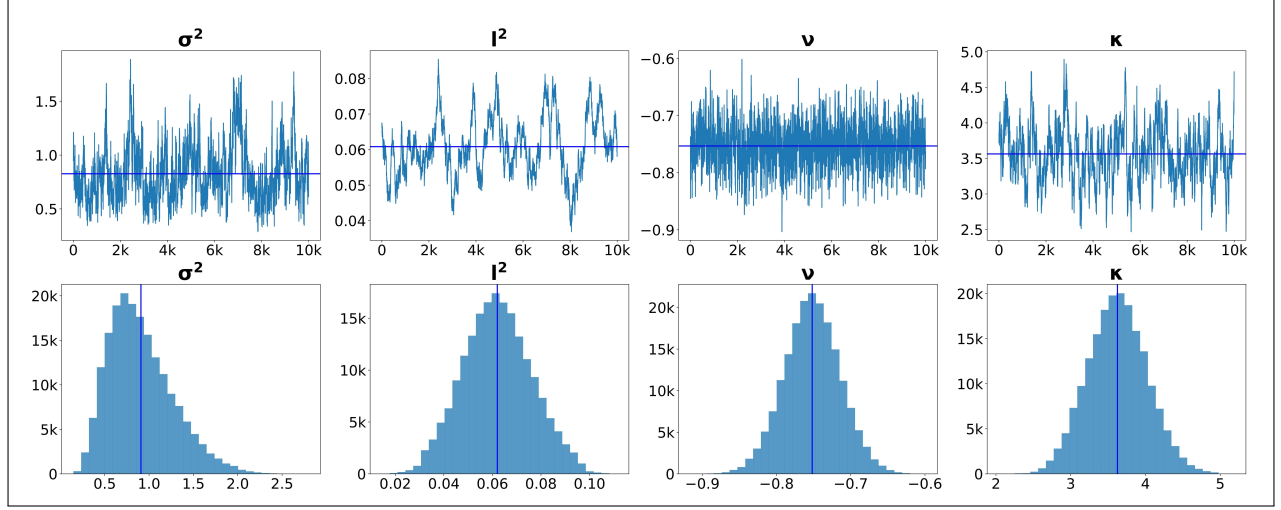
Figure S1: Sample traces and histograms of the parameters of the vMQP model (1) with kernel (39) in the experiment of wind directions prediction from Section 6.3.

C   Details of the experiments

The experiments were performed on a desktop computer with an Intel Core i5-13400F CPU at 2.90GHz.

C.1   Experiment: Wind directions in Germany

We run the sampler for 200K iterations, discarding the first 20K as burn-in.

- Parameter priors:

$$\sigma^2 \sim \mathcal{N}(\sigma^2; 0, 1) \qquad \text{truncated to } \sigma^2 > 0 \tag{72}$$

$$l^2 \sim \mathcal{N}(l^2; 0, 1) \qquad \text{truncated to } l^2 > 0 \tag{73}$$

$$\kappa \sim \mathcal{N}(\kappa; 0, 1) \qquad \text{truncated to } \kappa > 0 \tag{74}$$

$$\nu \sim \text{Uniform}[-\pi, +\pi] \tag{75}$$

- MH Proposals: random walk on all the parameters, with variances adjusted for reasonable acceptance rates.

C.2   Experiment: Prediction of percentage of running cycle from joints angles

- Parameter priors:

$$\sigma^2 \sim \mathcal{N}(\sigma^2; 0, 1) \qquad \text{truncated to } \sigma^2 > 0 \tag{76}$$

$$l^2 \sim \mathcal{N}(l^2; 0, 1) \qquad \text{truncated to } l^2 > 0 \tag{77}$$

$$g^2 \sim \mathcal{N}(g^2; 0, 1) \qquad \text{truncated to } g^2 > 0 \tag{78}$$

- MH Proposals: random walk on all the parameters, with variances adjusted for reasonable acceptance rates.

We run the sampler for 600K iterations, discarding the first 20K as burn-in.

- Parameter priors: positively truncated Gaussians for σ^2, l^2, g^2.

Experimental protocol used to collect the data. In the analysis, we used running data from Shkedy Rabani et al. (2022) who tested 16 healthy adults, with no lower-limb injuries or impairments: 9 males and 7 females
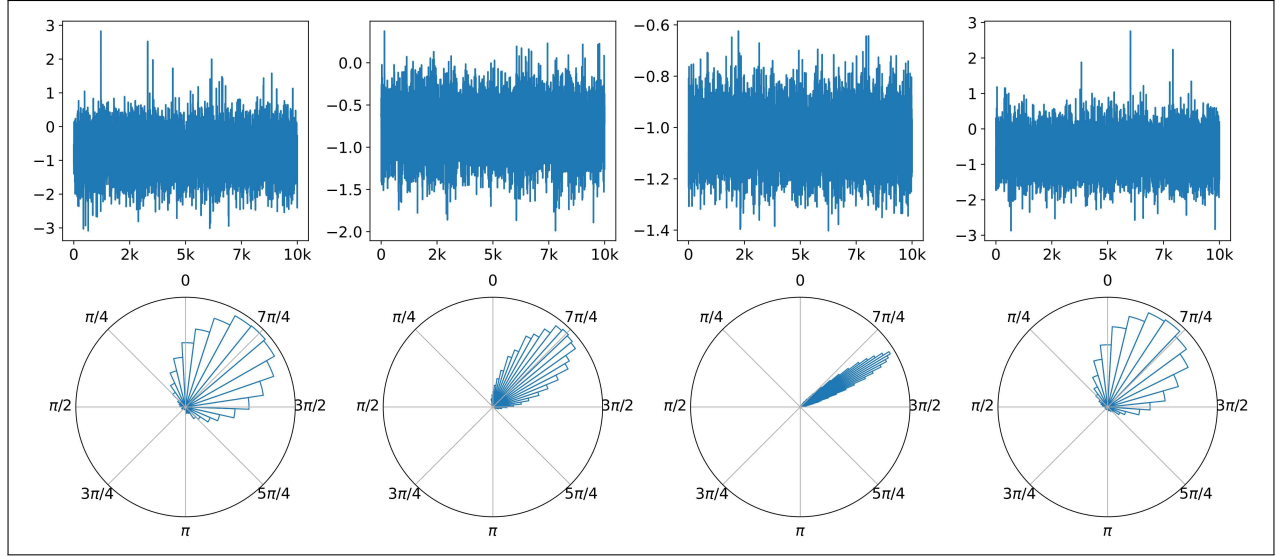
Figure S2: Sample traces and circular histograms of the predicted distribution of wind directions from Section 6.3 at four test locations.

(age: 24.56 ± 3.16 years, [Range 18–28 years]; height: 1.73 ± 0.09 m, [range: 1.55–1.86 m]; mass: 68.01 ± 13.98 kg, [range: 45.0–88.7 kg]). All participants run on an instrumented split-belt treadmill at a speed of 2.25 m/s and at five different surface gradients (-10%, -5%, 0%, +5%, +10%). Motion and force data were collected at each surface gradient using a minimum of 7 gait cycles, with an average of 16 cycles per condition. Using this data for each participant, the average joint kinematics and kinetics were calculated. Motion data were collected using a motion capture system at a frequency of 120 Hz to capture the positions of 22 reflective markers attached to each participant's pelvis and right leg (modification of the Calibration Anatomical System Technique (CAST) marker set (Cappozzo et al., 1995)). The raw data of marker positions was low-pass filtered (Butterworth second order forward and backward passes) with a cut-off frequency of 10 Hz.

Using the Visual 3D software (C-Motion Inc., Germantown, MD, USA), three lower-limb joint angles were obtained. these were ankle-, knee-, and hip-joint angles in sagittal plane (side view), as defined in Figure 6 (Right). The angles were calculated for the right leg only (since both legs behaved symmetrically). The joint angles were normalized in time as percentages of one stride cycle using spline interpolation. All calculated joint angles for each participant were averaged across all the measured gait cycles.
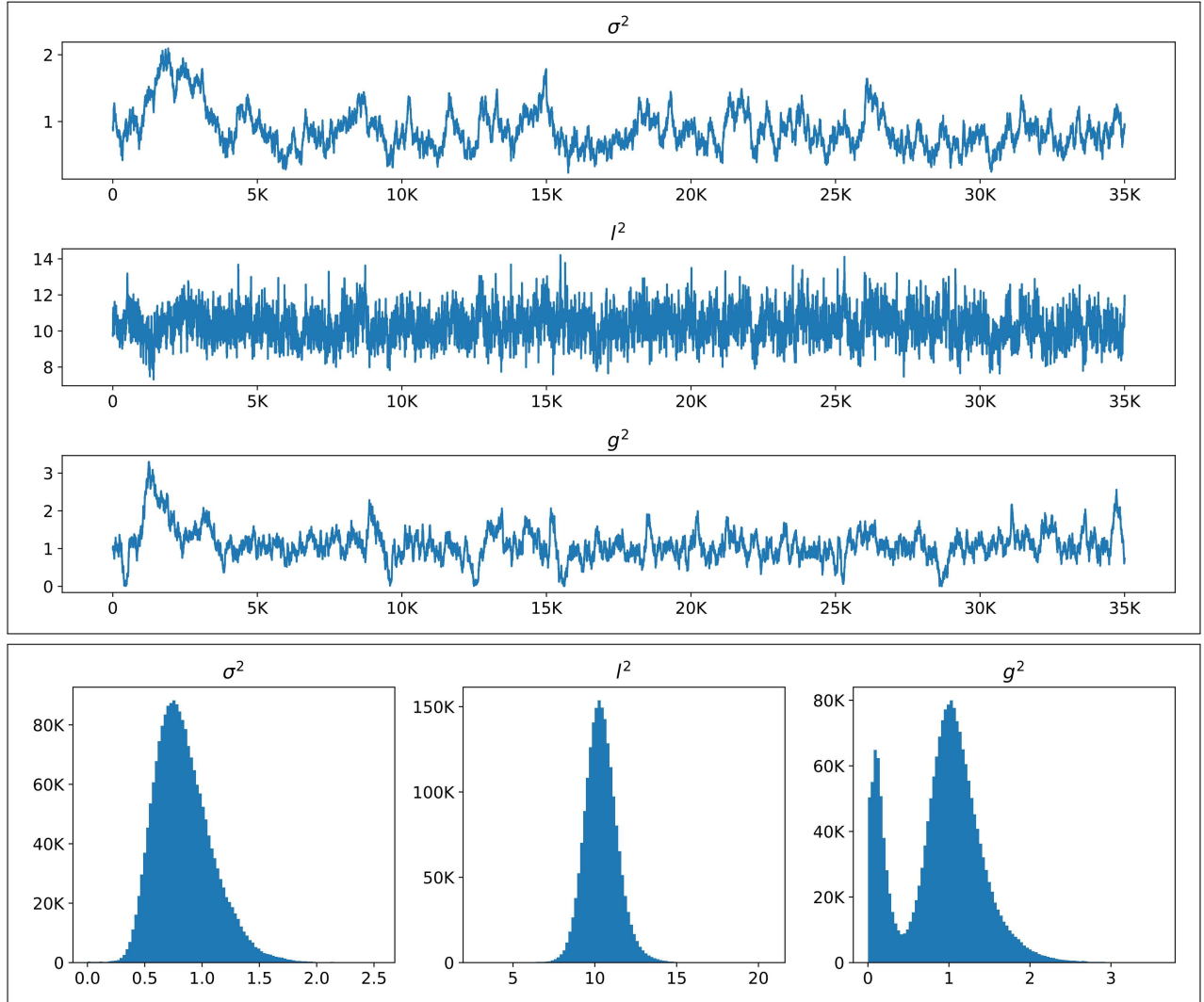
Figure S3: Sample traces and histograms of the parameters of the kernel (41) in the experiment of gait running cycle prediction from Section 6.4. Note the bimodality of the third parameter, correctly captured by our Bayesian approach.