eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Langevin diffusions on the torus: estimation and applications

**Eduardo García-Portugués · Michael Sørensen · Kanti V. Mardia ·
Thomas Hamelryck**

Abstract We introduce stochastic models for continuous-time evolution of angles and develop their estimation. We focus on studying Langevin diffusions with stationary distributions equal to well-known distributions from directional statistics, since such diffusions can be regarded as toroidal analogues of the Ornstein-Uhlenbeck process. Their likelihood function is a product of transition densities with no analytical expression, but that can be calculated by solving the Fokker-Planck equation numerically through adequate schemes. We propose three approximate likelihoods that are computationally tractable: (*i*) a likelihood based on the stationary distribution; (*ii*) toroidal adaptations of the Euler and Shoji-Ozaki pseudo-likelihoods; (*iii*) a likelihood based on a specific approximation to the transition density of the wrapped normal process. A simulation study compares, in dimensions one and two, the approximate transition densities to the exact ones, and investigates the empirical performance of the approximate likelihoods. Finally, two diffusions are used to model the evolution of the backbone angles of the protein G (PDB identifier 1GB1) during a molecular dynamics simulation. The software package `sdetorus` implements the estimation methods and applications presented in the paper.

Eduardo García-Portugués
Department of Mathematical Sciences, University of Copenhagen, Denmark
Present affiliations:
Department of Statistics, Carlos III University of Madrid, Spain
UC3M-BS Institute of Financial Big Data, Carlos III University of Madrid, Spain
E-mail: edgarcia@est-econ.uc3m.es

Michael Sørensen
Department of Mathematical Sciences, University of Copenhagen, Denmark

Kanti V. Mardia
Department of Statistics, University of Leeds, UK
Department of Statistics, University of Oxford, UK

Thomas Hamelryck
Bioinformatics Centre, Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Denmark
Image Section, Department of Computer Science, University of Copenhagen, Denmark

Keywords Circular data · Directional statistics · Likelihood · Protein structure · Stochastic differential equation · Wrapped normal

1 Introduction

Useful proposals of stochastic processes must take into account the particular features of the data that they aim to model. This is so for *toroidal* data, where observations are elements on the torus $\mathbb{T}^p = [-\pi, \pi) \times \overset{p}{\cdots} \times [-\pi, \pi)$ (with $-\pi$ and π identified). Models and inference for *circular* data ($p = 1$) are notably different from the Euclidean case; see Mardia and Jupp (2000) or Jammalamadaka and SenGupta (2001) for a comprehensive description and a review of applications. One of the first continuous-time processes on the circle was proposed by Kent (1975). It is defined as the solution

to the Stochastic Differential Equation (SDE)

$$d\Theta_t = \alpha \sin(\mu - \Theta_t)dt + \sigma dW_t, \tag{1}$$

where $\{W_t\}$ is a Wiener process, $\alpha > 0$ is the strength of the drift towards $\mu \in [-\pi, \pi)$, and $\sigma > 0$ is the diffusion coefficient. This process, referred to below as the von Mises (vM) process, can be regarded as a circular analogue of the Ornstein-Uhlenbeck (OU) process. The process is attracted to μ and, in the neighbourhood of μ, the drift is approximately linear. Moreover, the process is ergodic (*i.e.*, it has a unique stationary distribution) and the stationary distribution (abbreviated as sdi) is $\mathrm{vM}\left(\mu, \frac{2\alpha}{\sigma^2}\right)$. $\mathrm{vM}(\mu, \kappa)$ denotes the vM distribution with probability density function (pdf)

$$f_{\mathrm{vM}}(\theta; \mu, \kappa) := \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi \mathcal{I}_0(\kappa)}, \quad \theta, \mu \in [-\pi, \pi), \quad \kappa \geq 0,$$

with $\mathcal{I}_\nu$ being the modified Bessel function of the first kind and order ν. Despite its similarities with the OU process, the vM process is not as tractable as the former: no analytical expression for its transition probability density (tpd) is known. The vM process has been applied in mathematical biology (Hill and Häder, 1997; Codling and Hill, 2005), and related extensions were studied in physics in the context of oscillators (see Section 5.3.3 in Frank (2005) and references therein).

The contributions of this paper are two-fold. Firstly, we propose ergodic diffusions on the torus whose sdis are well-established distributions from directional statistics. These diffusions can be regarded as toroidal analogues of the OU process. Specifically, we introduce several Langevin diffusions, each defined as the wrapping of a p-dimensional Euclidean diffusion solving the time-homogeneous SDE

$$d\mathbf{X}_t = b(\mathbf{X}_t)dt + \sigma(\mathbf{X}_t)d\mathbf{W}_t, \tag{2}$$

where $b : \mathbb{R}^p \to \mathbb{R}^p$ is the drift, $\sigma : \mathbb{R}^p \to \mathbb{R}^{p \times p}$ is the diffusion coefficient, and $\mathbf{W}_t = (W_{t,1}, \ldots, W_{t,p})'$ is a vector of p independent standard Wiener processes ($'$ denotes transposition). We provide insights on the wrapping of (2) and study the properties of the new diffusions. We give particular emphasis to the Langevin diffusion with Wrapped Normal (WN) sdi, since this is a toroidal OU analogue with more tractable estimation.

Secondly, we present estimation procedures for discretely observed toroidal diffusions. The likelihood function involves the evaluation of the tpd $p_t(\cdot \mid \mathbf{x}_s)$, the density function of the conditional distribution of $\mathbf{X}_{t+s}$ given $\mathbf{X}_s = \mathbf{x}_s$. The tpd solves the Fokker-Planck or Kolmogorov's Forward equation, this is, the Partial Differential Equation (PDE)

$$\frac{\partial}{\partial t} p_t(\mathbf{x} \mid \mathbf{x}_s) = -\sum_{i=1}^p \frac{\partial}{\partial x_i}(b_i(\mathbf{x})p_t(\mathbf{x} \mid \mathbf{x}_s))$$
$$+ \frac{1}{2} \sum_{i,j=1}^p \frac{\partial^2}{\partial x_i \partial x_j}(V_{ij}(\mathbf{x})p_t(\mathbf{x} \mid \mathbf{x}_s)), \tag{3}$$

with $\mathbf{x}, \mathbf{x}_s \in \mathbb{R}^p$, $V(\cdot) := \sigma(\cdot)\sigma(\cdot)'$ and initial condition $p_0(\mathbf{x} \mid \mathbf{x}_s) = \delta(\mathbf{x} - \mathbf{x}_s)$ ($\delta(\cdot)$ represents Dirac's delta). This PDE has no explicit solution except for very few particular choices of b and V. We consider maximum likelihood estimation based on the numerical solution of (3). This method is computationally costly, but serves as a benchmark to which other computationally more expedient methods can be compared. A simple solution is to replace the unknown tpd by the known sdi, hence reducing the problem to maximum likelihood estimation with independent and identically distributed data, but this is usually inefficient and only allows for the estimation of the parameters appearing in the sdi. We therefore develop better approximations to the tpd that are relatively easy to compute. For general diffusions, we introduce toroidal versions of the Euler and Shoji-Ozaki pseudo-likelihoods. For the WN process, we derive a specific, sdi-correct and computationally efficient tpd approximation. We investigate the quality of these estimators by calculating the Kullback-Leibler divergences of the approximating tpds with respect to the tpd obtained by numerically solving (3). Furthermore, in a simulation study for different discretization steps we compare, in the one- and two-dimensional cases, the performance of the proposed approximate likelihoods.

Next, we describe relevant literature to our contributions. Diffusions featuring trigonometric drifts were presented in Kessler and Sørensen (1999), Larsen and Sørensen (2007) and Sørensen (2012), although these processes are not designed to capture periodicity, but rather to have a bounded interval as their state space. Wrapped Gaussian processes have been considered by Jona-Lasinio et al (2012) in the context of spatial modelling of wave directions. In a different setting, processes where the time-inhomogeneous drift $b(t, X_t)$ is a periodic function of time have been studied by Dehay (2015) and Dehling et al (2010). Discrete time processes on the circle include the circular autoregressive models by Breckling (1989) and the Markov processes on the circle by Wehrly and Johnson (1979), Kato (2010) and Yeh et al (2013). In a broader perspective, stochastic calculus on manifolds has been extensively developed, see for example Émery (1989), Stroock (2000) and Hsu (2002). For the case of the torus, a flat and

compact manifold, the modelling challenges do not reside in the curvature of the manifold, but rather in capturing angular dependencies, a non-trivial and ubiquitous problem in directional statistics, consequence of the complex behaviour of rotations on the torus. Finally, we refer to Rogers and Williams (2000), Steele (2001) and Øksendal (2003) for an exhaustive introduction to SDEs, and to Kloeden and Platen (1992) and Iacus (2008) for a more applied perspective.

The rest of this paper is organized as follows. Section 2 introduces diffusions on the torus. Section 3 presents and analyses several estimation procedures for them, whilst the empirical estimation performance is assessed in a simulation study in Section 4. Section 5 gives an application to modelling the evolution of protein backbone angles. Conclusions and final comments are given in Section 6.

2 Toroidal diffusions

The state space of a stochastic process $\{\boldsymbol{\Theta}_t\}$ on the torus is $\mathbb{T}^p = [-\pi, \pi) \times \overset{p}{\cdots} \times [-\pi, \pi)$. The space $\mathbb{R}^p$ also plays a relevant role, since $\{\boldsymbol{\Theta}_t\}$ can be regarded as a Euclidean process $\{\mathbf{X}_t\}$ that is wrapped into its principal angles by $\mathrm{cmod}\,(\cdot) := ((\cdot + \pi) \mod 2\pi) - \pi$. This approach eases the interpretation of crossings through boundaries and motivates the following definition.

Definition 1 (Toroidal diffusion) *The stochastic process $\{\boldsymbol{\Theta}_t\} \subset \mathbb{T}^p$ is said to be a* toroidal diffusion *if it arises as the wrapping $\boldsymbol{\Theta}_t = \mathrm{cmod}\,(\mathbf{X}_t)$ of a diffusion (2) such that b and σ are 2π-periodic:*

$$b(\mathbf{x}+2\mathbf{k}\pi) = b(\mathbf{x}),\ \sigma(\mathbf{x}+2\mathbf{k}\pi) = \sigma(\mathbf{x}),\ \forall \mathbf{k} \in \mathbb{Z}^p,\ \forall \mathbf{x} \in \mathbb{R}^p.$$

The toroidal diffusion coming from the wrapping of (2) is denoted as $\mathrm{d}\boldsymbol{\Theta}_t = b(\boldsymbol{\Theta}_t)\mathrm{d}t + \sigma(\boldsymbol{\Theta}_t)\mathrm{d}\mathbf{W}_t$.

The periodicity of b and σ are required to make $\{\boldsymbol{\Theta}_t\}$ a diffusion, since $\{\boldsymbol{\Theta}_t\}$ can only be Markovian if $\{\mathbf{X}_t\}$ is non-ergodic in $\mathbb{R}^p$, as the next result shows.

Proposition 1 (Wrapped ergodic diffusion) *Let $\{\mathbf{X}_t\}$ be an ergodic diffusion on $\mathbb{R}^p$ with stationary density ν and tpd $p_t(\cdot \,|\, \mathbf{x}_s)$. The following statements hold for the wrapped process $\boldsymbol{\Theta}_t := \mathrm{cmod}\,(\mathbf{X}_t)$:*

i. *$\{\boldsymbol{\Theta}_t\}$ is ergodic on $\mathbb{T}^p$, with stationary density $\nu^{\mathrm{W}}(\cdot) := \sum_{\mathbf{k} \in \mathbb{Z}^p} \nu(\cdot + 2\pi \mathbf{k})$.*
ii. *If $\mathbf{X}_s$ is distributed with density ν, then the conditional density of $\boldsymbol{\Theta}_{t+s} \,|\, \boldsymbol{\Theta}_s = \boldsymbol{\theta}_s$ is*

$$p_t^{\mathrm{W}}(\cdot \,|\, \boldsymbol{\theta}_s) := \sum_{\mathbf{k},\mathbf{m} \in \mathbb{Z}^p} p_t(\cdot + 2\mathbf{k}\pi \,|\, \boldsymbol{\theta}_s + 2\mathbf{m}\pi) w_{\mathbf{m}}(\boldsymbol{\theta}_s),$$

$$w_{\mathbf{m}}(\cdot) := \frac{\nu(\cdot + 2\mathbf{m}\pi)}{\nu^{\mathrm{W}}(\cdot)}. \qquad (4)$$

iii. *If $\{\mathbf{X}_t\}$ is time-reversible, i.e., $p_t(\mathbf{x} \,|\, \mathbf{y})\nu(\mathbf{y}) = p_t(\mathbf{y} \,|\, \mathbf{x})$ $\nu(\mathbf{x})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, then $p_t^{\mathrm{W}}(\boldsymbol{\theta} \,|\, \boldsymbol{\varphi})\nu^{\mathrm{W}}(\boldsymbol{\varphi}) = p_t^{\mathrm{W}}(\boldsymbol{\varphi} \,|\, \boldsymbol{\theta})$ $\nu^{\mathrm{W}}(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta}, \boldsymbol{\varphi} \in \mathbb{T}^p$.*
iv. *The wrapped process is* not *Markovian.*

Proof The statements can be easily checked, so we only illustrate the non-Markovianity. Recall that for $t_0 < t_1 < t_2$ and using the Markovianity of $\{\mathbf{X}_t\}$,

$$\mathbb{P}\{\boldsymbol{\Theta}_{t_2} \in B \,|\, \boldsymbol{\Theta}_{t_1} = \boldsymbol{\theta}_{t_1}, \boldsymbol{\Theta}_{t_0} = \boldsymbol{\theta}_{t_0}\}$$
$$= \sum_{\mathbf{m},\mathbf{n} \in \mathbb{Z}^p} \mathbb{P}\{\boldsymbol{\Theta}_{t_2} \in B \,|\, \mathbf{X}_{t_1} = \boldsymbol{\theta}_{t_1} + 2\mathbf{m}\pi\}$$
$$\times \frac{p_{t_1-t_0}(\boldsymbol{\theta}_{t_1} + 2\mathbf{m}\pi \,|\, \boldsymbol{\theta}_{t_0} + 2\mathbf{n}\pi)}{p_{t_1-t_0}^{\mathrm{W}}(\boldsymbol{\theta}_{t_1} \,|\, \boldsymbol{\theta}_{t_0})} w_{\mathbf{n}}(\boldsymbol{\theta}_{t_0}).$$

This clearly depends on $\boldsymbol{\theta}_{t_0}$ unless p_t is periodic on both arguments, impossible for a density in $\mathbb{R}^p$. □

Thus, a wrapped ergodic diffusion is *not* a diffusion. In particular, the family of conditional distributions given by (4) does not define a semi-group of transition operators. The non-Markovianity arises because $\boldsymbol{\Theta}_{t_2} \,|\, (\boldsymbol{\Theta}_{t_1}, \boldsymbol{\Theta}_{t_0})$, with $t_2 > t_1 > t_0$, does not depend only on $\boldsymbol{\Theta}_{t_1}$ but also on the *winding number* $\mathrm{wind}(\mathbf{X}_{t_1})$ $:= \lfloor \frac{\mathbf{X}_{t_1}+\pi}{2\pi} \rfloor \in \mathbb{Z}^p$ of $\mathbf{X}_{t_1} = \boldsymbol{\Theta}_{t_1} + 2\mathrm{wind}(\mathbf{X}_{t_1})\pi$, hence the requirement for periodic b and σ in Definition 1.

Remark 1 *The density of $\boldsymbol{\Theta}_{t+s} \,|\, \boldsymbol{\Theta}_s = \boldsymbol{\theta}_s$ is remarkably different from the density of $\boldsymbol{\Theta}_{t+s} \,|\, \mathbf{X}_s = \boldsymbol{\theta}_s$, given by $\sum_{\mathbf{k} \in \mathbb{Z}^p} p_t(\cdot + 2\mathbf{k}\pi \,|\, \mathbf{X}_s = \boldsymbol{\theta}_s)$. To make this point clearer, let $\{X_t\}$ be the OU process $\mathrm{d}X_t = \alpha(\mu - X_t)\mathrm{d}t + \sigma\mathrm{d}W_t$, $\{\Theta_t\}$ its wrapped version with sdi $\sum_{k \in \mathbb{Z}} \phi_{\sigma/\sqrt{2\alpha}}(\theta - \mu + 2k\pi)$ (ϕ_σ is the pdf of a $\mathcal{N}(0, \sigma^2)$), and assume $X_s \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\alpha}\right)$. Then:*

i. *The density of $X_{t+s} \,|\, X_s = \theta_s$ is $\phi_{\sigma_t}(\cdot - \mu_t)$, with $\mu_t := \mu + (\theta_s - \mu)e^{-\alpha t}$ and $\sigma_t^2 := \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})$. This is the usual tpd of the OU process.*
ii. *The density of $X_{t+s} \,|\, \Theta_s = \theta_s$ is $\sum_{m \in \mathbb{Z}} \phi_{\sigma_t}(\cdot - \mu_t^m)$ $w_m(\theta_s)$, where $\mu_t^m := \mu + (\theta_s + 2\pi m - \mu)e^{-\alpha t}$ and $w_m(\theta_s) = \frac{\phi_{\sigma/\sqrt{2\alpha}}(\theta_s - \mu + 2\pi m)}{\sum_{k \in \mathbb{Z}} \phi_{\sigma/\sqrt{2\alpha}}(\theta_s - \mu + 2k\pi)}$.*
iii. *The density of $\Theta_{t+s} \,|\, \Theta_s = \theta_s$ is $p_t^{\mathrm{W}}(\cdot \,|\, \Theta_s = \theta_s) = \sum_{k,m \in \mathbb{Z}} \phi_{\sigma_t}(\cdot - \mu_t^m + 2k\pi)\,w_m(\theta_s)$. This circular density can exhibit two modes describing the drift of $\{\Theta_t\}$ towards μ whenever θ_s and μ are antipodal.*
iv. *The density of $\Theta_{t+s} \,|\, X_s = \theta_s$ is $\sum_{k \in \mathbb{Z}} \phi_{\sigma_t}(\cdot - \mu_t + 2k\pi)$, which is unimodal. Whenever the circular shortest distance between θ_s and μ happens across the boundary, this circular density pushes the probability mass in the opposite direction.*

Remark 2 *Liu (2013) stated a similar density to iv above, with $2k\pi e^{-t}$ instead of $2\pi k$, as the "tpd function*

of the OU process on the circle" and proved it satisfied the Chapman-Kolmogorov equation[1]. That density is not circular (it has a time-shrinking period $2k\pi e^{-t}$).

The rest of the section is devoted to the introduction and analysis of notable toroidal diffusions.

2.1 Langevin toroidal diffusions

Let f be a pdf over $\mathbb{R}^p$. The so-called *Langevin diffusions* are a family of multivariate diffusions of the form (2), where the entries of b are given by

$$b_i(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{p} V_{ij}(\mathbf{x})\frac{\partial}{\partial x_j}\log f(\mathbf{x})$$

$$+ \det V(\mathbf{x})^{\frac{1}{2}}\sum_{j=1}^{p}\frac{\partial}{\partial x_j}\left(V_{ij}(\mathbf{x})\det V(\mathbf{x})^{-\frac{1}{2}}\right),\ (5)$$

with $i = 1, \ldots, p$. The most important property of these diffusions is that, under mild regularity conditions on f and σ, they are ergodic with stationary density f. This is particularly convenient since one of the first steps in modelling a given trajectory is to compare its empirical distribution with the sdi of the candidate diffusion model. Remarkably, the family of Langevin diffusions characterizes the family of ergodic diffusions with a given sdi that are time-reversible. The result is due to Kolmogoroff (1937) and was later extended by Kent (1978) using symmetric diffusions on manifolds (see Theorems 4.2 and 6.1 *ibid*). In particular, the OU process is identified as the unique time-reversible diffusion with Gaussian sdi and constant diffusion coefficient. This characterization is key for constructing analogues of the OU process in $\mathbb{T}^p$ by means of Langevin diffusions driven by Gaussian-like toroidal distributions.

The construction of Langevin toroidal diffusions is achieved by wrappings of Langevin diffusions, where now f is a toroidal density, that is, $\int_{\mathbb{T}^p}f(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$ and $f(\boldsymbol{\theta} + 2\mathbf{k}\pi) = f(\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\mathbb{T}^p$, $\mathbf{k}\in\mathbb{Z}^p$.

Proposition 2 *Assume $\{\boldsymbol{\Theta}_t\}$ is obtained from the wrapping of a Langevin diffusion $\{\mathbf{X}_t\}$ with drift (5), given by a strictly positive toroidal density f. Assume that the second derivatives of both f and the entries of V are Hölder continuous, and that V is 2π-periodic. Then, for the given V, $\{\boldsymbol{\Theta}_t\}$ is the unique toroidal time-reversible diffusion that is ergodic with stationary density f and squared diffusion coefficient V.*

---

[1] Note that $-(x_2 - x_1 e^{-(t_2 - t_1)} + 2k\pi e^{-t_2})^2$ should be in the exponential's denominator of Liu (2013)'s (15) and (16).

Proof We provide a sketch. The time-reversibility with equilibrium density f follows from Theorem 10.1 in Kent (1978) using the compactness (makes $\{\boldsymbol{\Theta}_t\}$ conservative), flatness, and global coordinates of $\mathbb{T}^p$. The equilibrium distribution is also the (unique) sdi, so $\{\boldsymbol{\Theta}_t\}$ is ergodic. To show the uniqueness, note that by Theorem 6.1 *ibid* a time-reversible diffusion must have an equilibrium density u and be u-symmetric, where necessarily $u = f$. By Theorem 4.2 *ibid* (and its proof) the only way a diffusion with a given V can be f-symmetric is if its drift is (5). □

As a consequence, any time-reversible toroidal diffusion with stationary density f and $V(\mathbf{x}) = \boldsymbol{\Sigma}$ is of the form

$$\mathrm{d}\boldsymbol{\Theta}_t = \frac{1}{2}\boldsymbol{\Sigma}\nabla\log f(\boldsymbol{\Theta}_t)\mathrm{d}t + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathrm{d}\mathbf{W}_t.$$

The rest of the paper focuses on diffusions of this form.

2.2 Analogues of the Ornstein-Uhlenbeck process

The vM process can be considered as the circular analogue of the OU process (Kent, 1975). Two arguments support this claim: (*i*) the vM process is the unique time-reversible diffusion with vM sdi and constant diffusion coefficient; (*ii*) the vM distribution is usually regarded as the Gaussian circular analogue due to important Gaussian-like characterizations (Jammalamadaka and SenGupta, 2001, Section 2.2.4). However, it is worth to note that a similar argument to (*ii*) holds for the WN: this distribution exhibits certain similarities with the Gaussian (*ibid*, Section 2.2.6) and, contrary to the vM distribution, it appears in Gaussian-related limit laws such as the wrapped version of the central limit theorem (Mardia, 1972, Section 4.3.2).

In this section we investigate the main properties of the Langevin diffusions driven by the multivariate versions of the vM and WN distributions. In addition, we consider two appealing extensions driven by more flexible sdis: the symmetric circular distribution of Jones and Pewsey (2005) and mixtures of (independent) vM distributions. These processes are later employed in Section 4.

2.2.1 Multivariate von Mises

The multivariate extension of the vM distribution is not immediate: several competing alternatives are described in the literature, see Mardia and Frellsen (2012) for a review focused on the bivariate case. Among the available proposals, we chose the Multivariate von Mises (MvM) with *sine interaction* (Mardia et al, 2008) due

to its pleasant modelling properties: simple unimodal characterization, ability of capturing positive/negative dependence within the same density formulation, and vM conditional distributions. The $\mathrm{MvM}(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})$ pdf is

$$f_{\mathrm{MvM}}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) := T(\boldsymbol{\kappa}, \boldsymbol{\Lambda})^{-1} \exp\left\{ \boldsymbol{\kappa}' \cos(\boldsymbol{\theta} - \boldsymbol{\mu}) \right. \\ \left. + \frac{1}{2} \sin(\boldsymbol{\theta} - \boldsymbol{\mu})' \boldsymbol{\Lambda} \sin(\boldsymbol{\theta} - \boldsymbol{\mu}) \right\},$$

where the trigonometric functions are understood as entry-wise operators, $\boldsymbol{\kappa} \geq 0$, $\boldsymbol{\Lambda}$ is a symmetric matrix with zero diagonal, and $T(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ is the normalizing constant. If $\boldsymbol{\Lambda} = \mathbf{0}$, then the MvM distribution is the product of independent vM, and hence $T(\boldsymbol{\kappa}, \mathbf{0}) = (2\pi)^p \prod_{j=1}^p \mathcal{I}_0(\kappa_j)$. A sufficient condition for unimodality is that $\mathbf{P} := \mathrm{diag}(\boldsymbol{\kappa}) - \boldsymbol{\Lambda}$ is positive definite (Mardia and Voss, 2014), a result related to the fact that, for large concentrations $\boldsymbol{\kappa}$, $\mathrm{MvM}(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) \approx \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{P}^{-1})$. The operator $\mathrm{diag}(\cdot)$ denotes either the diagonal extraction or the diagonal matrix construction, depending on its argument.

The non-linear dependence structure of the MvM distribution forces $\boldsymbol{\Sigma}$ in the associated Langevin diffusion to be isotropic (*i.e.*, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$) if separability between the drift and diffusion coefficients is desired. We opted to preserve separability and to generalize (1) by having a $\mathrm{MvM}\big(\boldsymbol{\mu}, \frac{2\boldsymbol{\alpha}}{\sigma^2}, \frac{2\mathbf{A}^*}{\sigma^2}\big)$ sdi:

$$\mathrm{d}\boldsymbol{\Theta}_t = \big[\boldsymbol{\alpha} \circ \sin(\boldsymbol{\mu} - \boldsymbol{\Theta}_t) \\ - (\mathbf{A}^* \sin(\boldsymbol{\mu} - \boldsymbol{\Theta}_t)) \circ \cos(\boldsymbol{\mu} - \boldsymbol{\Theta}_t)\big]\mathrm{d}t + \sigma\mathrm{d}\mathbf{W}_t,$$

where $\circ$ denotes the element-wise product of matrices, $\boldsymbol{\alpha} := \mathrm{diag}(\mathbf{A})$, $\mathbf{A}^* := \mathrm{diag}(\boldsymbol{\alpha}) - \mathbf{A}$, and $\mathbf{A}$ is a positive definite matrix. The equilibrium points of drift are located at $\boldsymbol{\mu} + \mathbf{k}_0\pi$, with $\mathbf{k}_0 \in \{-1, 0, 1\}^p$ (we assume implicit wrapping by cmod in the sums of angles in this section), and are unstable if any component is antipodal, this is, unless $\mathbf{k}_0 = \mathbf{0}$ (see Figures 1 and 2). The drift is approximately linear in a neighbourhood of $\boldsymbol{\mu}$ and has Jacobian $-\mathbf{A}$. For the unstable points, the drift has Jacobian $-\mathbf{A} \circ (\mathbf{ss}')$, with $\mathbf{s} = \cos(\mathbf{k}_0\pi)$ a vector of signs. In the circular case, the maximal drifts (in absolute value) towards μ are placed at $\mu \pm \frac{\pi}{2}$ (see Figure 1). For the general case, the maximal *marginal* drifts for the j-th component happen at $\mu_j - \tan^{-1}\big(A_{jj}\big[\sum_{k \neq j} A_{jk} \sin(\mu_k - \theta_k)\big]^{-1}\big) + k_0\pi$, $k_0 \in \{-1, 0, 1\}$.

2.2.2 Wrapped normal

The pdf of a (multivariate) wrapped normal, $\mathrm{WN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is given by $f_{\mathrm{WN}}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \sum_{\mathbf{k} \in \mathbb{Z}^p} \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\theta} - \boldsymbol{\mu} + 2\mathbf{k}\pi)$,

with $\boldsymbol{\mu} \in \mathbb{T}^p$, $\boldsymbol{\Sigma}$ a covariance matrix, and $\phi_{\boldsymbol{\Sigma}}$ the pdf of a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. For the sake of clarity of exposition, we first introduce the circular case and then the multivariate extension. Using the OU parametrization, the circular WN process with $\mathrm{WN}\big(\mu, \frac{\sigma^2}{2\alpha}\big)$ sdi is defined as

$$\mathrm{d}\Theta_t = \left[\alpha \sum_{k \in \mathbb{Z}} (\mu - \Theta_t - 2k\pi) w_k(\Theta_t)\right]\mathrm{d}t + \sigma\mathrm{d}W_t, \quad (6)$$

$$w_k(\theta) = \frac{\phi_{\sigma/\sqrt{2\alpha}}(\theta - \mu + 2k\pi)}{\sum_{m \in \mathbb{Z}} \phi_{\sigma/\sqrt{2\alpha}}(\theta - \mu + 2m\pi)}.$$

Despite the similar shape of the vM and WN densities in the main bulk of the probability, their behaviour is substantially different at antipodality, a fact strengthened in log scale. The WN process drift is a smoothed "sawtooth wave" that has negative slope at μ and crosses the x-axis at $\mu + k\pi$, $k \in \{-1, 0, 1\}$. Hence, the drift behaves almost linearly in a neighbourhood of μ (equilibrium point, stable) and rapidly decays to pass across $\mu \pm \pi$ (equilibrium point, unstable). This neighbourhood is larger than for the vM process. There is no separability between α and σ and both alter the drift non-trivially. For example, the drift maxima are implicitly given by $\sum_{k \in \mathbb{Z}} k^2 w_k(\theta) - \big[\sum_{k \in \mathbb{Z}} k w_k(\theta)\big]^2 = \frac{\sigma^2}{8\alpha\pi^2}$, and vary from $\mu \pm \pi$ (if $\frac{\sigma^2}{2\alpha} \to 0$, the sdi is degenerate at μ) to $\mu \pm \frac{\pi}{2}$ (if $\frac{\sigma^2}{2\alpha} \to \infty$, the sdi is uniform and the drift is null). Thereby, the maximum drifts always happen closer to antipodality than in the vM process (see Figure 1). The slopes of the drift at μ and $\mu \pm \pi$ are $-\alpha + a_1(\alpha, \sigma)$ and $a_2(\alpha, \sigma)$, respectively, where

$$0 \leq a_1(\alpha, \sigma) := \frac{8\pi^2\alpha^2}{\sigma^2} \sum_{k \in \mathbb{Z}} k^2 w_k(\mu) \leq \alpha,$$

$$0 \leq a_2(\alpha, \sigma) := -\alpha + \frac{2\pi^2\alpha^2}{\sigma^2}\left[4\sum_{k \in \mathbb{Z}} k^2 w_k(\mu + \pi) - 1\right].$$

The lower and upper bounds for $a_1(\alpha, \sigma)$ (respectively, $a_2(\alpha, \sigma)$) are attained, with α fixed, when $\sigma \to 0$ ($\sigma \to \infty$) and $\sigma \to \infty$ ($\sigma \to 0$), respectively.

The multivariate extension of (6) is the diffusion

$$\mathrm{d}\boldsymbol{\Theta}_t = \left[\mathbf{A} \sum_{\mathbf{k} \in \mathbb{Z}^p} (\boldsymbol{\mu} - \boldsymbol{\Theta}_t - 2\mathbf{k}\pi) w_{\mathbf{k}}(\boldsymbol{\Theta}_t)\right]\mathrm{d}t \\ + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathrm{d}\mathbf{W}_t, \quad (7)$$

$$w_{\mathbf{k}}(\boldsymbol{\theta}) = \frac{\phi_{\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma}}(\boldsymbol{\theta} - \boldsymbol{\mu} + 2\mathbf{k}\pi)}{\sum_{\mathbf{m} \in \mathbb{Z}^p} \phi_{\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma}}(\boldsymbol{\theta} - \boldsymbol{\mu} + 2\mathbf{m}\pi)}.$$

This diffusion has $\mathrm{WN}\big(\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma}\big)$ sdi, provided that $\mathbf{A}$ is invertible and such that $\mathbf{A}^{-1}\boldsymbol{\Sigma}$ is a covariance matrix. The drift is null at $\boldsymbol{\mu} + \mathbf{k}_0\pi$, with $\mathbf{k}_0 \in \{-1, 0, 1\}^p$, since $\sum_{\mathbf{k} \in \mathbb{Z}^p}(2\mathbf{k} + \mathbf{k}_0)w_{\mathbf{k}}(\boldsymbol{\mu} + \mathbf{k}_0\pi) = \mathbf{0}$ due to the
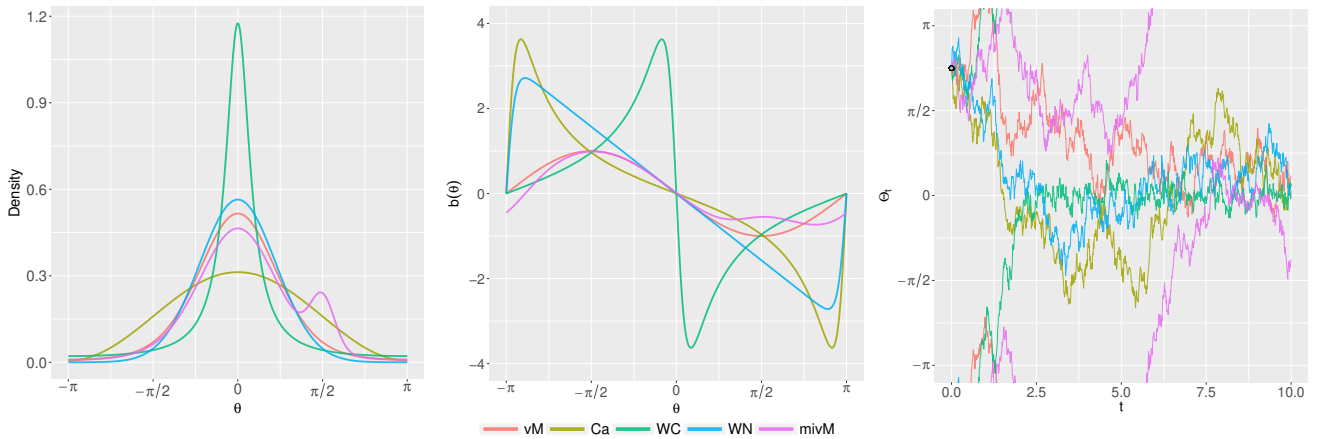
Fig. 1 Stationary densities (left panel), drifts (center), and sample trajectories (right) of the one-dimensional vM, WN, WC, Ca, and mivM diffusions. The parameters for the unimodal diffusions are $\mu = 0$ and $\alpha = \sigma = 1$. The parameters for the mivM diffusion are $\boldsymbol{\alpha} = (1, 10)$, $\boldsymbol{\mu} = (0, \frac{\pi}{2})$, $\mathbf{p} = (0.9, 0.1)$, and $\sigma = 1$. The sample trajectories are simulated in the interval $[0, 10]$ with initial point $\frac{3\pi}{4}$.
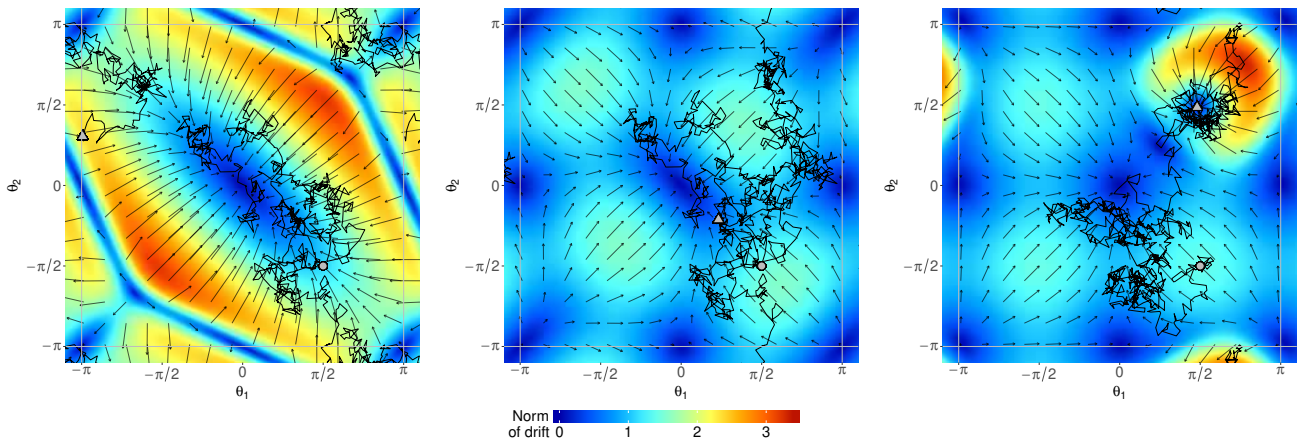

Fig. 2 Vector fields of the drift for the two-dimensional diffusions WN (left panel), MvM (center), and mivM (right). The vector field is coloured by the Euclidean norm of the drift. Parameter $\sigma = 1.5$ is common to all panels and $\boldsymbol{\mu} = (0, 0)$ is used for first and second. For the MvM and WN processes, $\mathbf{A} = (1, 0.5; 0.5, 1)$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ are considered. For the mivM process, $\mathbf{M} = (0, 0; \frac{\pi}{2}, \frac{\pi}{2})$, $\mathbf{A} = (1, 1; 4, 4)$, and $\mathbf{p} = (0.8, 0.2)$. A sample trajectory starting at $(-\frac{\pi}{2}, -\frac{\pi}{2})$ and running in the time interval $[0, 5]$ is drawn, with round and triangular facets denoting the start and end of the simulated path, respectively.

symmetry of $w_{\mathbf{k}}(\boldsymbol{\mu})$ as a function of $\mathbf{k} \in \mathbb{Z}^p$. Properties similar to the circular case can be obtained using that $\nabla w_{\mathbf{k}}(\boldsymbol{\theta}) = 4\pi \boldsymbol{\Sigma}^{-1} \mathbf{A} w_{\mathbf{k}}(\boldsymbol{\theta}) \left[\sum_{\mathbf{m} \in \mathbb{Z}^p} \mathbf{m} w_{\mathbf{m}}(\boldsymbol{\theta}) - \mathbf{k}\right]$. For instance, the Jacobian of the drift at $\boldsymbol{\mu}$ is $-\mathbf{A} + 8\pi^2 \mathbf{A} \left[\sum_{\mathbf{k} \in \mathbb{Z}^p} \mathbf{k} \mathbf{k}' w_{\mathbf{k}}(\boldsymbol{\mu})\right] \mathbf{A}' \boldsymbol{\Sigma}^{-1}$.

The vector field of the drift has a characteristic tessellated structure that, in the two-dimensional case, is formed by hexagonal-like tiles (see Figure 2). $\boldsymbol{\Sigma}$ alters the tessellation that binds the drifts $\mathbf{A}(\boldsymbol{\mu} - \boldsymbol{\theta} - 2\mathbf{k}\pi)$ by modifying $\{w_{\mathbf{k}}(\boldsymbol{\theta}) : \mathbf{k} \in \mathbb{Z}^p\}$. This set is the distribution of the winding numbers of $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$, since $\mathbb{P}\{\text{wind}(\mathbf{X}) = \mathbf{k} \,|\, \text{cmod}(\mathbf{X}) = \boldsymbol{\theta}\} = w_{\mathbf{k}}(\boldsymbol{\theta})$ and satisfies that $\arg\max_{\mathbf{k} \in \mathbb{Z}^p} w_{\mathbf{k}}(\boldsymbol{\theta}) = \text{wind}(\boldsymbol{\mu} - \boldsymbol{\theta})$. Under isotropy (*i.e.* $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$), the larger (respectively, smaller) σ, the more spread (concentrated) the distribution of wind-

ing numbers is, resulting in flat (peaked) drifts with smooth (rough) transitions in the limits defining the tessellation.

2.2.3 Jones and Pewsey (2005) circular distribution

The Jones and Pewsey (2005) distribution, $\text{JP}(\mu, \kappa, \psi)$, is a tractable family of symmetric and unimodal circular distributions that contains the Wrapped Cauchy (WC, $\psi = -1$), Cardioid (Ca, $\psi = 1$), and von Mises ($\psi \to 0$) distributions. Its pdf is $f_{\text{JP}}(\theta; \mu, \kappa, \psi) := (2\pi P_{1/\psi}(\cosh(\kappa\psi)))^{-1}(\cosh(\kappa\psi) + \sinh(\kappa\psi)\cos(\theta - \mu))^{1/\psi}$, with $\mu \in [-\pi, \pi)$, $\kappa \geq 0$, $\psi \in \mathbb{R}$, and P_ν the Legendre function of the first kind and order ν.

The diffusion with $\mathrm{JP}\big(\mu, \frac{2\alpha}{\sigma^2}, \psi\sigma^2\big)$ sdi, parametrised to yield (1) as a particular case, is

$$\mathrm{d}\Theta_t = \frac{\sinh(2\alpha\psi)\sin(\mu - \Theta_t)}{\psi(\cosh(2\alpha\psi) + \sinh(2\alpha\psi)\cos(\mu - \Theta_t))}\mathrm{d}t + \sigma\mathrm{d}W_t.$$

The maximal drifts, located at $\mu \pm (\frac{\pi}{2} + \sin^{-1}(\tanh(2\alpha\psi)))$, are closer to the equilibrium mean μ when $\psi < 0$ and to the antipodal mean when $\psi > 0$. The slope of the drift at μ is $\frac{e^{4\alpha\psi}-1}{4\psi}$. At $\mu \pm \pi$, it is $\frac{e^{-4\alpha\psi}-1}{4\psi}$. This relates to the fact that the drifts with $\psi < 0$ equal the ones with $\psi > 0$, once translated by $\pm\pi$ and reflected around μ. Hence, whilst the WC diffusion features a drift attracting the process towards a tight neighbourhood around μ, the Ca diffusion repulses the process from $\mu \pm \pi$ and weakly attracts it towards μ (see trajectories and drifts in Figure 1).

2.2.4 Mixtures of independent von Mises

The density of an m-mixture of independent von Mises distributions, $\mathrm{mivM}(\mathbf{M}, \mathbf{K}, \mathbf{p})$, is given by $f_{\mathrm{mivM}}(\boldsymbol{\theta}; \mathbf{M}, \mathbf{K}, \mathbf{p}) := \sum_{j=1}^{m} p_j f_{\mathrm{MvM}}(\boldsymbol{\theta}; \boldsymbol{\mu}_j, \boldsymbol{\kappa}_j, \mathbf{0})$, with $\mathbf{M} := (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m)'$, $\mathbf{K} := (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_m)'$, $\mathbf{p} := (p_1, \ldots, p_m)'$, and $p_j \geq 0$, $j = 1, \ldots, m$ and $\sum_{j=1}^{m} p_j = 1$. The mivM distribution is a highly flexible tool for modelling multimodal and skewed circular data (Banerjee et al, 2005), and has tractability as a key advantage: the normalizing constant is known and estimation by the Expectation-Maximization (EM) algorithm is relatively easy. Setting $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)'$ and $\boldsymbol{\kappa}_j = \frac{2\boldsymbol{\alpha}_j}{\sigma^2}$,

$$\mathrm{d}\boldsymbol{\Theta}_t = \left[\sum_{j=1}^{m} \boldsymbol{\alpha}_j \circ \sin(\boldsymbol{\mu}_j - \boldsymbol{\Theta}_t) v_j(\boldsymbol{\Theta}_t) \right] \mathrm{d}t + \sigma\mathrm{d}\mathbf{W}_t,$$

$$v_j(\boldsymbol{\theta}) = \frac{p_j f_{\mathrm{MvM}}\big(\boldsymbol{\theta}; \boldsymbol{\mu}_j, \frac{2}{\sigma^2}\boldsymbol{\alpha}_j, \mathbf{0}\big)}{\sum_{l=1}^{m} p_l f_{\mathrm{MvM}}\big(\boldsymbol{\theta}; \boldsymbol{\mu}_l, \frac{2}{\sigma^2}\boldsymbol{\alpha}_l, \mathbf{0}\big)}, \tag{8}$$

has $\mathrm{mivM}\big(\mathbf{M}, \frac{2\mathbf{A}}{\sigma^2}, \mathbf{p}\big)$ sdi. The mivM process drift is a weighted average of the corresponding component drifts, whose weights are the posterior probabilities of drawing $\boldsymbol{\Theta}_t$ from the mixture components of the sdi. The drift behaves locally around $\boldsymbol{\mu}_j$ as $\boldsymbol{\alpha}_j \circ \sin(\boldsymbol{\theta} - \boldsymbol{\mu}_j) v_j(\boldsymbol{\mu}_j) + \mathbf{b}_j$, with $\mathbf{b}_j = \sum_{k \neq j} \boldsymbol{\alpha}_k \circ \sin(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j) v_k(\boldsymbol{\mu}_j)$ (Figures 1 and 2). Then, $\boldsymbol{\mu}_j$ is only an *asymptotic* equilibrium point for $\sigma \to 0$, since $\lim_{\sigma \to 0} v_k(\boldsymbol{\mu}_j) = \delta_{jk}$. The larger σ, the smoother the binding of the component drifts is.

3 Estimation for toroidal diffusions

We focus now on the estimation of the vector parameter $\boldsymbol{\lambda}$ of a toroidal diffusion

$$\mathrm{d}\boldsymbol{\Theta}_t = b(\boldsymbol{\Theta}_t; \boldsymbol{\lambda})\mathrm{d}t + \sigma(\boldsymbol{\Theta}_t; \boldsymbol{\lambda})\mathrm{d}\mathbf{W}_t, \tag{9}$$

when the data are observations at discrete time points, $\{\boldsymbol{\Theta}_{\Delta i}\}_{i=0}^{N}$. For simplicity, we assume that the time points are equidistant in the time interval $[0, T]$, $T = N\Delta$. The Maximum Likelihood Estimator (MLE) for $\boldsymbol{\lambda} \in \Lambda$ is given by

$$\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}} := \arg\max_{\boldsymbol{\lambda} \in \Lambda} l(\boldsymbol{\lambda}; \{\boldsymbol{\Theta}_{\Delta i}\}_{i=0}^{N}),$$

where, using the Markovianity of (9), the log-likelihood is given by

$$l\left(\boldsymbol{\lambda}; \{\boldsymbol{\Theta}_{\Delta i}\}_{i=0}^{N}\right) = \log p(\boldsymbol{\Theta}_0; \boldsymbol{\lambda})$$
$$+ \sum_{i=1}^{N} \log p_\Delta(\boldsymbol{\Theta}_{\Delta i} \,|\, \boldsymbol{\Theta}_{\Delta(i-1)}; \boldsymbol{\lambda}). \tag{10}$$

Here $p_\Delta(\cdot \,|\, \cdot; \boldsymbol{\lambda})$ is the tpd of (9). The first term in (10) is often disregarded or set to the sdi of (9). Maximum likelihood estimation is, under weak regularity conditions, consistent and asymptotically efficient when $N \to \infty$ with fixed Δ (Dacunha-Castelle and Florens-Zmirou, 1986), or when $\Delta \to 0$ and $T \to \infty$ (Sørensen, 2008). However, it can rarely be readily performed, as usually no explicit expression for the tpd exists and this tpd is only given implicitly as the solution to (3) on $\mathbb{T}^p$.

In the following we present and analyse several estimation strategies to circumvent the unavailability of the tpd when dealing with toroidal diffusions. All these methods rely on an *approximate likelihood function*, where the unknown tpd is replaced by an approximation. For the sake of brevity, we suppress the, implicitly assumed, dependence on $\boldsymbol{\lambda}$ in the notation.

3.1 Estimation based on the stationary distribution

The simplest approximate likelihood function is obtained by replacing the tpd by the stationary density of (9). Usually, the sdi depends only on a function $\boldsymbol{\lambda}^\nu$ of $\boldsymbol{\lambda}$. For instance, for the WN process, $\boldsymbol{\lambda} = (\mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\lambda}^\nu = (\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$. Therefore, we denote the stationary density by $\nu(\cdot; \boldsymbol{\lambda}^\nu)$ and state the *stationary* MLE of $\boldsymbol{\lambda}^\nu$ as

$$\hat{\boldsymbol{\lambda}}_{\mathrm{SMLE}}^\nu := \arg\max_{\boldsymbol{\lambda}^\nu \in \Lambda^\nu} \sum_{i=0}^{N} \log \nu(\boldsymbol{\Theta}_i; \boldsymbol{\lambda}^\nu). \tag{11}$$

For the vM process, SMLE is semi-explicit (Mardia and Jupp, 2000, page 198). The JP distribution has implicit SMLE and is discussed in Jones and Pewsey (2005, Section 3). Effective estimation in MvM distributions involves pseudo-likelihood (Mardia et al, 2008). Finally, inference for mivM distributions can be carried out by the EM algorithm (Banerjee et al, 2005). The simple SMLE is of interest for three reasons: (i) for stationary ergodic processes, it is consistent for $\boldsymbol{\lambda}^{\nu}$ as $N \to \infty$ for fixed Δ (Kessler, 2000); (ii) $\hat{\boldsymbol{\lambda}}_{\mathrm{SMLE}}^{\nu}$ can be used as a sensible starting value in the optimization routines of more sophisticated procedures; (iii) it can be supplemented by estimators of the rest of $\boldsymbol{\lambda}$ (see Bibby and Sørensen (2001), for example).

When the unidentifiability of $\boldsymbol{\lambda}$ by SMLE involves the diffusion matrix $\boldsymbol{\Sigma}$, an estimator of $\boldsymbol{\lambda}$ can be obtained by an estimator of $\boldsymbol{\Sigma}$ that is unrelated to the SMLE. Conditionally on $\boldsymbol{\Theta}_{\Delta(i-1)}$, $\boldsymbol{\Theta}_{\Delta i}$ is approximately distributed as $\mathrm{WN}(\boldsymbol{\Theta}_{\Delta(i-1)}, \Delta\boldsymbol{\Sigma})$ when $\Delta \approx 0$ (high frequency observations). This, plus the high concentration of such WN distribution (see Remark 3 below), gives

$$p_\Delta(\boldsymbol{\Theta}_{\Delta i} \mid \boldsymbol{\Theta}_{\Delta(i-1)}) \approx f_{\mathrm{WN}}(\boldsymbol{\Theta}_{\Delta i}; \boldsymbol{\Theta}_{\Delta(i-1)}, \Delta\boldsymbol{\Sigma})$$
$$\approx \phi_{\Delta\boldsymbol{\Sigma}}\left(\mathrm{cmod}\left(\boldsymbol{\Theta}_{\Delta i} - \boldsymbol{\Theta}_{\Delta(i-1)}\right)\right).$$

Thus, an approximate MLE of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{HF}} := \frac{1}{N\Delta} \sum_{i=1}^{N} \mathrm{cmod}\left(\boldsymbol{\Theta}_{\Delta i} - \boldsymbol{\Theta}_{\Delta(i-1)}\right)$$
$$\cdot \mathrm{cmod}\left(\boldsymbol{\Theta}_{\Delta i} - \boldsymbol{\Theta}_{\Delta(i-1)}\right)'. \tag{12}$$

Under isotropy, $\hat{\sigma}_{\mathrm{HF}}^2 := p^{-1}\mathrm{tr}[\hat{\boldsymbol{\Sigma}}_{\mathrm{HF}}]$. The Euclidean counterpart of (12) is well-known to be a consistent estimator of $\boldsymbol{\Sigma}$ as $\Delta \to 0$ (for fixed T) due to the convergence in probability to the quadratic variation. The consistency for $\hat{\boldsymbol{\Sigma}}_{\mathrm{HF}}$ follows easily from this result.

The estimator (12) gives a practical method to disentangle the unidentifiability inherent to SMLE. We illustrate this with the WN process. The SMLE $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{S}})_{\mathrm{SMLE}}$ for $(\boldsymbol{\mu}, \mathbf{S})$, where $\mathbf{S} = \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma}$, can be found by optimizing (11). The circular means, $\hat{\boldsymbol{\mu}}_c := \mathrm{atan2}\left(\sum_{i=1}^{N} \sin(\boldsymbol{\Theta}_{i\Delta}), \sum_{i=1}^{N} \cos(\boldsymbol{\Theta}_{i\Delta})\right)$, and the high-concentration estimate of $\mathbf{S}$, $\frac{1}{N}\sum_{i=1}^{N} \mathrm{cmod}\left(\boldsymbol{\Theta}_{i\Delta} - \hat{\boldsymbol{\mu}}_c\right)\mathrm{cmod}\left(\boldsymbol{\Theta}_{i\Delta} - \hat{\boldsymbol{\mu}}_c\right)'$, can be used as starting values. $(\hat{\boldsymbol{\mu}}, \hat{\mathbf{S}})_{\mathrm{SMLE}}$ and (12) give $\hat{\mathbf{A}} := \frac{1}{2}\hat{\boldsymbol{\Sigma}}_{\mathrm{HF}}\mathbf{S}_{\mathrm{SMLE}}^{-1}$, resulting in $\hat{\boldsymbol{\lambda}} = (\hat{\mathbf{A}}, \hat{\boldsymbol{\mu}}_{\mathrm{SMLE}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{HF}})$. Similar approaches can be followed for the rest of the diffusions presented in Section 2.

3.2 Adapted Euler and Shoji-Ozaki pseudo-likelihoods

The well-known Euler pseudo-likelihood can be adapted for toroidal diffusions with minor changes. The Euler scheme arises as the first order discretization of the process, where the drift and diffusion coefficient are approximated constantly. After wrapping, the scheme becomes

$$\boldsymbol{\Theta}_{\Delta i} = \mathrm{cmod}\big(\boldsymbol{\Theta}_{\Delta(i-1)} + b(\boldsymbol{\Theta}_{\Delta(i-1)})\Delta$$
$$+ \sqrt{\Delta}\sigma(\boldsymbol{\Theta}_{\Delta(i-1)})\mathbf{Z}^i\big),$$

where $\mathbf{Z}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $i = 1, \ldots, N$. The wrapping yields the Euler pseudo-tpd

$$p_\Delta^{\mathrm{E}}(\boldsymbol{\theta} \mid \boldsymbol{\varphi}) := f_{\mathrm{WN}}\left(\boldsymbol{\theta}; \boldsymbol{\varphi} + b(\boldsymbol{\varphi})\Delta, V(\boldsymbol{\varphi})\Delta\right), \quad \boldsymbol{\theta}, \boldsymbol{\varphi} \in \mathbb{T}^p.$$

When $\Delta \to \infty$, the Euler pseudo-tpd converges to the uniform distribution in $\mathbb{T}^p$ by spreading its probability mass whilst the mean moves along the wrapped line $\{\mathrm{cmod}\,(\boldsymbol{\varphi} + b(\boldsymbol{\varphi})\Delta) : \Delta > 0\}$. The Euler pseudo-likelihood is obtained from (10) by replacing the tpd by the Euler pseudo-tpd.

The Shoji-Ozaki (Shoji and Ozaki, 1998) scheme uses a linear approximation for the drift and assumes the diffusion coefficient constant between observation times: for $t \in [s, s+\Delta)$, $b(\mathbf{X}_t) \approx b(\mathbf{X}_s) + \mathbf{J}_s(\mathbf{X}_t - \mathbf{X}_s)$, where $\mathbf{J}_s = J(\mathbf{X}_s)$ denotes the Jacobian of b at $\mathbf{X}_s$. This gives the linear approximating SDE

$$\mathrm{d}\mathbf{X}_t = (b(\mathbf{X}_s)+\mathbf{J}_s(\mathbf{X}_t-\mathbf{X}_s))\mathrm{d}t+\boldsymbol{\sigma}_s\mathrm{d}\mathbf{W}_t, \quad t \in [s, s+\Delta).$$

Conditionally on $\mathbf{X}_s$, this is a multivariate OU process. Hence, $\mathbf{X}_t \mid \mathbf{X}_s \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Gamma}_t)$, with $\boldsymbol{\mu}_t := \mathbf{J}_s^{-1}(\exp\{\mathbf{J}_s(t-s)\} - \mathbf{I})b(\mathbf{X}_s)$, $\boldsymbol{\Gamma}_t := \int_s^t \exp\{\mathbf{J}_s(t-u)\}\mathbf{V}_s \exp\{\mathbf{J}_s'(t-u)\}\mathrm{d}u$, and $\mathbf{V}_s = \boldsymbol{\sigma}_s\boldsymbol{\sigma}_s'$. If $\mathbf{J}_s$ has no pair of reverse-sign eigenvalues, then

$$\mathrm{vec}\left(\boldsymbol{\Gamma}_t\right) = (\mathbf{I} \otimes \mathbf{J}_s + \mathbf{J}_s \otimes \mathbf{I})^{-1}\mathbf{v}_t, \tag{13}$$
$$\mathbf{v}_t := \exp\{\mathbf{J}_s(t-s)\}\mathbf{V}_s \exp\{\mathbf{J}_s'(t-s)\} - \mathbf{V}_s.$$

If $\mathbf{V}_s^{-1}\mathbf{J}_s$ is symmetric, then $\boldsymbol{\Gamma}_t$ admits a more explicit form[2]:

$$\boldsymbol{\Gamma}_t = \int_s^t \exp\{\mathbf{J}_s(t-u)\}\mathbf{V}_s\mathbf{V}_s^{-1}$$
$$\cdot \exp\{\mathbf{V}_s\mathbf{J}_s'\mathbf{V}_s^{-1}(t-u)\}\mathbf{V}_s\mathrm{d}u$$
$$= \frac{1}{2}\mathbf{J}_s^{-1}(\exp\{2\mathbf{J}_s(t-s)\} - \mathbf{I})\mathbf{V}_s. \tag{14}$$

Interestingly, for the Langevin family of diffusions, $\mathbf{V}_s^{-1}\mathbf{J}_s$ is guaranteed to be symmetric as long as the diffusion

---

[2] Note the similar argument given in Roberts and Stramer (2002), albeit in their equation (24) the covariance matrix is not symmetric, probably because of a typo in (25), which should have been $(J(x)a_{x,h})' = J(x)a_{x,h}$.

coefficient is constant. This is due to the particular form of (5), which gives $\mathbf{J}_s = \frac{1}{2}\mathbf{V}_s\boldsymbol{\mathcal{H}}_s$, where $\boldsymbol{\mathcal{H}}_s$ stands for the Hessian of $\log f$ at $\mathbf{X}_s$. Therefore, (14) simplifies notably the evaluation of the Shoji-Ozaki pseudo-likelihood for all the toroidal diffusions considered in this paper.

The Shoji-Ozaki pseudo-tpd for toroidal diffusions is[3]

$$p_\Delta^{\mathrm{SO}}(\boldsymbol{\theta}\,|\,\boldsymbol{\varphi}) := f_{\mathrm{WN}}\left(\boldsymbol{\theta}; E_\Delta(\boldsymbol{\varphi}), V_\Delta(\boldsymbol{\varphi})\right), \quad \boldsymbol{\theta},\,\boldsymbol{\varphi} \in \mathbb{T}^p,$$

where, assuming that $V(\boldsymbol{\varphi})^{-1}J(\boldsymbol{\varphi})$ is symmetric (otherwise use (13) instead of (14)),

$$E_\Delta(\boldsymbol{\varphi}) = \boldsymbol{\varphi} + J(\boldsymbol{\varphi})^{-1}(\exp\{J(\boldsymbol{\varphi})\Delta\} - \mathbf{I})b(\boldsymbol{\varphi}),$$
$$V_\Delta(\boldsymbol{\varphi}) = \frac{1}{2}J(\boldsymbol{\varphi})^{-1}(\exp\{2J(\boldsymbol{\varphi})\Delta\} - \mathbf{I})V(\boldsymbol{\varphi}).$$

When $J(\boldsymbol{\varphi})$ shrinks to $\mathbf{0}$, then $E_\Delta(\boldsymbol{\varphi}) \approx \boldsymbol{\varphi} + b(\boldsymbol{\varphi})\Delta$ and $V_\Delta(\boldsymbol{\varphi}) \approx V(\boldsymbol{\varphi})\Delta$, so the Euler scheme follows by continuity. If all the real parts of the eigenvalues of $J(\boldsymbol{\varphi})$ are negative, then $p_\Delta^{\mathrm{SO}}(\boldsymbol{\theta}\,|\,\boldsymbol{\varphi}) \underset{\Delta\to\infty}{\longrightarrow} f_{\mathrm{WN}}\left(\boldsymbol{\theta}; \boldsymbol{\varphi} - J(\boldsymbol{\varphi})b(\boldsymbol{\varphi}), -\frac{1}{2}J(\boldsymbol{\varphi})^{-1}V(\boldsymbol{\varphi})\right)$ and the pseudo-tpd does not degenerate into a uniform density as Euler's does. Otherwise, the pseudo-tpd converges to the uniform distribution in $\mathbb{T}^p$ exponentially fast (see Figure 3), at a rate controlled by the maximum positive real part of the eigenvalues.

A disadvantage of these pseudo-likelihoods is that they are unimodal, so they cannot capture the multimodality of the tpd, a distinctive feature of toroidal diffusions.

Remark 3 *Evaluating $f_{\mathrm{WN}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the above pseudo-tpds is a computationally demanding task. Several approximations are possible:*

i. **High-concentration.** *Use the closest winding number as a one-term truncation of the series, i.e., $\phi_{\boldsymbol{\Sigma}}(\mathrm{cmod}\,(\cdot - \boldsymbol{\mu}))$.*

ii. **Fixed truncation.** *Mardia and Jupp (2000, page 50) suggests (for $p = 1$) $\sum_{\mathbf{k}\in\{-1,0,1\}^p} \phi_{\boldsymbol{\Sigma}}(\cdot - \boldsymbol{\mu} + 2\mathbf{k}\pi)$, which is usually enough for practical purposes if the argument lays in $\mathbb{T}^p$.*

iii. **Von Mises moment matching.** *Uses the approximation $\mathrm{WN}(\mu, \sigma^2) \approx \mathrm{vM}(\mu, A_1^{-1}(e^{-\sigma^2/2}))$, with*

$A_1(\kappa) = \mathcal{I}_1(\kappa)/\mathcal{I}_0(\kappa)$ *(Mardia and Jupp, 2000, page 38). This approximation generalizes easily to the multivariate case only if $\boldsymbol{\Sigma}$ is diagonal. For the bivariate case, an alternative is to use a von Mises score matching (Mardia, 2017).*

iv. **Adaptive truncation.** *The Jona-Lasinio et al (2012)'s "3σ adaptive truncation" can be generalized to the multivariate case by Bonferroni: $\sum_{\mathbf{k}=\mathbf{k}_L}^{\mathbf{k}_U} \phi_{\boldsymbol{\Sigma}}(\cdot - \boldsymbol{\mu} + 2\mathbf{k}\pi)$ with $\mathbf{k}_U = -\mathbf{k}_L = 1 + \lfloor z_{1-\alpha/(2p)}\sqrt{\mathrm{diag}\,(\boldsymbol{\Sigma})}$ $/(2\pi)\rfloor$, where z_α is the upper α-quantile of a $\mathcal{N}(0,1)$, ensures a probability mass in $\mathbb{T}^p$ larger than $1 - \alpha$.*

For $p = 1, 2$, a simple compromise between tractability and accuracy is combining i and ii into $\sum_{\mathbf{k}\in\{-1,0,1\}^p}$ $\phi_{\boldsymbol{\Sigma}}(\mathrm{cmod}\,(\cdot - \boldsymbol{\mu}) + 2\mathbf{k}\pi)$, which has a probability coverage of $\mathbb{T}^p$ larger than $1 - 2\sum_{j=1}^p \Phi(-\frac{3\pi}{\sigma_j})$.

3.3 Wrapped Ornstein-Uhlenbeck approximation of the WN process

We now present a specific approximation for the tpd of the WN process that allows to model the multimodality in the tpd. Multimodality is not uncommon for toroidal diffusions since each coordinate can move towards its mean value in two directions and, contrary to what happens with the OU process, this implies that neither the WN nor the MvM processes have tpds within the parametric families of the sdis.

The approximation relies on the connection of the WN process with the tractable multivariate OU process:

$$\mathrm{d}\mathbf{X}_t = \mathbf{A}(\boldsymbol{\mu} - \mathbf{X}_t)\mathrm{d}t + \boldsymbol{\Sigma}^{\frac{1}{2}}\mathrm{d}\mathbf{W}_t, \tag{15}$$

with $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma}$ a covariance matrix, and $\mathbf{A}$ such that $\mathbf{A}^{-1}\boldsymbol{\Sigma}$ is a covariance matrix. The last assumption ensures that the OU process is ergodic and time-reversible, and as a consequence, implies a simple expression for the covariance matrix of the tpd (see below). Under this setting, the process is ergodic, time-reversible, and has stationary density $\mathcal{N}(\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$. We denote by WOU, standing for Wrapped multivariate OU process, to the wrapping of (15). Assuming that $\mathbf{X}_s \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$, the conditional density of WOU is given by Proposition 1 and the tpd of (15):

$$p_t^{\mathrm{WOU}}(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_s) := \sum_{\mathbf{m}\in\mathbb{Z}^p} f_{\mathrm{WN}}(\boldsymbol{\theta}; \boldsymbol{\mu}_t^{\mathbf{m}}, \boldsymbol{\Gamma}_t)w_{\mathbf{m}}(\boldsymbol{\theta}_s), \tag{16}$$

$$w_{\mathbf{m}}(\mathbf{x}) = \frac{\phi_{\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma}}(\mathbf{x} - \boldsymbol{\mu} + 2\mathbf{m}\pi)}{\sum_{\mathbf{k}\in\mathbb{Z}^p} \phi_{\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma}}(\mathbf{x} - \boldsymbol{\mu} + 2\mathbf{k}\pi)}$$

---

[3] In Shoji and Ozaki (1998) the drift approximation is done by Itô's formula. To obtain a simpler pseudo-likelihood, we use a local linear approximation of b as in Ozaki (1985) (for the case $p = 1$). Without this extra simplification, the expectation becomes $\tilde{E}_\Delta(\boldsymbol{\varphi}) = E_\Delta(\boldsymbol{\varphi}) + J(\boldsymbol{\varphi})^{-2}(\exp\{J(\boldsymbol{\varphi})\Delta\} - \mathbf{I} - J(\boldsymbol{\varphi})\Delta)M(\boldsymbol{\varphi})$ with $M(\boldsymbol{\varphi}) = \frac{1}{2}(\mathrm{tr}\,[\mathbf{V}(\boldsymbol{\varphi})\mathbf{H}_1(\boldsymbol{\varphi})], \ldots, \mathrm{tr}\,[\mathbf{V}(\boldsymbol{\varphi})\mathbf{H}_n(\boldsymbol{\varphi})])'$ and $\mathbf{H}_i(\boldsymbol{\varphi}) = \left(\frac{\partial^2 b_i(\boldsymbol{\varphi})}{\partial\phi_k\partial\phi_l}\right)_{1\le k,l\le p}$, $i = 1, \ldots, p$.
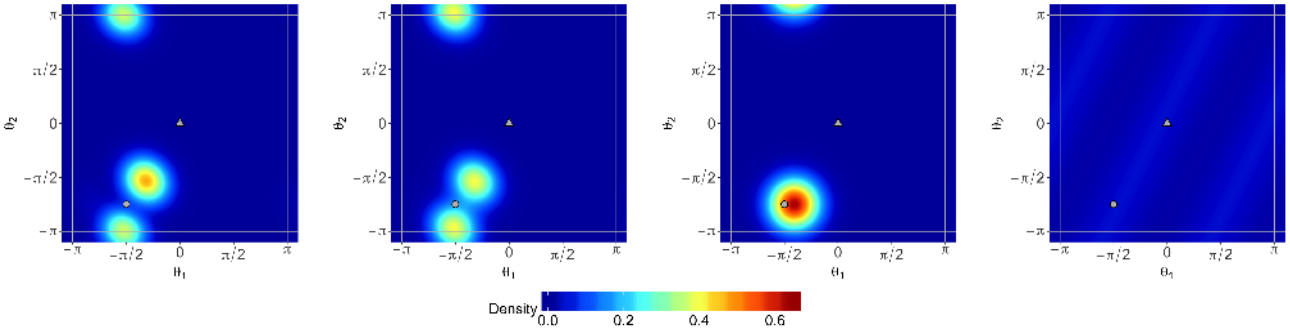
Fig. 3 The four approximations to the tpd $p_t(\cdot \,|\, \boldsymbol{\theta}_0)$ for the two-dimensional WN process, with $\boldsymbol{\theta}_0 = (-\frac{\pi}{2}, -\frac{3\pi}{4})$ (round facet), and $t = 0.25$. From left to right: PDE solution with $\sigma_0 = 0.01$, WOU tpd approximation, Euler pseudo-tpd, and Shoji-Ozaki pseudo-tpd. The WN process has parameters $\boldsymbol{\alpha} = (1, 1, 0.5)$, $\boldsymbol{\mu} = (0, 0)$ (triangular facet), and $\boldsymbol{\Sigma} = \mathrm{diag}(1, 1)$.

where, by the same argument used in (14),

$$\boldsymbol{\mu}_t^{\mathbf{m}} := \boldsymbol{\mu} + e^{-t\mathbf{A}}(\boldsymbol{\theta}_s - \boldsymbol{\mu} + 2\mathbf{m}\pi),$$

$$\boldsymbol{\Gamma}_t = \int_0^t e^{-s\mathbf{A}} \boldsymbol{\Sigma} e^{-s\mathbf{A}'} \mathrm{d}s = \frac{1}{2}\mathbf{A}^{-1}(\mathbf{I} - \exp\{-2t\mathbf{A}\})\boldsymbol{\Sigma}.$$

The conditional density (16) can be seen as a wrapping of the tpd of (15) weighted by the sdi of the winding numbers, which resembles the structure of the WN process drift: a weighting of linear drifts like (15) according to the winding number sdi in order to achieve periodicity. Albeit (16) and the tpd of the WN process are different, they behave similarly in many situations. The next corollary from Proposition 1 formalizes these arguments.

Corollary 1 Suppose $\boldsymbol{\Theta}_t$ solves (7) with $\boldsymbol{\Theta}_0 = \boldsymbol{\theta}_0$, and let $\boldsymbol{\Theta}_t^{\mathrm{WOU}}$ be the wrapping of the solution to (15), where $\mathbf{X}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$. We condition, moreover, on $\boldsymbol{\Theta}_0^{\mathrm{WOU}} = \boldsymbol{\theta}_0$. Then:

i. As $t \to 0$, $\boldsymbol{\Theta}_t \to \boldsymbol{\theta}_0$ and $\boldsymbol{\Theta}_t^{\mathrm{WOU}} \to \boldsymbol{\theta}_0$ in probability.
ii. As $t \to \infty$, both $\boldsymbol{\Theta}_t$ and $\boldsymbol{\Theta}_t^{\mathrm{WOU}}$ converge in distribution to a $\mathrm{WN}(\boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$.
iii. When $\mathbf{A}^{-1}\boldsymbol{\Sigma} \to \mathbf{0}$ with $\boldsymbol{\Sigma}$ bounded, $\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_t^{\mathrm{WOU}} \to \mathbf{0}$ in probability, so the distributions of $\boldsymbol{\Theta}_t$ and $\boldsymbol{\Theta}_t^{\mathrm{WOU}}$ are similar in the limit.
iv. p_t^{WOU} satisfies $f_{\mathrm{WN}}(\boldsymbol{\theta}_0; \boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})p_t^{\mathrm{WOU}}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0) = f_{\mathrm{WN}}(\boldsymbol{\theta}; \boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})p_t^{\mathrm{WOU}}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{\theta}), \forall \boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \mathbb{T}^p$ (just like p_t).

Proof The first two statements for the WN process are well-known for any diffusion, and it follows from (16) that, for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $\lim_{t \to 0} p_t^{\mathrm{WOU}}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0) = 0$ and that $\lim_{t \to \infty} p_t^{\mathrm{WOU}}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0) = f_{\mathrm{WN}}(\boldsymbol{\theta}; \boldsymbol{\mu}, \frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma})$. The last statement follows from (16) and the fact that the OU process is time-reversible when $\mathbf{A}^{-1}\boldsymbol{\Sigma}$ is positive definite. We give a rough sketch of a proof of the third statement. The result follows because the tpd of the WN process is asymptotically equal to the tpd of the OU process in the high concentration limit. To see this,

suppose that $\mathbf{X}_t$ solves (15) with $\mathbf{X}_0 = \boldsymbol{\theta}_0$ (we can ignore the other starting points), and that $\boldsymbol{\Theta}_t$ solves (7) with $\boldsymbol{\Theta}_0 = \boldsymbol{\theta}_0$, both driven by the same Wiener process. Then $\mathbf{Y}_t := \mathbf{X}_t - \boldsymbol{\Theta}_t$ solves $\mathrm{d}\mathbf{Y}_t = -\mathbf{A}\mathbf{Y}_t\mathrm{d}t + \mathrm{d}\mathbf{Z}_t$, with $\mathbf{Y}_0 = \mathbf{0}$, where

$$\mathbf{Z}_t = -\mathbf{A} \sum_{\mathbf{k} \in \mathbb{Z}^p \setminus \{\mathbf{0}\}} \int_0^t (\boldsymbol{\mu} - \boldsymbol{\Theta}_s - 2\mathbf{k}\pi) w_{\mathbf{k}}(\boldsymbol{\Theta}_s)\mathrm{d}s$$

$$+ \mathbf{A} \int_0^t (\boldsymbol{\mu} - \boldsymbol{\Theta}_s)(1 - w_{\mathbf{0}}(\boldsymbol{\Theta}_s))\mathrm{d}s.$$

If $\mathbf{Y}_t = \int_0^t e^{-\mathbf{A}(t-s)}\mathrm{d}\mathbf{Z}_s \to \mathbf{0}$ in probability, then the two distributions of $\boldsymbol{\Theta}_t$ and $\mathbf{X}_t$ will be the same in the limit. This follows because $\mathbf{A}w_{\mathbf{k}}(\boldsymbol{\theta}) \to \mathbf{0}$ and $\mathbf{A}(1 - w_{\mathbf{0}}(\boldsymbol{\theta})) \to \mathbf{0}$ for $\mathbf{k} \in \mathbb{Z}^p \setminus \{\mathbf{0}\}$ and $-\pi < \boldsymbol{\theta} < \pi$ (we consider only the case $\mathrm{wind}(\boldsymbol{\mu} - \boldsymbol{\theta}) = \mathbf{0}$ because $\mathbb{P}[|\boldsymbol{\Theta}_s - \boldsymbol{\mu}| \leq \boldsymbol{\pi}] \to 1$), and hence $\mathbf{Z}_s \to \mathbf{0}$ in probability for all $s \leq t$. $\square$

The tractability of (16) degenerates quickly with the dimension, but it can be readily computed for $p = 1, 2$, two highly relevant situations in practice. We focus our attention on implementation matters for the nontrivial case $p = 2$. The first point of inquiry is what parametrization of $\mathbf{A}$ and $\boldsymbol{\Sigma}$ leads to a covariance matrix $\mathbf{A}^{-1}\boldsymbol{\Sigma}$, which guarantees a non-degenerate WN sdi.

Lemma 1 *Let $\mathbf{A}$ and $\boldsymbol{\Sigma}$ be 2×2 matrices, $\boldsymbol{\Sigma} = (\sigma_1^2, \rho\sigma_1\sigma_2; \rho\sigma_1\sigma_2, \sigma_2^2)$ positive-definite. Assume $\alpha_1, \alpha_2 > 0$. Any matrix $\mathbf{A}$ such that $\mathbf{A}^{-1}\boldsymbol{\Sigma}$ is a covariance matrix has the form*

$$\mathbf{A} = \begin{pmatrix} \alpha_1 & \frac{\sigma_1}{\sigma_2}\left(\alpha_3 + \frac{1}{2}\rho(\alpha_2 - \alpha_1)\right) \\ \frac{\sigma_2}{\sigma_1}\left(\alpha_3 - \frac{1}{2}\rho(\alpha_2 - \alpha_1)\right) & \alpha_2 \end{pmatrix},$$

with $\alpha_3^2 < \frac{\rho^2(\alpha_1 - \alpha_2)^2}{4} + \alpha_1\alpha_2$.

The parametrization with $\rho = 0$ provides a compromise between flexibility and simplicity, and will be employed throughout (first occurrences in Figures 3 and 5). With

$\rho = 0$ the dependence between components is modelled by α_3, which is clear from

$$\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma} = \frac{1}{2(\alpha_1\alpha_2 - \alpha_3^2)} \begin{pmatrix} \alpha_2\sigma_1^2 & -\alpha_3\sigma_1\sigma_2 \\ -\alpha_3\sigma_1\sigma_2 & \alpha_1\sigma_2^2 \end{pmatrix}.$$

The second point is the efficient computation of $e^{-t\mathbf{A}}$ and $\boldsymbol{\Gamma}_t$. In virtue of Corollary 2.4 of Bernstein and So (1993), $e^{t\mathbf{A}} = a(t)\mathbf{I} + b(t)\mathbf{A}$ with $b(t) := e^{s(\mathbf{A})t}\frac{\sinh(q(\mathbf{A})t)}{q(\mathbf{A})}$ (if $q(\mathbf{A}) = 0$, then, by continuity, $b(t) = e^{s(\mathbf{A})t}t$), $a(t) := e^{s(\mathbf{A})t}\cosh(q(\mathbf{A})t) - s(\mathbf{A})b(t)$, $s(\mathbf{A}) := \frac{\text{tr}[\mathbf{A}]}{2}$, and $q(\mathbf{A}) := \sqrt{|\det(\mathbf{A} - s(\mathbf{A})\mathbf{I})|}$. Therefore,

$$\boldsymbol{\Gamma}_t = s(t)\frac{1}{2}\mathbf{A}^{-1}\boldsymbol{\Sigma} + i(t)\boldsymbol{\Sigma}, \tag{17}$$

with $s(t) = 1 - a(-2t)$ and $i(t) = -\frac{1}{2}b(-2t)$. Expression (17) shows neatly the interpolation between the infinitesimal and stationary covariance matrices and is especially useful if it is required to compute the tpd for several t's.

To conclude, we highlight some of the advantages of the WOU approximation over the Euler and Shoji-Ozaki pseudo-likelihoods for the WN process. Firstly, WOU is able to capture the multimodality of the tpd (see Figure 3) and has the correct sdi. Secondly, WOU is faster to compute than Shoji-Ozaki, as it does not require exponentiation and inversion of the Jacobian matrix for each observation, but only once.

3.4 Likelihood by numerical PDE solution

An alternative to approximate likelihoods is to compute the "exact" (up to a prescribed accuracy) MLE by a numerical solution of (3). This approach is computationally expensive, but remains valid for arbitrary diffusions and discretization times. Moreover, it provides insightful visualizations of the tpd. In the following, we discuss how to solve numerically (3) for dimensions $p = 1, 2$.

3.4.1 One-dimensional case

We consider a state grid $\mathcal{G} := \{x_1, \ldots, x_{M_x}\}$ in $[-\pi, \pi)$ constructed with step $\Delta x := \frac{2\pi}{M_x}$, and such that $x_{M_x+1} := x_1 = -\pi$ and $x_0 := x_{M_x} = \pi - \Delta x$. We also consider a time grid in $[0, T]$ with $\Delta t := \frac{T}{M_t}$. For consistency with the common notation for PDEs, we refer by $u(x, t)$ to $p_t(x \,|\, p_s) := \int_{\mathbb{T}^1} p_t(x \,|\, \phi)p_s(\phi)\mathrm{d}\phi$, the solution of the PDE for the initial condition (at time s) given by a circular density p_s. The vector $\mathbf{u}^n$, $n = 0, \ldots, M_t$, denotes the tpd evaluated at $\mathcal{G}$ at time $s + n\Delta t$. We write $b_i := b(x_i)$ and $\sigma_i^2 := \sigma^2(x_i)$, $i = 1, \ldots, M_x$.

We employ the so-called *Crank-Nicolson scheme* for discretizing (3), which can be rewritten as

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x}(b(x)u(x, t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2(x)u(x, t))$$
$$= F\left(u, x, t, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}\right).$$

Crank-Nicolson is a well-known scheme for diffusion and convection-diffusion PDEs such as (3). It is based on a trapezoidal-like approximation of the forward difference of the time derivative that is combined with a centered finite differences of the state derivatives:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{1}{2}\left[F_i^{n+1} + F_i^n\right],$$
$$F_i^n := \frac{2}{\Delta t}\left(\gamma_i u_{i+1}^n - 2\beta_i u_i^n + \alpha_i u_{i-1}^n\right),$$

with $r := \frac{\Delta t}{4(\Delta x)^2}$, $\gamma_i := \left(-b_{i+1}\Delta x + \sigma_{i+1}^2\right)r$, $\beta_i := \sigma_i^2 r$, and $\alpha_i := \left(b_{i-1}\Delta x + \sigma_{i-1}^2\right)r$. The next step in time of the solution, $\mathbf{u}^{n+1}$, is implicitly given by the system

$$(\mathbf{F} - \mathbf{I})\mathbf{u}^{n+1} = \mathbf{d}^n, \quad \mathbf{F} := \begin{pmatrix} -2\beta_1 & \gamma_1 & & & \alpha_1 \\ \alpha_2 & -2\beta_2 & \ddots & & \\ & \ddots & \ddots & \gamma_{M_x-1} & \\ \gamma_{M_x} & & & \alpha_{M_x} & -2\beta_{M_x} \end{pmatrix},$$
$$\mathbf{d}^n := -(\mathbf{F} + \mathbf{I})\mathbf{u}^n$$
$$= -\boldsymbol{\gamma} \circ \mathbf{u}_+^n + (2\boldsymbol{\beta} - 1) \circ \mathbf{u}^n - \boldsymbol{\alpha} \circ \mathbf{u}_-^n, \tag{18}$$

with subscript $\pm$ denoting the vector with entries circularly shifted ∓ 1 position. It is well-known (Thomas, 1995, page 225) that this periodic tridiagonal system can be solved efficiently by tacking the tridiagonal systems $\mathbf{B}\mathbf{y}_1 = \mathbf{d}_n$ and $\mathbf{B}\mathbf{y}_2 = \mathbf{w}$ (where $\mathbf{F} = \mathbf{B} - \mathbf{w}\mathbf{z}'$), and using the Sherman-Morrison formula: $\mathbf{u}^{n+1} = \mathbf{y}_1 + \frac{\mathbf{z}'\mathbf{y}_1}{1-\mathbf{z}'\mathbf{y}_2}\mathbf{y}_2$. The latter tridiagonal systems can be jointly solved by a modification of the Thomas algorithm, since they share coefficient matrix. The cost of the solution is $\mathcal{O}(M_t M_x)$. In addition, since $\mathbf{F}$ is constant with respect to time, the *tridiagonal* LU factorization underlying the Thomas algorithm can be reused, yielding a complexity factor reduction of $5/8$ on the tridiagonal solver.

3.4.2 Two-dimensional case

We consider now two grids $\mathcal{G}_x$ and $\mathcal{G}_y$ analogous to $\mathcal{G}$, but of sizes M_x and M_y, and steps Δx and Δy. We refer by $u(x, y, t)$ to $p_t(x, y \,|\, p_s) := \int_{\mathbb{T}^2} p_t(x, y \,|\, \boldsymbol{\varphi})p_s(\boldsymbol{\varphi})\mathrm{d}\boldsymbol{\varphi}$, where p_s is a toroidal density giving the initial condition (at time s). The matrix $\mathbf{U}^n$, $n = 0, \ldots, M_t$, denotes the tpd evaluated at $\mathcal{G}_x \times \mathcal{G}_y$ at time $s + n\Delta t$. We write $b_{z;i,j} := b_z(x_i, y_j)$, $\sigma_{z;i,j}^2 := \sigma_z^2(x_i, y_j)$, $\sigma_{xy;i,j}^2 :=$

$\sigma_{xy}^2(x_i, y_j)$, with z standing for x or y, and $i = 1, \ldots, M_x$, $j = 1, \ldots, M_y$. With this notation, (3) becomes

$$\frac{\partial u}{\partial t} = \sum_{z \in \{x,y\}} \Bigl[-\frac{\partial}{\partial z}(b_z(x,y)u(x,y,t))$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial z^2}(\sigma_z^2(x,y)u(x,y,t)) \Bigr]$$
$$+ \frac{\partial^2}{\partial x \partial y}(\sigma_{xy}(x,y)u(x,y,t))$$

The Crank-Nicolson scheme proceeds as in the one-dimensional case:

$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t} = \frac{1}{2}\left[F_{i,j}^{n+1} + F_{i,j}^n \right], \tag{19}$$

with finite differences that can be collected into three terms associated to the partial and mixed derivatives:

$$F_{i,j}^n := F_{x;i,j}^n + F_{y;i,j}^n + F_{xy;i,j}^n,$$
$$F_{x;i,j}^n := \frac{2}{\Delta t}\left(\Gamma_{x;i,j} u_{i+1,j}^n - 2B_{x;i,j} u_{i,j}^n + A_{x;i,j} u_{i-1,j}^n \right),$$
$$F_{y;i,j}^n := \frac{2}{\Delta t}\left(\Gamma_{y;i,j} u_{i,j+1}^n - 2B_{y;i,j} u_{i,j}^n + A_{y;i,j} u_{i,j-1}^n \right),$$
$$F_{xy;i,j}^n := \frac{2}{\Delta t}\bigl(C_{i,j}^{+,+} u_{i+1,j+1}^n - C_{i,j}^{+,-} u_{i+1,j-1}^n$$
$$- C_{i,j}^{-,+} u_{i-1,j+1}^n + C_{i,j}^{-,-} u_{i-1,j-1}^n \bigr).$$

We have denoted $r_z := \frac{\Delta t}{4(\Delta z)^2}$, $r_{xy} := \frac{\Delta t}{8\Delta x \Delta y}$, and

$$\Gamma_{x;i,j} := (-b_{x;i+1,j}\Delta x + \sigma_{x;i+1,j}^2)r_x,$$
$$\Gamma_{y;i,j} := (-b_{y;i,j+1}\Delta y + \sigma_{y;i,j+1}^2)r_y,$$
$$A_{x;i,j} := (b_{x;i-1,j}\Delta z + \sigma_{x;i-1,j}^2)r_x,$$
$$A_{y;i,j} := (b_{y;i,j-1}\Delta y + \sigma_{y;i,j-1}^2)r_y,$$
$$B_{z;i,j} := \sigma_{z;i,j}^2 r_z, \quad C_{i,j}^{\pm,\pm} := \sigma_{xy;i\pm1,j\pm1}r_{xy}.$$

Let $F = F_x + F_y + F_{xy}$ be the linear functions mapping $\mathbf{U}^n$ into $F(\mathbf{U}^n) = F_x(\mathbf{U}^n) + F_y(\mathbf{U}^n) + F_{xy}(\mathbf{U}^n) = (F_{x;i,j}^n + F_{y;i,j}^n + F_{xy;i,j}^n) = (F_{i,j}^n)$ and I the identity function. Then, we can express (19) as

$$(F_x + F_y + F_{xy} - I)(\mathbf{U}^{n+1}) = -(F + I)(\mathbf{U}^n). \tag{20}$$

If the left and right hand sides of (20) are stacked column-wise, (20) becomes an $M_x M_y \times M_x M_y$ periodic 9-diagonal system. This system cannot be solved so efficiently as in the tridiagonal case, requiring a more complex algorithm or a generic sparse LU factorization.

An alternative approach that reduces drastically the computational burden of solving (20) is to adopt an Alternating Direction Implicit (ADI) scheme. ADI schemes split the multidimensional finite differences in a series of univariate discretizations with simpler associated systems. Originally developed for the diffusion equation,

they were extended to the convection-diffusion equations with a mixed derivative term by McKee et al (1996), in the so-called *Douglas scheme*. This scheme proceeds with an explicit multivariate step corrected by two unidimensional Crank-Nicolson steps, whose purpose is to stabilize the explicit step:

$$\begin{cases} \mathbf{Y}_0 = \mathbf{U}^n + \Delta t F(\mathbf{U}^n) & \text{(explicit)} \quad (21) \\ \mathbf{Y}_1 - \frac{\Delta t}{2}F_x(\mathbf{Y}_1) = \mathbf{Y}_0 - \frac{\Delta t}{2}F_x(\mathbf{U}^n) & \text{(implicit)} \quad (22) \\ \mathbf{Y}_2 - \frac{\Delta t}{2}F_y(\mathbf{Y}_2) = \mathbf{Y}_1 - \frac{\Delta t}{2}F_y(\mathbf{U}^n) & \text{(implicit)} \quad (23) \\ \mathbf{U}^{n+1} = \mathbf{Y}_2 \end{cases}$$

Consequently, if the matrix equations in (21)–(23) are transformed into linear systems by column-wise stacking for (21) and (22), and row-wise stacking for (23), the Douglas scheme transforms the difficult task of solving (20) into solving two periodic tridiagonal systems of size $M_x M_y$. Specifically, the steps in (21)–(23) are carried out using

$$\text{vec}(F_{xy}(\mathbf{Y})) = \text{vec}(\mathbf{C}^{+,+}) \circ \mathbf{y}_{+,+,c}$$
$$- \text{vec}(\mathbf{C}^{+,-}) \circ \mathbf{y}_{+,-,c}$$
$$- \text{vec}(\mathbf{C}^{-,+}) \circ \mathbf{y}_{-,+,c}$$
$$+ \text{vec}(\mathbf{C}^{-,-}) \circ \mathbf{y}_{-,-,c},$$
$$\mathbf{y}_{\pm,\pm,c} := \text{vec}(\mathbf{Y}_{\pm,\pm})$$
$$\text{vec}(F_x(\mathbf{Y})) = \text{vec}(\mathbf{\Gamma}_x) \circ \mathbf{y}_{c,+} - 2\text{vec}(\mathbf{B}_x) \circ \mathbf{y}_c$$
$$+ \text{vec}(\mathbf{A}_x) \circ \mathbf{y}_{c,-},$$
$$\text{vec}(F_y(\mathbf{Y})') = \text{vec}(\mathbf{\Gamma}_y') \circ \mathbf{y}_{r,+} - 2\text{vec}(\mathbf{B}_y') \circ \mathbf{y}_r$$
$$+ \text{vec}(\mathbf{A}_y') \circ \mathbf{y}_{r,-},$$
$$\mathbf{y}_c := \text{vec}(\mathbf{Y}), \quad \mathbf{y}_r := \text{vec}(\mathbf{Y}').$$

$\mathbf{U}^{n+1}$ is obtained by setting $\mathbf{Y}$ equal to $\mathbf{U}^n$, $\mathbf{Y}_1$ or $\mathbf{Y}_2$ in the above expressions and by solving (22) and (23) as (18) was. Then, the total cost of the solution is $\mathcal{O}(M_t M_x M_y)$. Note that the row-stacking vector $\mathbf{y}_r^n$ can be directly obtained from $\mathbf{y}_c^n$ by extracting the indexes $((k_c - 1) \mod M_y)M_x + \lfloor \frac{k_c - 1}{M_y} \rfloor + 1$, $k_c = 1, \ldots, M_x M_y$, of the latter (analogous for the converse). We refer to the neat expository paper of In 't Hout and Foulon (2010) for further description of ADI schemes.

3.4.3 Remarks on the discretization schemes

The Crank-Nicolson and Douglas schemes are tailored solutions for solving (3) that exploit the particular PDE structure. It is worth to note that, among other methods, a well-known approach to solve PDEs is the *method of lines*. This method is prone to create stiff systems, which need to be handled properly by a meta-solver that chooses between stiff and non-stiff solvers (e.g., the `lsoda` implementation in Soetaert et al (2012)).

Not surprisingly, in our application we found that the efficiency and reliability of the tailored solutions were superior to the latter, much more general, meta-solver.

Some theoretical remarks about the schemes employed are given as follows. The Crank-Nicolson scheme is *conservative* (hence the Douglas scheme is too), which can be easily seen from the periodic tridiagonal system. It is also second-order consistent in time and space (with the discretization used), with the appealing property of being *unconditionally stable* with respect to Δt. The Douglas scheme is first-order consistent and unconditionally stable when applied to two-dimensional convection-diffusion equations with a mixed derivative term. See In 't Hout and Foulon (2010) for the description of second-order ADI schemes of the same computational complexity (but with a factor increase of at least two). Both unconditional stabilities refer to the usual framework of constant coefficients.

3.4.4 Likelihood evaluation

The PDE numerical solutions approximate $p_t(\boldsymbol{\theta} \,|\, p_s) = \int_{\mathbb{T}^p} p_t(\boldsymbol{\theta} \,|\, \boldsymbol{\varphi}) p_s(\boldsymbol{\varphi}) \mathrm{d}\boldsymbol{\varphi}$, where p_s is a density over $\mathbb{T}^p$ giving the initial condition. Therefore, $p_t(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0)$ can be approximated by considering a concentrated $\mathrm{WN}(\boldsymbol{\theta}_0, \sigma_0^2 \mathbf{I})$ as the initial condition. For a fixed grid, σ_0 must not be set to an arbitrarily small value, as it will create a sharp initial condition poorly discretized and prone to raise numerical errors. A possible rule of thumb is to choose a small σ_0 such that the periodic trapezoidal rule of the discretized $\mathrm{WN}(\boldsymbol{\theta}_0, \sigma_0^2 \mathbf{I})$ is close to one.

We illustrate the evaluation of the log-likelihood (10) from the PDE solution for $p = 1$. The extension to $p = 2$ is conceptually straightforward, albeit cumbersome in notation. Given the sample $\{\Theta_{\Delta i}\}_{i=1}^N$ and the grid $\mathcal{G}$, let denote by $\mathbf{P} := p_t(\mathcal{G} \,|\, \mathcal{G})$ the $M_x \times M_x$ *tpd matrix* of the process discretized in $\mathcal{G}$. The j-th column of $\mathbf{P}$ is obtained by solving the PDE with initial condition $\mathrm{WN}(x_j, \sigma_0^2)$. We can approximate $p_\Delta(\Theta_{i\Delta} \,|\, \Theta_{(i-1)\Delta})$ from $\mathbf{P}$ by linear interpolation:

$$p_\Delta(\Theta_{i\Delta} \,|\, \Theta_{(i-1)\Delta})$$
$$\approx \sum_{k,l=0}^{1} \omega_k(\Theta_{\Delta i}) P_{g_0(i)+k, g_0(i-1)+l} \omega_l(\Theta_{\Delta(i-1)}), \quad (24)$$

with $g_0(i) := \lceil \frac{\Theta_{\Delta i} + \pi}{\Delta x} \rceil$, $\omega_0(\theta) = \frac{x_{g_0(i)+1} - \theta}{\Delta x}$, and $\omega_1(\theta) = \frac{\theta - x_{g_0(i)}}{\Delta x}$. The log-likelihood is obtained by plugging (24) into (10). The advantage of doing so is that the number of PDE solutions required for a single log-likelihood evaluation remains bounded by M_x, irrespectively of N. In addition, we only need to compute the columns of $\mathbf{P}$

corresponding to the unique set of indexes $\{g_0(i) + l : i = 0, \ldots, N - 1, l = 0, 1\}$. A simpler, though less precise, alternative to (24) is to use constant interpolation for $\Theta_{\Delta(i-1)}$. This results in a lower number of PDE solutions, specially in the two-dimensional case. Finally, if the drift is antisymmetric around a point μ, then $p_t(\theta \,|\, \varphi) = p_t(2\mu - \theta \,|\, 2\mu - \varphi)$. Hence, if $\mathcal{G}$ is circularly centered at μ, half of the columns of $\mathbf{P}$ contain redundant information. The situation is analogous for $p = 2$: if $b(\theta_1 - \mu_1, \theta_2 - \mu_2) = -b(\mu_1 - \theta_1, \mu_2 - \theta_2)$, $\forall \theta_1, \theta_2 \in [-\pi, \pi)$, and $\mathcal{G}_x$ and $\mathcal{G}_y$ are both centered at μ_1 and μ_2, respectively, then only half of the columns of $\mathbf{P}$ are required. If the drift is isotropic, then only one fourth of the columns are needed.

4 Simulation study

We measure now the performance of the likelihood approximations given in Section 3. Two types of empirical analysis are employed. First, we compare the divergences between the true tpd of a diffusion and its approximations across time. Second, we examine the errors of the approximate likelihoods in estimating $\boldsymbol{\lambda}$ in several diffusions.

4.1 Kullback-Leibler divergences for WN and vM processes

All the estimation approaches described on Section 3 share a common root: the substitution of the true tpd p_t by an approximation p_t^{A}. The goodness-of-fit of these approximations has a direct influence on MLE since, for a general parametric setting, MLE is equivalent to minimizing the Kullback-Leibler divergence of the parametric pdf from the empirical pdf. We propose to measure the Kullback-Leibler divergence of $p_t^{\mathrm{A}}(\cdot \,|\, \boldsymbol{\theta}_s)$ from $p_t(\cdot \,|\, \boldsymbol{\theta}_s)$ by *weighting* with the stationary density the contributions of each initial point $\boldsymbol{\theta}_s$:

$$\mathrm{D}_t^{\mathrm{A}} = \int_{\mathbb{T}^p} \int_{\mathbb{T}^p} p_t(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_s) \log \left(\frac{p_t(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_s)}{p_t^{\mathrm{A}}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_s)} \right) \nu(\boldsymbol{\theta}_s) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\theta}_s.$$

The curve $\mathrm{D}_t^{\mathrm{A}}$ gives a succinct summary of the goodness-of-fit of any approximation to the tpd across time. Its effective computation – when no analytical expression for the tpd exists – can be done with the PDE solution to the tpd. Some care is needed though. The PDE solution involves the initial condition in the form of a concentrated $\mathrm{WN}(\boldsymbol{\theta}_0, \sigma_0^2 \mathbf{I})$. This initial condition implies that the PDE solution is approximating $p_{t,\sigma_0^2}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0) := \int_{\mathbb{T}^p} p_t(\boldsymbol{\theta} \,|\, \boldsymbol{\varphi}) f_{\mathrm{WN}}(\boldsymbol{\varphi}; \boldsymbol{\theta}_0, \sigma_0^2 \mathbf{I}) \mathrm{d}\boldsymbol{\varphi}$ rather than p_t. Therefore,
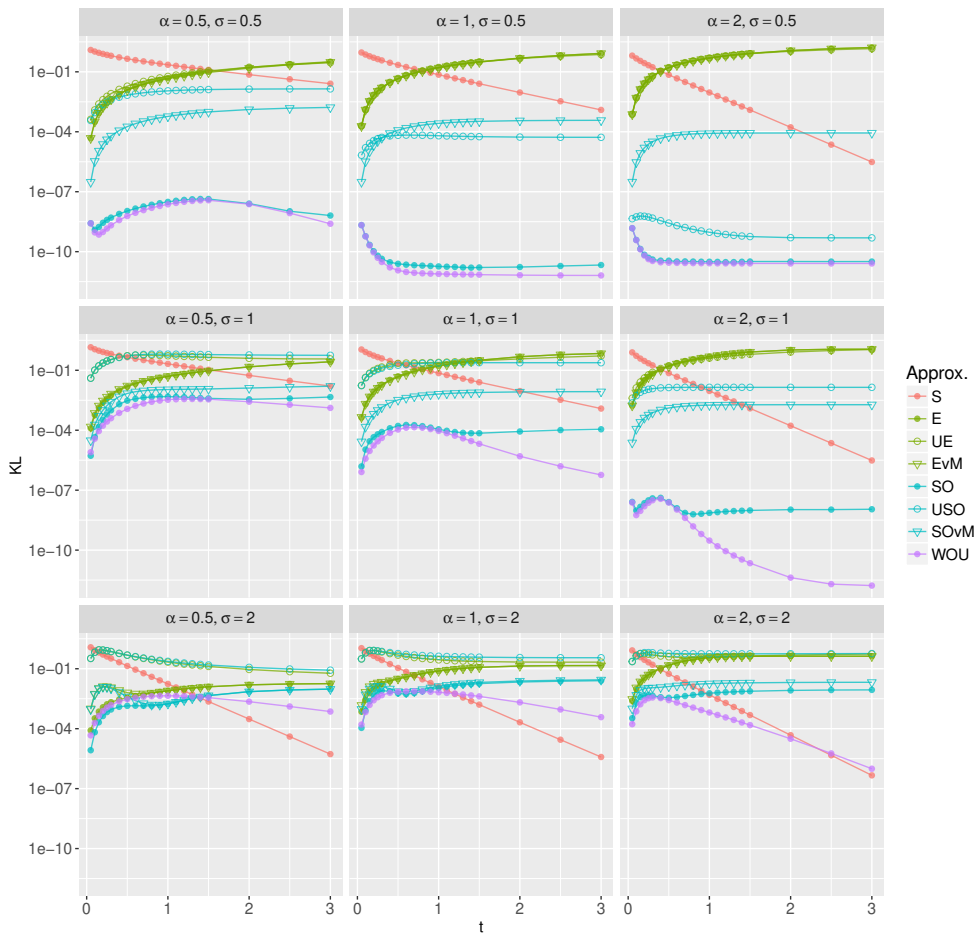
Fig. 4 $\mathrm{D}^{\mathrm{A}}_{t,\sigma_0^2}$ curves for $p = 1$ on the WN process, on vertical log-scale. From top to bottom, rows represent small, moderate, and high diffusivities. The PDE was solved with $M_x = 3000$, $M_t = \lceil 1500t \rceil$, and $\sigma_0 = 0.1$.

a more adequate approach is to smooth also the approximations in the computation of D_t to perform a fair comparison:

$$\mathrm{D}^{\mathrm{A}}_{t,\sigma_0^2} = \int_{\mathbb{T}^p} \int_{\mathbb{T}^p} u_{t,\sigma_0^2}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_s) \log \left(\frac{u_{t,\sigma_0^2}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_s)}{p^{\mathrm{A}}_{t,\sigma_0^2}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_s)} \right)$$
$$\times \nu(\boldsymbol{\theta}_s) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\theta}_s.$$

We explore the $\mathrm{D}^{\mathrm{A}}_{t,\sigma_0^2}$ curves for several variants of the approximations given in Section 3, denoted as S (Stationary density), E (Euler), SO (Shoji-Ozaki), UE (Unwrapped Euler – the usual Euler pseudo-likelihood), USO (Unwrapped Shoji-Ozaki), EvM, SOvM, and WOU. Suffix vM denotes the use of a vM distribution (*moment* if one-dimensional; *score* if two-dimensional) *matching* approximation to the WN distribution appearing in the pseudo-likelihoods.

Figures 4 and 5 show the Kullback-Leibler curves for the WN process with $p = 1$ and $p = 2$, under different drift strengths and diffusivities. We highlight as follows their main features. First, WOU outperforms in almost all scenarios and times the other approximations. The main exceptions are the lower left scenarios of both figures, representing processes with a high diffusivity (small drifts and large diffusivities), where WOU is outperformed by SO and E for a significant range of intermediate times. In addition to S, WOU is the only approximation whose accuracy improves as time increases, above a certain local maximum in the Kullback-Leibler divergence. Second, the Euler and Shoji-Ozaki pseudo-likelihoods deteriorate or stabilize as time increases, except for scenarios with low and moderate diffusivity where SO is close to WOU (and both are close to the true tpd). E is systematically behind SO in performance, usually by several orders of magnitude. S is, as expected, giving a poor performance unless t is large. Third, the wrapped versions of the pseudo-likelihoods dominate uniformly the unwrapped ones, both having
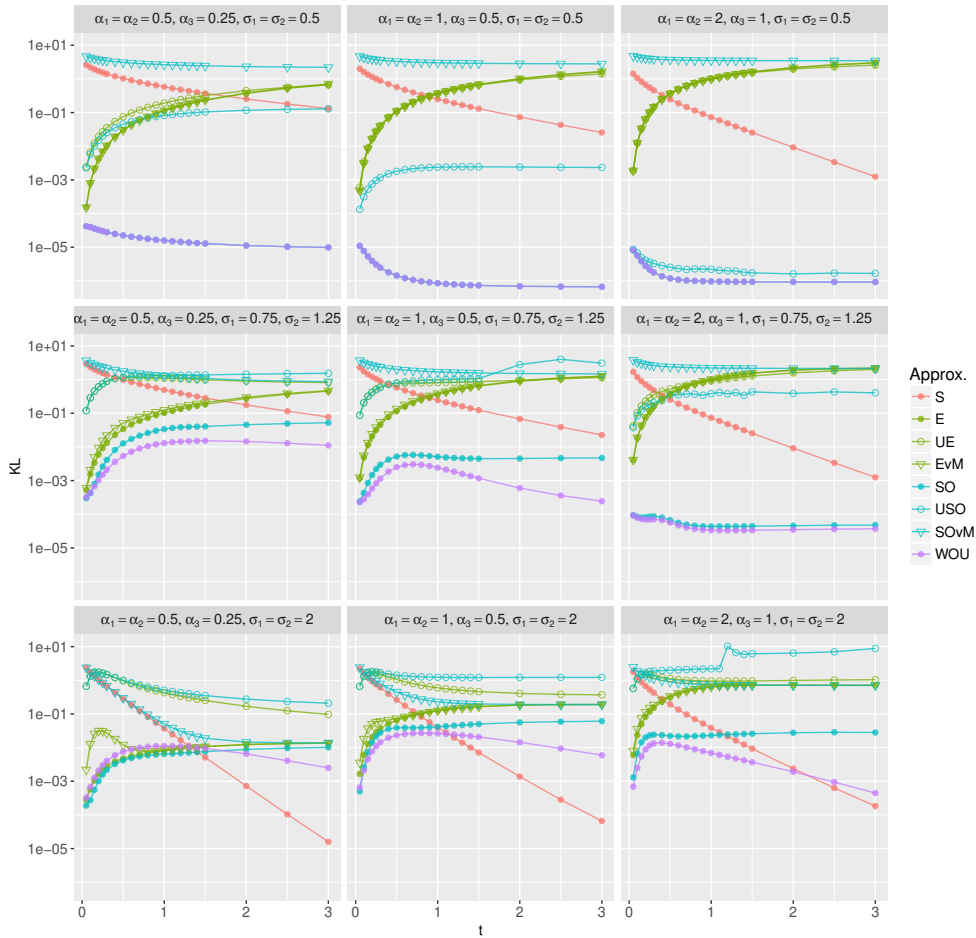
Fig. 5 $D_{t,\sigma_0^2}^A$ curves for $p = 2$ on the WN process. Note the vertical log-scale. From top to bottom, rows represent small, moderate, and high diffusivities. The PDE was solved with $M_x = M_y = 240$, $M_t = \lceil 1500t \rceil$, and $\sigma_0 = 0.1$.

similar performances if the process is highly concentrated. Indeed, the wrapping of SO is key in preventing the spread of probability mass outside $\mathbb{T}^p$ when the Jacobian of the drift has positive eigenvalues and t grows, which raises numerical instabilities (e.g., lower right panel of Figure 5). Finally, matching the WN distribution of E and SO by a vM has different effects depending on the method. For E, the results are similar for both E and EvM, except for a bump in small times with high diffusivities. However, SOvM consistently adds a high bias to SO, resulting in significant higher divergences. As a general advice, we recommend to approximate the tpd of the WN process by WOU, SO and E, in this order.

We reproduce the same experiment on the vM process, with results collected in Figures 6 and 7. The highlights are similar except for the following differences. First, the good properties that WOU has for the WN process do not hold any more, evincing its process-specificity. S is now the only approximation whose accuracy improves over time. Second, SO is systematically above E in performance, yet this difference is reduced as SO is not the true tpd under high-concentration. Third, the vM distribution match does not provide a better approximation to the tpd, despite the sdi being vM. EvM is again close to E and EvM except for small t's where EvM adds a substantial bias for scenarios with moderate and high diffusivities. The same happens for SOvM in $p = 1$, whereas for $p = 2$ SOvM increases the Kullback-Leibler divergence by several orders of magnitude when compared to SO in the scenarios with high diffusivity. Our general advice is to approximate the tpd by SO and E, in this order.
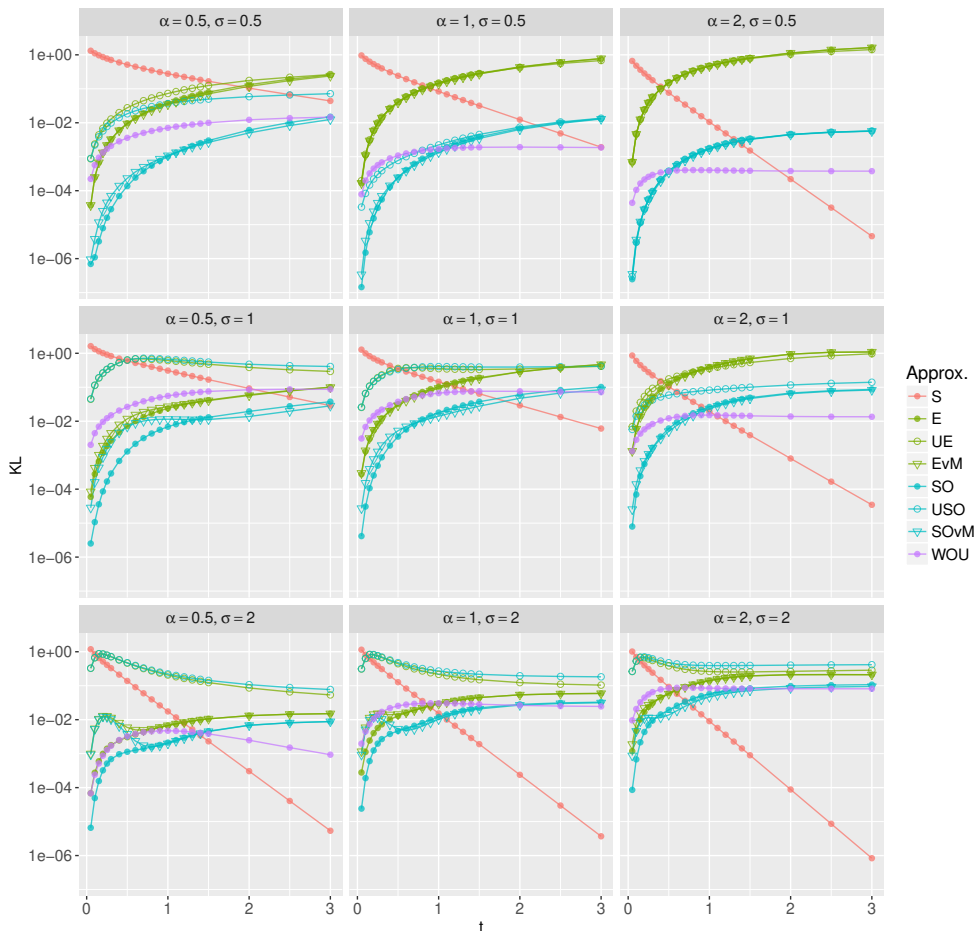
Fig. 6 $\mathrm{D}^{\mathrm{A}}_{t,\sigma_0^2}$ curves for $p = 1$ on the vM process. The description in Figure 4 applies.

4.2 Empirical performance of the approximate likelihoods

We compare now the efficiency of WOU, SO, and E – the best performing tpd approximations, according to the weighted Kullback-Leibler divergences – in estimating the unknown parameters of the diffusion (9) from a trajectory $\{\boldsymbol{\Theta}_{\Delta i}\}_{i=0}^N$. In this section, we set $N = 250$ and assume that $\sigma(\cdot; \boldsymbol{\lambda}) = \boldsymbol{\Sigma}^{\frac{1}{2}}$ is known in order to avoid the inherent unidentifiabilities of $\boldsymbol{\lambda}$ when Δ is large and the tpd converges to the sdi. We explore the behaviour of the estimators for dimensions $p = 1, 2$, time steps $\Delta = 0.05, 0.20, 0.50, 1.00$, and for representative parameter choices of the WN process and of two challenging diffusions. For $p = 1$, we also consider the PDE-based approximation to the likelihood. The trajectories are simulated using the E method with time step 0.001 and then subsampled for given Δ's.

In order to summarize the overall performance of a collection $\{\hat{\boldsymbol{\lambda}}_j = (\hat{\lambda}_{j,1}, \ldots, \hat{\lambda}_{j,K}) : j = 1 \ldots, J\}$ of K-variate estimators of $\boldsymbol{\lambda}$, we consider a global measure of relative performance. This measure is the componentwise average of Relative Efficiency (RE), where the relative efficiency is measured with respect to the best estimator at a given component in terms of Mean Squared Error (MSE):

$$\mathrm{RE}(\hat{\boldsymbol{\lambda}}_j) := \frac{1}{K} \sum_{k=1}^{K} \mathrm{RE}(\hat{\lambda}_{j,k} | \hat{\lambda}_{\star,k}),$$

$$\mathrm{RE}(\hat{\lambda}_{j,k} | \hat{\lambda}_{\star,k}) := \frac{\mathrm{MSE}(\hat{\lambda}_{\star,k})}{\mathrm{MSE}(\hat{\lambda}_{j,k})},$$

$$\mathrm{MSE}(\hat{\lambda}_{\star,k}) := \min_{j=1,\ldots,J} \mathrm{MSE}(\hat{\lambda}_{j,k}).$$

Hence, if $\hat{\boldsymbol{\lambda}}_j$ is the best estimator for all the components of $\boldsymbol{\lambda}$, then $\mathrm{RE}(\hat{\boldsymbol{\lambda}}_j) = 1$. We estimate $\mathrm{RE}(\hat{\boldsymbol{\lambda}}_j)$ by Monte Carlo with 1000 replicates, where $\hat{\boldsymbol{\lambda}}_j$ is obtained by maximizing the approximate likelihood with a com-
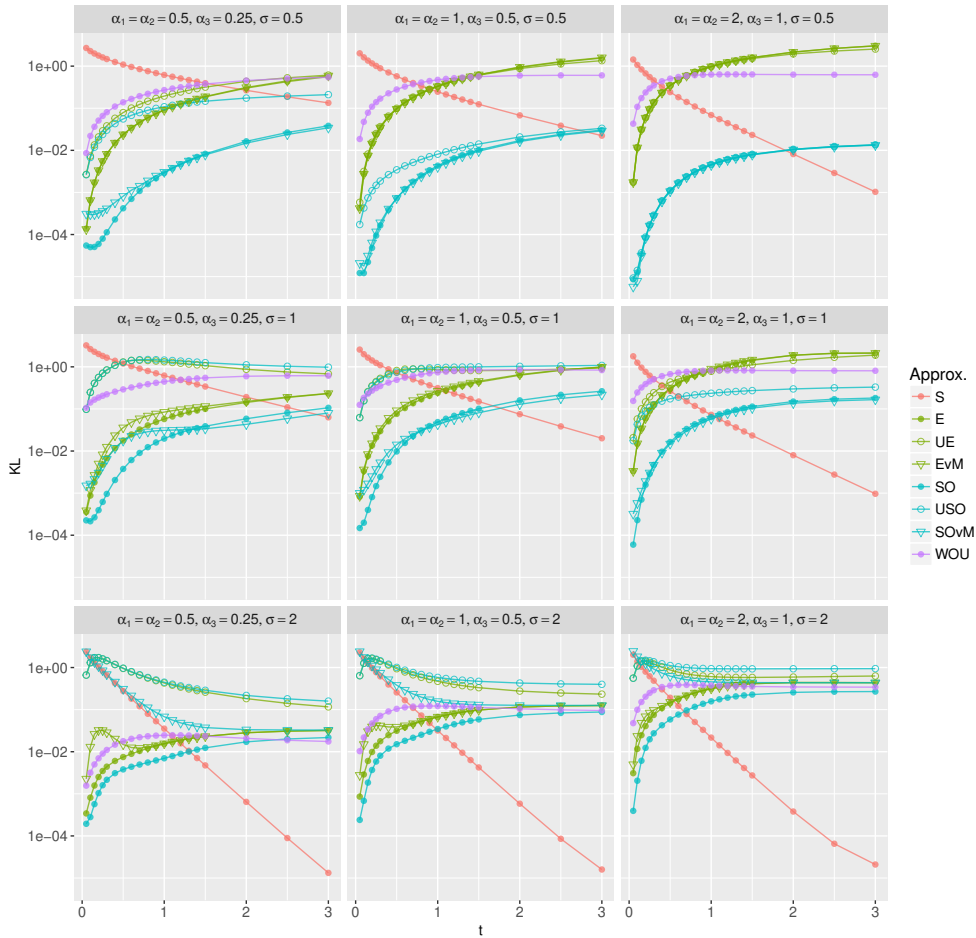
Fig. 7 D^{A}_{t,σ^2_0} curves for $p = 2$ on the MvM process. The description in Figure 5 applies.

mon optimization procedure that employs (11) as starting values.

4.2.1 WN process

Table 1 shows the relative efficiencies for E, SO, WOU, and PDE with $p = 1$. When averaging across scenarios and discretization times, the global ranking of performance is: WOU (0.9195), PDE (0.8831), SO (0.8766), and E (0.7642). On average, E is the best performing method for $\Delta = 0.05$, followed closely by SO. However, the relative performance of E severely decays as Δ increases. A similar pattern is present for SO, although the decay in relative efficiency is less severe, being by a narrow margin the best performing method for $\Delta = 0.20$ (above E and WOU with an absolute difference lower than 0.5%). PDE is significantly underperforming for $\Delta = 0.05, 0.20$, which is explained by the bias induced by the initial condition: $\sigma_0 = 0.1$ was considered as a compromise between tractability

$(M_x = 500, M_t = \lceil 100\Delta \rceil)$ and accuracy. PDE becomes the best performer on average for $\Delta = 0.50, 1.00$, where the effects of the initial condition become less important. WOU shows an intermediate profile with an indubitable advantage: on average, its relative efficiency has an absolute difference with respect to the best performing method of less than 2.5%. This fact is what makes it the best method on the global ranking of performance.

Table 2 gives the relative efficiencies for E, SO, and WOU in $p = 2$. When averaging across scenarios and discretization times, the global ranking of performance is: WOU (0.9608), SO (0.8372), and E (0.6607). Similarly to $p = 1$, E is the best performing method for $\Delta = 0.05$ and its relative efficiency quickly decays as Δ increases. SO and WOU perform similarly for low diffusive scenarios ($\sigma = 1$), but for $\sigma = 2$ WOU significantly outperforms SO for $\Delta = 0.20, 050, 1.00$, a fact explained by the proneness of the tpd to be multimodal in those situations. The competitive performance of WOU for

Δ	$\alpha = 0.5$, $\sigma = 1$				$\alpha = 1$, $\sigma = 1$			
	E	SO	WOU	PDE	E	SO	WOU	PDE
0.05	**0.9799**	0.9392	0.9608	0.7596	**0.9888**	0.9229	0.9241	0.7276
0.20	**0.9631**	0.8554	0.8878	0.8319	**0.9937**	0.8425	0.8422	0.7852
0.50	0.8941	0.7444	0.9016	**0.9340**	0.6907	0.9904	**1.0000**	0.9826
1.00	0.5685	0.7504	0.8978	**1.0000**	0.5329	0.9763	**0.9972**	0.9969

Δ	$\alpha = 0.5$, $\sigma = 2$				$\alpha = 1$, $\sigma = 2$			
	E	SO	WOU	PDE	E	SO	WOU	PDE
0.05	0.9688	**0.9972**	0.9700	0.9098	0.9392	**0.9431**	0.9205	0.8760
0.20	0.7586	**0.9740**	0.8586	0.7805	0.8040	0.8670	0.9319	**0.9380**
0.50	0.6272	0.9565	**1.0000**	0.8535	0.6297	0.7321	0.8368	**1.0000**
1.00	0.2784	0.6904	**1.0000**	0.8578	0.6090	0.8437	0.7823	**0.8964**

Table 1 Relative efficiencies for WN diffusion with $p = 1$ and $\mu = \frac{\pi}{2}$.

$p = 1, 2$ under all scenarios and Δ's, in addition to its affordable computational cost, places it as the preferred estimation method for the WN process.

Δ	$\alpha = 1$, $\sigma = 1$			$\alpha = 2$, $\sigma = 1$		
	E	SO	WOU	E	SO	WOU
0.05	**0.9765**	0.9244	0.8999	**0.9920**	0.8452	0.8460
0.20	**0.9985**	0.8214	0.8229	0.7234	0.9978	**0.9993**
0.50	0.5679	0.9868	**0.9972**	0.4370	**1.0000**	0.9980
1.00	0.4296	0.9872	**0.9998**	0.3467	**1.0000**	0.9970

Δ	$\alpha = 1$, $\sigma = 2$			$\alpha = 2$, $\sigma = 2$		
	E	SO	WOU	E	SO	WOU
0.05	0.9297	**1.0000**	0.9422	**0.9635**	0.8752	0.8793
0.20	0.8249	0.9573	**0.9916**	0.6017	0.7333	**1.0000**
0.50	0.6050	0.6607	**1.0000**	0.3797	0.6406	**1.0000**
1.00	0.5254	0.5432	**1.0000**	0.2690	0.4214	**1.0000**

Table 2 Relative efficiencies for WN diffusion with $p = 2$, $\mu = \left(\frac{\pi}{2}, -\frac{\pi}{2}\right)$, $\alpha_1 = \alpha_2 = \alpha$, $\alpha_3 = \frac{\alpha}{2}$, and $\Sigma = \sigma^2 \mathbf{I}$.

Tpd approx.	$t \to 0$	$t \in \mathbb{R}^+$	$t \to \infty$	Comput. expediency
E	★★★★★	★★	★	★★★★★
SO	★★★★	★★★	★★★	★★★
WOU	★★★★	★★★★	★★★★★	★★★★
PDE	★★★	★★★★★	★★★★★	★

Table 3 Comparative of estimation methods for the WN process in $p = 1, 2$. The number of stars ranges from one to five. The more stars, the better performance in the category. The first three columns give the behaviour of the tpd approximation when t is small, intermediate, and large, respectively.

4.2.2 Other processes

The WC diffusion has a remarkably different drift from the WN process (Figure 1). As a consequence, the tpd of the WC diffusion quickly becomes highly non-WN (multimodal, "heavy tails", peaked), both the opposite defining features of the pseudo-tpds. This affects the relative efficiencies for E, SO, and PDE given in Table 4, whose global performance is: PDE (0.9727), SO (0.4587), and E (0.4131). The supremacy of the PDE, except for small drift ($\alpha = 0.5$) and $\Delta = 0.05$, is evident. Thus, Table 4 is an illustration of the low efficiency of applying the Euler and Shoji-Ozaki pseudo-likelihoods for highly non-WN processes at arbitrary Δ's.

Δ	$\alpha = 0.5$, $\sigma = 1$			$\alpha = 1$, $\sigma = 1$		
	E	SO	PDE	E	SO	PDE
0.05	0.9277	**1.0000**	0.7682	0.5715	0.5938	**0.9309**
0.20	0.5968	0.7315	**1.0000**	0.3418	0.3524	**1.0000**
0.50	0.3548	0.4264	**1.0000**	0.2923	0.3030	**1.0000**
1.00	0.3068	0.3295	**1.0000**	0.2865	0.2774	**1.0000**

Δ	$\alpha = 0.5$, $\sigma = 2$			$\alpha = 1$, $\sigma = 2$		
	E	SO	PDE	E	SO	PDE
0.05	**0.9686**	0.8947	0.8646	0.7325	0.6734	**1.0000**
0.20	0.8114	0.8720	**1.0000**	0.0213	0.1196	**1.0000**
0.50	0.1867	0.3634	**1.0000**	0.0258	0.0880	**1.0000**
1.00	0.1417	0.2396	**1.0000**	0.0441	0.0750	**1.0000**

Table 4 Relative efficiencies for the WC diffusion ($p = 1$) with $\mu = \frac{\pi}{2}$.

Finally, Table 5 shows the relative efficiencies of E and SO for a mivM diffusion with antipodal means. In order to avoid spurious maximums, q was estimated

by SMLE and then kept fixed when optimizing the approximate likelihood. The global performances are: SO (0.9655), and E (0.8920). The analysis by Δ's shows that, as in the WC diffusion, SO is performing better than E except for $\Delta = 0.05$. However, inspection of the tpd shows a prevalent multimodality, which points towards a low efficiency of the pseudo-likelihoods when Δ is not small.

Δ	$q = 0.25$		$q = 0.50$		$q = 0.75$	
	E	SO	E	SO	E	SO
0.05	0.9282	**0.9595**	**0.9851**	0.9620	**0.9716**	0.9527
0.20	0.8678	**0.9901**	0.8999	**0.9616**	**0.9517**	0.9296
0.50	0.8312	**0.9825**	0.8223	**0.9454**	0.9448	**0.9640**
1.00	0.8867	**0.9984**	0.8625	**0.9742**	0.7525	**0.9661**

Table 5 Relative efficiencies for the mivM diffusion with $p = 2$, $\mathbf{M} = \left(\frac{\pi}{2}, \frac{\pi}{2}; -\frac{\pi}{2}, -\frac{\pi}{2}\right)$, $\mathbf{A} = \left(\frac{3}{4}, \frac{3}{4}; \frac{3}{2}, \frac{3}{2}\right)$, $\mathbf{p} = (q, 1-q)$, and $\sigma = 1$.

5 Application to molecular dynamics

Toroidal data arises from the representation of the backbone of a protein made of n amino acids as a sequence of $n-2$ pairs of dihedral angles (ϕ, ψ), thus as a point in $\mathbb{T}^{2(n-2)}$. The dihedral angles capture the rotations around the N–C_α and C_α–C bonds, which are the remaining degrees of freedom of the backbone (if the bond angles and bond lengths are assumed fixed to their ideal values). Molecular dynamics simulations are widely employed to study the folding and the dynamical properties of proteins, providing ultra high frequency trajectories of protein structures. The dihedral angles of the time-varying backbone result in a trajectory $\{(\phi_{1,i\Delta}, \psi_{1,i\Delta}, \ldots, \phi_{n-2,i\Delta}, \psi_{n-2,i\Delta})\}_{i=0}^{N}$. Diffusive models on the torus are appropriate tools to summarize these trajectories and, once fitted, can be used as computationally affordable emulators of the physical process.

We consider data from molecular dynamics simulations of the protein G (Protein Data Bank identifier 1GB1) around its native state. This protein contains $n = 56$ amino acids and, due to its relatively small size and availability of extensive experimental data, is commonly considered in the molecular dynamics literature. The molecular dynamics simulations were done using the CHARMM36 force field with the EEF1-SB solvent model (Bottaro et al, 2013) during $T = 100$ nanoseconds equally discretized in 10000 time cuts, which afterwards were subsampled to $N = 1000$. For

the sake of illustration, we study two specific trajectories: $\{\psi_{\Delta i}\}_{i=0}^{N}$ of the 9-th amino acid (Glycine, between Asparagine and Lysine), and $\{(\phi_{\Delta i}, \psi_{\Delta i})\}_{i=0}^{N}$ of the 14-th amino acid (Glycine, between Lysine and Glutamate). These one- and two-dimensional trajectories exhibit multi- and unimodal patterns that are representative of the general case.

The one-dimensional multimodal trajectory was modelled with a diffusion driven by a mixture of two vM distributions, as given in (8). The fitting was done with the PDE method with $M_x = 500$, $M_t = 20$, and $\sigma_0 = 0.01$. We used SMLE and (12) as starting values, and fixed the mixture proportions to the stationary estimates to avoid spurious minima. The optimization took 115 seconds in a 1.7 GHz core for 566 likelihood evaluations and gave $\hat{\boldsymbol{\alpha}} = (9.06, 5.00)$, $\hat{\boldsymbol{\mu}} = (0.23, -2.91)$, $\hat{\sigma} = 1.08$, and $\hat{p} = 0.56$. The first row of Figure 8 presents a graphical summary of the parametric fit. The first panel shows the observed data and a simulated trajectory from the fitted model, which captures the main patterns of the observed data, except for some outliers.

In order to evaluate the goodness-of-fit of the parametric model – and due to the absence of formal tests directly applicable in this setting, to the best of the authors' knowledge – we compared graphically the parametric fits of the drift and diffusion coefficient with their nonparametric estimations. To that aim, we considered the following Nadaraya-Watson estimator for the drift

$$\hat{b}_h(\theta) := \sum_{i=0}^{N-1} W_h(\theta, \Theta_{\Delta i}) Y_i, \qquad (25)$$

$$W_h(\theta, \Theta_{\Delta i}) := \frac{e^{\cos(\theta - \Theta_{\Delta i})/h^2}}{\sum_{j=0}^{N-1} e^{\cos(\theta - \Theta_{\Delta j})/h^2}},$$

with $Y_i := \mathrm{cmod}\left(\Theta_{\Delta(i+1)} - \Theta_{\Delta i}\right)/\Delta$ and h as the bandwidth parameter. For the diffusion coefficient, we set $Y_i := \left(\mathrm{cmod}\left(\Theta_{\Delta(i+1)} - \Theta_{\Delta i}\right)\right)^2/\Delta$ and then took the square root in the estimate. To remove the smoothing bias of (25), we smoothed the parametric estimate by considering $Y_i = b(\Theta_{i\Delta}; \hat{\boldsymbol{\lambda}})$ in (25), hence equating both biases under the correct specification of the model. The second panel in first row of Figure 8 compares the nonparametric and smoothed parametric estimates of the drift. Both drifts are shadowed according to a kernel density estimate that emphasizes the regions were the data is present. For those regions, there is a close match between both estimates. The third panel shows a similar analysis for the diffusion coefficient, whose nonparametric estimate exhibits mild departures from $\hat{\sigma}$ in the regions with high density.
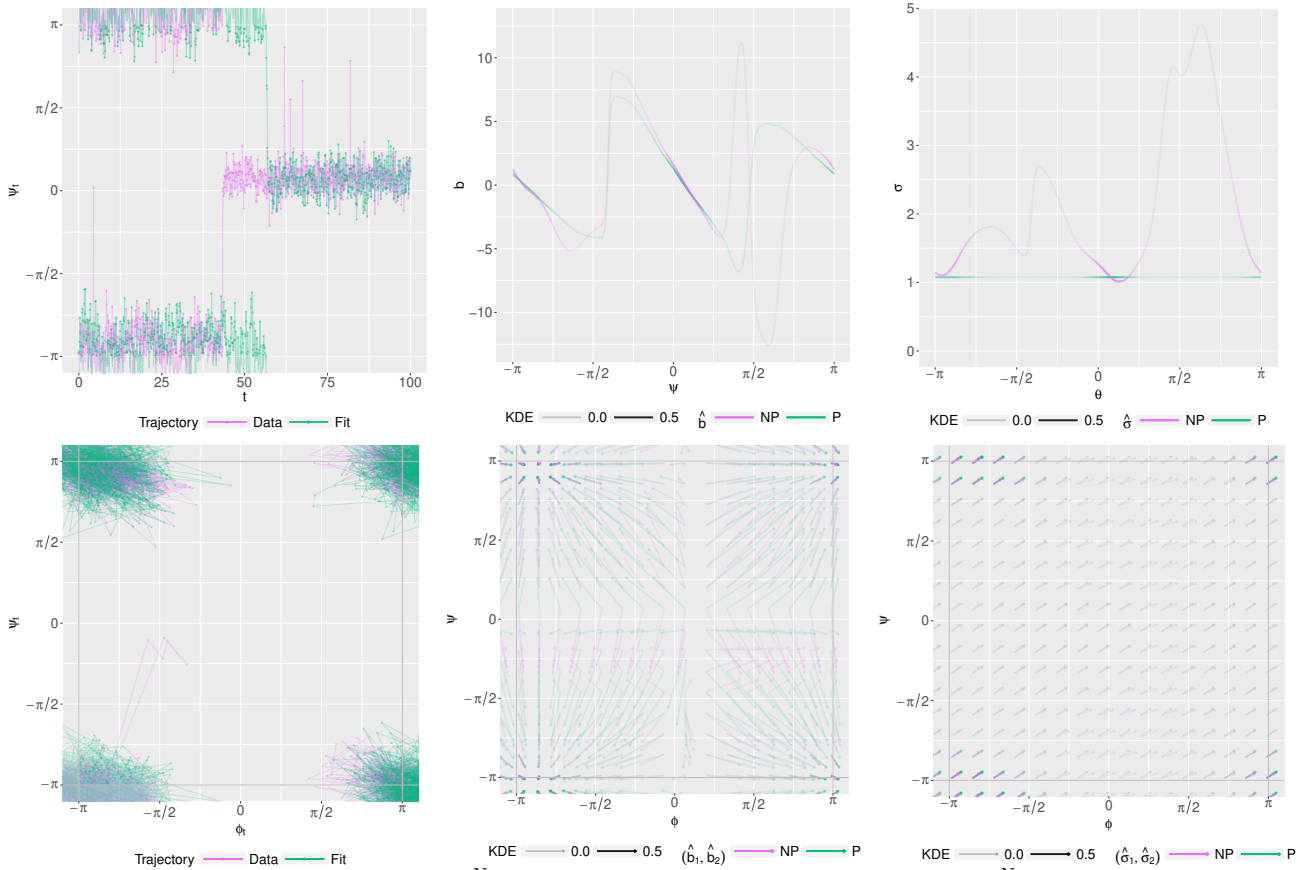
Fig. 8 Summary plots for the fits of $\{\psi_{\Delta i}\}_{i=0}^{N}$ (first row; 9-th amino acid) and $\{(\phi_{\Delta i}, \psi_{\Delta i})\}_{i=0}^{N}$ (second row; 14-th amino acid). From left to right, by columns: observed trajectory and a sample from the fitted model; NonParametric (NP) and Parametric (P) drift estimates; NP and P diffusion estimates. Shading is done according to a Kernel Density Estimate (KDE) of the observed data.

For modelling the two-dimensional and unimodal trajectory we employed a bivariate WN diffusion with unconstrained $\boldsymbol{\Sigma}$. The fitting was done with the WOU approximation using SMLE and (12) for starting values. The optimization took 14 seconds for 2739 approximate likelihood evaluations. The first panel in the second row of Figure 8 shows the correct match between the simulated and the observed trajectories, again except for some outliers from the latter. The next panel shows the comparison between the vector fields for the smoothed parametric and nonparametric drifts. They show a strong agreement on the drift structure at regions with presence of data, both in magnitude and direction. The parametric vector field $(\sigma_1(\phi, \psi; \hat{\boldsymbol{\lambda}}), \sigma_2(\phi, \psi; \hat{\boldsymbol{\lambda}}))$ and the nonparametric $(\hat{\sigma}_{1,h_1}(\phi, \psi), \hat{\sigma}_{2,h_2}(\phi, \psi))$ have a proper match for the regions with data, the latter being constant in most of $\mathbb{T}^p$. The nonparametric estimates were constructed by considering product kernels on the covariates. All the bandwidths were automatically selected by cross-validation.

6 Conclusions

We introduced ergodic diffusions on the torus as the natural processes with stationary distributions equal to well-known toroidal distributions. The WN process, with an available analytical approximation to its tpd, is shown to be the most tractable OU-like toroidal process among the different proposals. This approximation outperforms the wrapped Euler and Shoji-Ozaki pseudo-likelihoods, and shows an affordable computational cost for one and two dimensions. In addition, we provide numerical solutions of the one- and two-dimensional Fokker-Planck PDEs for approximating the true tpd, which serve as benchmarks of the accuracy of the approximating tpds. A thorough simulation study explored the performance of the approximate likelihoods under different scenarios. Finally, a data application illustrated the usefulness of the new diffusive models for modelling molecular dynamics simulations.

We summarize some important practical conclusions. For estimating the WN process, we recommend to use

WOU as a first option for a fast and accurate approximation in dimensions $p = 1, 2$. For a general process, we advise to employ PDE with $p = 1$ if accuracy is a priority, and SO in case speed is. For $p = 2$, SO is preferred to E, but both are prone to underperform severely for highly non-WN tpds, which can be visualized using the PDE solution.

The development of a general and computationally fast method for approximating an arbitrary tpd, that is able to cope with multimodality, remains an open challenge. A promising avenue is methods based on simulation, which have been successful for Euclidean diffusions; see e.g. Beskos et al (2006), Papaspiliopoulos and Roberts (2012), Sermaidis et al (2012), Bladt et al (2016), and references in these papers. The simplest algorithm by Beskos et al (2006a) is well suited for exact simulation of the transient diffusion (*i.e.*, before wrapping) because of the periodicity of the coefficients, and the method in Sermaidis et al (2012) is applicable to Langevin diffusions. It is therefore likely that the exact simulation methods can be adapted to toroidal Langevin diffusions by finding ways to deal with the wrapping when simulating diffusion bridges. It is also of interest to study whether the coupling methods underlying the diffusions bridge simulation technique in Bladt et al (2016) can be adapted to the torus setting. Another interesting approach would be to include the winding number for each observation as a latent variable and apply methods like the EM algorithm or the Gibbs sampler for likelihood inference.

Software

The software `sdetorus`, available at `https://github.com/egarpor/sdetorus`, contains the implementations of the methods described in the paper and the files required for reproducing all the empirical analyses.

References

Banerjee A, Dhillon IS, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. J Mach Learn Res 6:1345–1382

Bernstein DS, So W (1993) Some explicit formulas for the matrix exponential. IEEE Trans Automat Control 38(8):1228–1232

Beskos A, Papaspiliopoulos O, Roberts GO, Fearnhead P (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. J R Stat Soc Ser B Stat Methodol 68(3):333–382

Beskos A, Papaspiliopoulos O, Roberts GO (2006) Retrospective exact simulation of diffusion sample paths with applications. Bernoulli 12(6):1077–1098

Bibby BM, Sørensen M (2001) Simplified estimating functions for diffusion models with a high-dimensional parameter. Scand J Statist 28(1):99–112

Bladt M, Finch S, Sørensen M (2016) Simulation of multivariate diffusion bridges. J R Stat Soc Ser B Stat Methodol 78(2):343–369

Bottaro S, Lindorff-Larsen K, Best RB (2013) Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. J Chem Theory Comput 9(12):5641–5652

Breckling J (1989) The analysis of directional time series: applications to wind speed and direction, Lecture Notes in Statistics, vol 61. Springer-Verlag, Berlin

Codling E, Hill N (2005) Calculating spatial statistics for velocity jump processes with experimentally observed reorientation parameters. J Math Biol 51(5):527–556

Dacunha-Castelle D, Florens-Zmirou D (1986) Estimation of the coefficients of a diffusion from discrete observations. Stochastics 19(4):263–284

Dehay D (2015) Parameter maximum likelihood estimation problem for time periodic modulated drift Ornstein Uhlenbeck processes. Stat Inference Stoch Process 18(1):69–98

Dehling H, Franke B, Kott T (2010) Drift estimation for a periodic mean reversion process. Stat Inference Stoch Process 13(3):175–192

Émery M (1989) Stochastic calculus in manifolds. Universitext, Springer-Verlag, Berlin

Frank TD (2005) Nonlinear Fokker-Planck equations: Fundamentals and applications. Springer Series in Synergetics, Springer-Verlag, Berlin

Hill N, Häder DP (1997) A biased random walk model for the trajectories of swimming micro-organisms. J Theor Biol 186(4):503–526

In 't Hout KJ, Foulon S (2010) ADI finite difference schemes for option pricing in the Heston model with correlation. Int J Numer Anal Model 7(2):303–320

Hsu EP (2002) Stochastic analysis on manifolds, Graduate Studies in Mathematics, vol 38. American Mathematical Society, Providence, RI

Iacus SM (2008) Simulation and inference for stochastic differential equations: with R examples. Springer Series in Statistics, Springer, New York

Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics, Series on Multivariate Analysis, vol 5. World Scientific Publishing, River Edge

Jona-Lasinio G, Gelfand A, Jona-Lasinio M (2012) Spatial analysis of wave direction data using wrapped Gaussian processes. Ann Appl Stat 6(4):1478–1498

Jones MC, Pewsey A (2005) A family of symmetric distributions on the circle. J Amer Statist Assoc 100(472):1422–1428

Kato S (2010) A Markov process for circular data. J R Stat Soc Ser B Stat Methodol 72(5):655–672

Kent J (1975) Discussion of paper by K. V. Mardia. J Roy Statist Soc Ser B 37(3):377–378

Kent J (1978) Time-reversible diffusions. Adv in Appl Probab 10(4):819–835

Kessler M (2000) Simple and explicit estimating functions for a discretely observed diffusion process. Scand J Statist 27(1):65–82

Kessler M, Sørensen M (1999) Estimating equations based on eigenfunctions for a discretely observed diffusion process. Bernoulli 5(2):299–314

Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations, Applications of Mathematics (New York), vol 23. Springer-Verlag, Berlin

Kolmogoroff A (1937) Zur Umkehrbarkeit der statistischen Naturgesetze. Math Ann 113(1):766–772

Larsen KS, Sørensen M (2007) Diffusion models for exchange rates in a target zone. Math Finance 17(2):285–306

Liu CS (2013) Ornstein-Uhlenbeck process, Cauchy process, and Ornstein-Uhlenbeck-Cauchy process on a circle. Applied Mathematics Letters 26(9):957–962

Mardia KV (1972) Statistics of directional data, Probability and Mathematical Statistics, vol 13. Academic Press, London

Mardia KV (2017) The magic of score matching estimators and approximations for distributions on manifolds and some cutting edge applications to molecular biology. In: Proceedings 61st ISI World Statistics Congress, Marrakech

Mardia KV, Frellsen J (2012) Statistics of bivariate von Mises distributions. In: Hamelryck T, Mardia KV, Ferkinghoff-Borg J (eds) Bayesian methods in structural bioinformatics, Statistics for Biology and Health, Springer, Berlin

Mardia KV, Jupp PE (2000) Directional statistics, 2nd edn. Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester

Mardia KV, Voss J (2014) Some fundamental properties of a multivariate von Mises distribution. Comm Statist Theory Methods 43(6):1132–1144

Mardia KV, Hughes G, Taylor CC, Singh H (2008) A multivariate von Mises distribution with applications to bioinformatics. Canad J Statist 36(1):99–109

McKee S, Wall DP, Wilson SK (1996) An alternating direction implicit scheme for parabolic equations with mixed derivative and convective terms. J Comput Phys 126(1):64–76

Øksendal B (2003) Stochastic differential equations: An introduction with applications, sixth edn. Universitext, Springer-Verlag, Berlin

Ozaki T (1985) Statistical identification of storage models with application to stochastic hydrology. J Am Water Resour Assoc 21(4):663–675

Papaspiliopoulos O, Roberts G (2012) Importance sampling techniques for estimation of diffusion models. In: Kessler M, Lindner A, Sørensen M (eds) Statistical methods for stochastic differential equations, Monographs on Statistics and Applied Probability, vol 124, Chapman & Hall/CRC Press, Boca Raton, FL

Roberts GO, Stramer O (2002) Langevin diffusions and Metropolis-Hastings algorithms. Methodol Comput Appl Probab 4(4):337–357 (2003), international Workshop in Applied Probability (Caracas, 2002)

Rogers LCG, Williams D (2000) Diffusions, Markov processes, and martingales. Vol. 1. Cambridge Mathematical Library, Cambridge University Press, Cambridge

Sermaidis G, Papaspiliopoulos O, Roberts GO, Beskos A, Fearnhead P (2012) Markov chain monte carlo for exact inference for diffusions. Scand J Statist 40(2):294–321

Shoji I, Ozaki T (1998) A statistical method of estimation and simulation for systems of stochastic differential equations. Biometrika 85(1):240–243

Soetaert K, Cash J, Mazzia F (2012) Solving differential equations in R. Use R!, Springer, New York

Sørensen M (2008) Efficient estimation for ergodic diffusions sampled at high frequency. Tech. rep., Department of Mathematical Sciences, University of Copenhagen

Sørensen M (2012) Estimating functions for diffusion-type processes. In: Kessler M, Lindner A, Sørensen M (eds) Statistical methods for stochastic differential equations, Monographs on Statistics and Applied Probability, vol 124, Chapman & Hall/CRC Press, Boca Raton, FL

Steele JM (2001) Stochastic calculus and financial applications, Applications of Mathematics (New York), vol 45. Springer-Verlag, New York

Stroock DW (2000) An introduction to the analysis of paths on a Riemannian manifold, Mathematical Surveys and Monographs, vol 74. American Mathematical Society, Providence, RI

Thomas JW (1995) Numerical partial differential equations: finite difference methods, Texts in Applied Mathematics, vol 22. Springer-Verlag, New York

Wehrly TE, Johnson RA (1979) Bivariate models for dependence of angular observations and a related Markov process. Biometrika 67(1):255–256

Yeh SY, Harris KDM, Jupp PE (2013) A drifting Markov process on the circle, with physical applications. P Roy Soc Lond A Mat 469(2156)