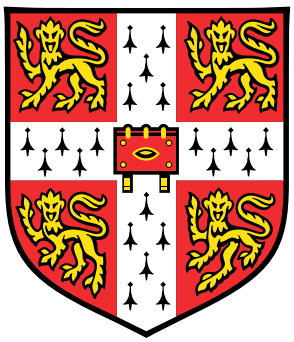
Probabilistic Machine Learning for Circular Statistics

Models and inference using the Multivariate Generalised von Mises distribution



Alexandre Khae Wu Navarro

Department of Engineering

University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy

Magdalene College 2017

I would like to dedicate this work to all those that gave me support to overcome difficult *moments*.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Alexandre Khae Wu Navarro
2017

Acknowledgements

This thesis would not be possible without the support of several people and organisations. Here I briefly note down some of the remarkable people whom I am utterly grateful for their role in this project.

I would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), the Cambridge Trusts and Sir Colin Corness Student Bursary, without which the research described in this thesis would have not been possible.

I have also to thank Dr. Richard Turner for his instrumental role in the completion of this project through his insightful remarks, support and guidance. Another major contributor was Dr. Jes Frellsen, whom I had many instigating discussions that spurred our collaboration.

Magdalene College and the Engineering Department also provided support for this thesis on multiple levels. In particular, my tutors Prof. Paul Dupree and Dr. Luke Skinner have been extremely helpful in times of unexpected hardships which I faced during the course of my studies. Cris Percival's efficiency was also paramount for surpassing these obstacles life hurdled at me. I have also to thank Dr. Sae Franklin and prof. David Franklin for their invaluable advice, as well as Louise Segar and Diane Hazell for interesting conversations.

Last, but definitely not least, I wish to thank my wife Manoela Carpenedo, my family and my friends for who stood by my side in the moments of joy and tears prompted by this project.

Abstract

Probabilistic machine learning and circular statistics—the branch of statistics concerned with data as angles and directions—are two research communities that have grown mostly in isolation from one another. On the one hand, probabilistic machine learning community has developed powerful frameworks for problems whose data lives on Euclidean spaces, such as Gaussian Processes, but have generally neglected other topologies studied by circular statistics. On the other hand, the approximate inference frameworks from probabilistic machine learning have only recently started to the circular statistics landscape. This thesis intends to redress the gap between these two fields by contributing to both fields with models and approximate inference algorithms. In particular, we introduce the multivariate Generalised von Mises distribution ($m\mathcal{G}v\mathcal{M}$), which allows the use of kernels in circular statistics akin to Gaussian Processes, and an augmented representation. These models account for a vast number of applications comprising both latent variable modelling and regression of circular data. Then, we propose methods to conduct approximate inference on these models. In particular, we investigate the use of Variational Inference, Expectation Propagation and Markov chain Monte Carlo methods. The variational inference route taken was a mean field approach to efficiently leverage the $m\mathcal{G}v\mathcal{M}$ tractable conditionals and create a baseline for comparison with other methods. Then, an Expectation Propagation approach is presented drawing on the Expectation Consistent Framework for Ising models and connecting the approximations used to the augmented model presented. In the final MCMC chapter, efficient Gibbs and Hamiltonian Monte Carlo samplers are derived for the $m\mathcal{G}v\mathcal{M}$ and the augmented model.

Table of contents

List of figures

List of tables

Part I

Introduction

Chapter 1

A primer on circular data, associated distributions and this thesis

Mainstream modelling in Probabilistic Machine Learning has targeted function approximation on Euclidean spaces, often disregarding available topological structure present in either observed or latent spaces. While it is true that flexible models with thousands of parameters such as Deep Neural Networks may approximate well manifolds and emulate their topological properties, these models require a substantial amount of data to be trained. Moreover, if the underlying manifold and associated topology are known *a priori*, bespoke models for these manifolds will perform better as they will not have to learn the manifold structure from the data.

A particularly important case where the modeller knows the nature of data or latent space topology is when the variables of interest are angles. Angles arise in a myriad of scientific and engineering contexts. For instance, controlling articulated or flying robots requires estimation of individual joint angles and bearings in the presence of noisy measurements. Likewise, to predict protein structures, it is paramount to be able to characterise and correctly estimate a vast number of dihedral and torsion angles between protein side chains and the main backbone. Equally important is to handle phase angles and the uncertainty surrounding them in numerous signal processing tasks from speech analysis to recovering corrupted signals in mobile phones. More generally, transformations in rigid body rotations, complex numbers and Fourier-representations rely on learning angles and invariant representations.

Representing angles as Euclidean variables neglect the invariance structure and constraints inherent to the topology conveyed by angles, e.g. the same angle is identified

by adding an integer multiple of 2π. Consequently, this has a detrimental effect on the model's predictions and uncertainty estimates. For example, when capturing the motion of a human arm, a model that does not account for the circular nature of the body joints may confidently predict configurations which would imply in body joints bending beyond their physical limits.

This thesis examines bespoke probabilistic models and approximate inference methods for systems comprised of angles, or as referred in statistical literature, circular variables. This chapter reviews circular distributions and their multivariate extensions. The remained of the thesis is divided into two major segments. Part II presents and discusses a new model for circular variables that leverages the existing machinery for Gaussian Processes in Chapter 2, as well as transformations and sparse models in Chapter 3. Part III provides methods for performing approximate inference and learning in the models of Part II, namely Variational Inference (VI) methods in Chapter 4, Expectation Propagation (EP) in Chapter 5 and Markov chain Monte Carlo in Chapter 6.

1.1 Distributions over the unit circle

In this section, we demonstrate how to generate unidimensional circular distributions. Circular distributions are identified with the topology of the unit circle, supported on $[0, 2\pi)$ and functions of trigonometric functions of a random angle ϕ. These distributions can be either generated from a transformation of a distribution over a Euclidean space or generated by defining a measurable function directly on the unit circle. The former case applies a transformation to random variables defined on the real line or plane and can be followed by an additional conditioning or marginalisation step. The latter procedure is often used to generalise existing distributions by re-defining a parameter as a function in the unit circle, as in the case of the Batschelet distribution (Batschelet, 1981), or by convolving an existing distribution with another function in the unit circle, as is the case of the Jones-Pewsey distribution (Jones and Pewsey, 2005).

The discussion presented in this section will focus on the construction from existing distributions over the real line. Traditionally, this approach is divided into three main constructions, namely *wrapped*, *projected* and *intrinsic*. These approaches are diagrammatically summarised in Figure 1.1 and explored in further detail in Section 1.1.1 and Section 1.1.2.

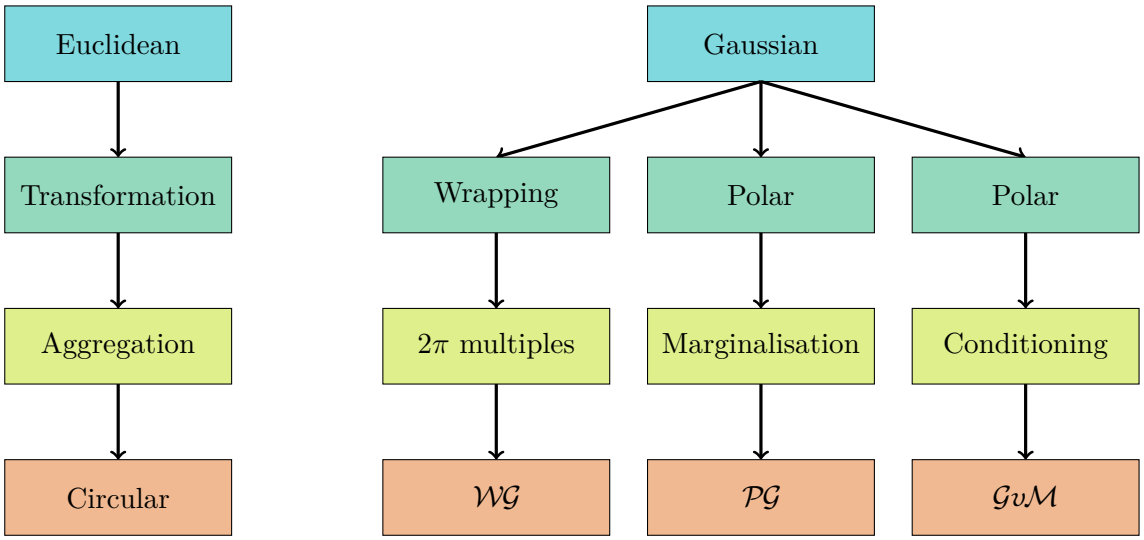


Fig. 1.1 Graphical summary of construction procedure for circular distributions from Euclidean distributions exemplified through the Gaussian distribution circular counterparts.

1.1.1 Constructions based on the *wrapping* transformation

The intuition behind the wrapping construction is that any point 2π apart in the real line correspond to the same location on the unit circle, modulo a complete clockwise revolution. Since any location on the unit circle added with a 2π multiple corresponds to the same location in the unit circle, both locations should have the same probability density. A distribution with such property can be constructed by reparametrizing each point of a distribution $p(\mathsf{x})$ with $\mathsf{x} \in \mathbb{R}$ through the modulo 2π transformation

$$\phi = \mathsf{x} \mod 2\pi. \tag{1.1}$$

Through this transformation[1], the probability mass of each 2π multiple of ϕ is condensed at the location ϕ. Therefore, any wrapped density will have the form

$$p(\phi) \propto \sum_{\mathsf{k}=-\infty}^{\infty} p(\mathsf{x} = \phi + 2\mathsf{k}\pi). \tag{1.2}$$

The name wrapped distribution arises from the fact that Equation (1.1) can be geometrically interpreted as winding, or wrapping, the real line around the unit circle. This genesis is presented in Equation (1.2) and is illustrated graphically in Figure 1.2.

[1]The transformation of Equation (1.1) can also be posited in terms of complex numbers. In this setting, the transformation becomes $\phi = e^{i\cdot\mathsf{x}}$, where e is Euler's number and $i = \sqrt{-1}$.

The univariate Gaussian analogue on the unit circle obtained from wrapping is the the

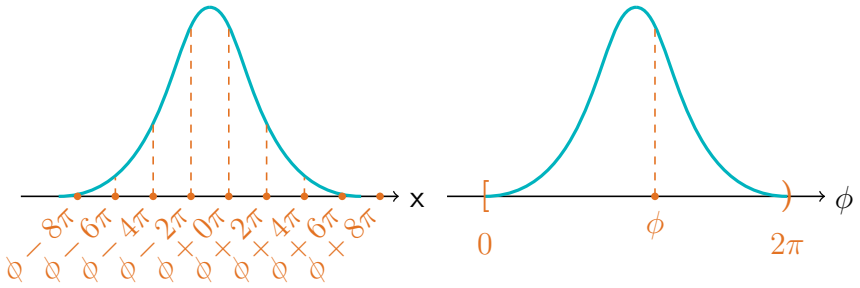


Fig. 1.2 Illustration of the construction of a wrapped distribution: the shaded regions in the distribution to the left represent 2π intervals with 2π multiples of an angle $\phi \in [0, 2\pi)$ and their respective probability (dashed coloured lines). A wrapped distribution, shown in the right figure, is constructed by condensing all the probability mass of the 2π multiples of ϕ from the distribution in the left figure to the point ϕ.

Wrapped Gaussian distribution (Pólya, 1927),

$$\mathcal{WG}(\phi; m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{\mathrm{k}=-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\frac{\phi + 2\mathrm{k}\pi - m}{\sigma}\right)^2\right\}. \tag{1.3}$$

where m and σ^2 are the mean and standard deviation of the Gaussian distribution on the real line that originated the wrapped distribution. Even though wrapped distributions can be obtained with relative ease, the infinite sums involved in their definition render them impractical in unapproximated form. For most uses, the Wrapped Gaussian truncated at the third harmonic, i.e., $\mathrm{k} = 3$ (Mardia, 1999), using adaptive truncation schemes (Jona-Lasinio et al., 2012) or modelling the truncation point as a latent variable (Jona-Lasinio et al., 2014). It is also trivial to show that wrapped distributions are not members of the exponential family, a useful property when deriving analytical relationships.

1.1.2 Constructions based on *polar* transformations

Another valid transformation to map a distribution on the real plane to the unit circle is the polar transformation, i.e. for every point in the plane $\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2]^\top$ becomes

$$\begin{aligned} \mathrm{x}_1 &= \mathrm{r}\cos(\phi) \\ \mathrm{x}_2 &= \mathrm{r}\sin(\phi) \end{aligned} \tag{1.4}$$

where $\mathrm{r} \in \mathbb{R}_+$ is a positive real scalar and ϕ is an angle in $[0, 2\pi)$. The determinant of the Jacobian for the transformation of Equation (1.4) is

$$\det \mathbf{J}_{\mathsf{x}_1,\mathsf{x}_2 \to \mathrm{r},\phi} = \det \begin{bmatrix} \frac{\partial \mathsf{x}_1}{\partial \mathrm{r}} & \frac{\partial \mathsf{x}_1}{\partial \phi} \\ \frac{\partial \mathsf{x}_2}{\partial \mathrm{r}} & \frac{\partial \mathsf{x}_2}{\partial \phi} \end{bmatrix} = \det \begin{bmatrix} \cos\phi & -\mathrm{r}\sin\phi \\ \sin\phi & \mathrm{r}\cos\phi \end{bmatrix} = \mathrm{r}. \tag{1.5}$$

The distribution yielded after the transformation is still on the real plane. Therefore, an additional step is required to eliminate the radial component r to obtain a distribution on the unit circle. There are two ways of to achieve this goal, either by *marginalising* r or by *conditioning* r to a fixed value.

Marginalisation of the radial component can be viewed geometrically as a projection, where the probability mass along a ray from the origin forming an angle ϕ with the x_1 axis is condensed to a point. This geometric interpretation illustrated in Figure 1.3 is the basis for categorising distributions constructed following this procedure as *projected* distributions.

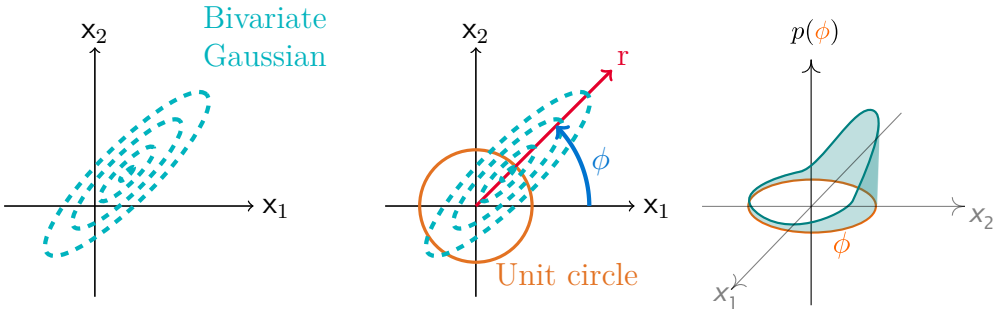


Fig. 1.3 Illustration of the genesis of projected distributions as marginalisation over a radial component. Here, the probability of ϕ is obtained by integrating over the ray forming an angle ϕ with the x_1 axis and renormalising.

The Gaussian distribution analogue obtained by the marginalisation of the radial component is the Projected Gaussian ($\mathcal{PG}$) (Mardia, 1975b). The expression for the Projected Gaussian is rather long winded and to simplify its presentation we define the quantities α, β, γ, and η as functions of ϕ, $[m_1, m_2]^\top$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{1.6}$$

of the original bivariate Gaussian distribution. These quantities are defined as

$$\alpha = \sigma_2^2 \cos^2\phi - 2\sigma_1\sigma_2 \cos\phi \sin\phi + \sigma_1^2 \sin^2\phi \tag{1.7}$$

$$\beta = \sigma_2^2 m_1 \cos\phi - \sigma_1\sigma_2\rho m_2 \cos\phi - \sigma_1\sigma_2\rho m_1 \sin\phi + \sigma_1^2 m_2 \sin\phi \tag{1.8}$$

$$\gamma = \sigma_2^2 m_1 - 2\sigma_1\sigma_2\rho m_1 m_2 + \sigma_1^2 m_2 \tag{1.9}$$

$$\eta = \sigma_1^2\sigma_2^2(1-\rho^2) \tag{1.10}$$

allow the Projected Gaussian distribution to be written as

$$\begin{aligned}\mathcal{PG}(\phi) = \frac{\exp\left\{\frac{\alpha}{\eta}\left(\frac{\gamma}{\alpha} - \frac{\beta^2}{\alpha^2}\right)\right\}}{\pi\sqrt{\det\boldsymbol{\Sigma}}} \Bigg(&\frac{\beta\sqrt{\eta\pi}}{2\alpha^{3/2}}\left(\operatorname{erf}\left[(\alpha-\beta)(\alpha\eta)^{-\frac{1}{2}}\right] - \operatorname{erf}\left[-\beta(\alpha\eta)^{-\frac{1}{2}}\right]\right) \\ &- \frac{d}{2\alpha}\left(\exp\{(\alpha-\beta)^2(\alpha\eta)^{-1}\} + \exp\{-\beta^2(\alpha\eta)^{-1}\}\right)\Bigg).\end{aligned} \tag{1.11}$$

As demonstrated by Equation (1.11), distributions obtained from marginalising of the radial component yield complicated and analytically intractable expressions when such expressions are available. More generally, the marginalisation step does not necessarily yield closed-form expressions for the circular distribution. Furthermore, theoretical properties of the base distribution, such as exponential family membership, are not preserved under the transform-then-marginalise approach.

An alternative approach to marginalising the radial component r is to condition r to a particular value, without loss of generality taken to be 1. This procedure is illustrated in Figure 1.4. Conditioning alleviates the problems associated with the marginalisation; the functional form of the base distribution is inherited by circular distribution[2] as well as exponential family membership where applicable. The preservation of these properties is a direct consequence of the Jacobian of the transformation collapsing to a scalar. The distributions arising from the transform-then-condition construction are traditionally referred to as the *intrinsic* class of circular distributions.

The intrinsic analogue for a Gaussian distribution is the Generalised von Mises distribution ($\mathcal{GvM}$) (Cox, 1975; Gatto and Jammalamadaka, 2007; Yfantis and Borgman, 1982),

$$\mathcal{GvM}(\phi; \boldsymbol{\kappa}, \boldsymbol{\mu}) = \frac{1}{2\pi\mathcal{G}(\boldsymbol{\kappa}, \boldsymbol{\mu})} \exp\left\{\kappa_1 \cos(\phi - \mu_1) + \kappa_2 \cos(2(\phi - \mu_2))\right\}, \tag{1.12}$$

[2]This procedure can result in over-parametrised forms of the distribution. The distributions often admit further simplification by applying trigonometric identities to yield minimal and compact forms.

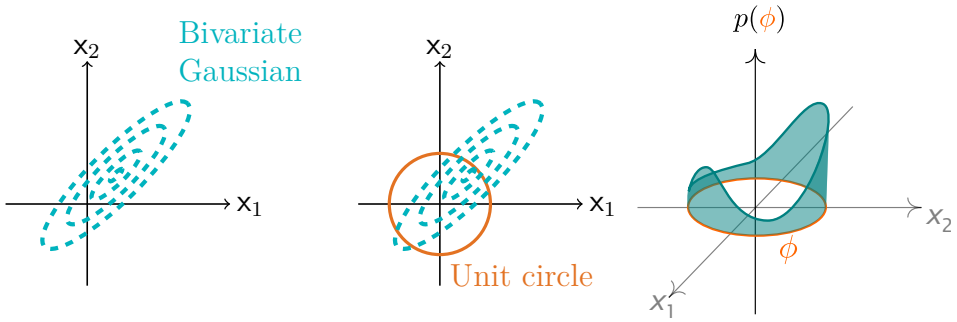


Fig. 1.4 Graphical interpretation of the construction of an *intrinsic* distribution: the Generalised von Mises distribution is constructed by conditioning a two-dimensional distribution to the unit circle, and then expressing it through an angle.

where $\mathcal{G}$ is a special function expressed as a series of modified Bessel functions of the first kind, $\mathcal{I}_j$, bearing the form

$$\mathcal{G}(\boldsymbol{\kappa}, \boldsymbol{\mu}) = \mathcal{I}_0(\kappa_1)\mathcal{I}_0(\kappa_2) + 2\sum_{\mathrm{j}=1}^{\infty} \mathcal{I}_{\mathrm{j}}(\kappa_1)\mathcal{I}_{2\mathrm{j}+1}(\kappa_2)\cos(2\mathrm{j}(\mu_1 - \mu_2)). \tag{1.13}$$

The $\mathcal{G}v\mathcal{M}$ can be either unimodal or bimodal, symmetric or asymmetric depending on the configuration of its parameters. The trigonometric moments[3] of the $\mathcal{G}v\mathcal{M}$ are available through series of modified Bessel functions of the first kind similar to the function $\mathcal{G}$ in its normalising constant, originally presented in (Gatto, 2008).

The Generalised von Mises was suggested independently by Vyacheslav Maksimov (Maksimov, 1967) in the context of distributions on groups and harmonic analysis, and Sir David Cox (Cox, 1975) as an extension of the von Mises distribution. Later, Yfantis and Bogman (Yfantis and Borgman, 1982) analysed the symmetry, modality and related properties of the Generalised von Mises. More recently, Gatto and Jammalamadaka (Gatto and Jammalamadaka, 2007) reintroduced this distribution, but expanded to an arbitrary number of cosine harmonics, that is,

$$\mathcal{G}v\mathcal{M}_{\mathrm{N}}(\phi; \boldsymbol{\kappa}, \boldsymbol{\mu}) \propto \exp\left\{\sum_{\mathrm{n}=1}^{\mathrm{N}} \kappa_{\mathrm{n}} \cos(\mathrm{n}(\phi - \mu_{\mathrm{n}}))\right\}. \tag{1.14}$$

[3]Trigonometric moments, or expectations over harmonics of sines and cosines obtained by complex exponentials

$$\left\langle e^{i\mathrm{n}\phi} \right\rangle_{p(\phi)} = \langle \cos(\mathrm{n}\phi) \rangle_{p(\phi)} + i \langle \sin(\mathrm{n}\phi) \rangle_{p(\phi)},$$

characterise circular distributions, as opposed to standard moments defined for distributions over Euclidean spaces.

However, most theoretical results and properties are provided only for the N = 2 case—including its connection with the Gaussian distribution. Therefore, the term $\mathcal{GvM}$ will be used interchangeably to $\mathcal{GvM}_2$ unless otherwise noted.

Drawing on the analytic inverse for a covariance matrix of a bivariate Gaussian,

$$\boldsymbol{\Sigma}^{-1} = \left(\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right)^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{1.15}$$

the GvM parameters $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ can be obtained by noting that

$$\mathcal{GvM}(\phi) \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \cos\phi - m_1 \\ \sin\phi - m_2 \end{bmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \cos\phi - m_1 \\ \sin\phi - m_2 \end{bmatrix} \right\}. \tag{1.16}$$

Equation (1.16) admits the expansion

$$\begin{aligned} \mathcal{GvM}(\phi) \propto \exp\Big\{ & -\frac{\cos^2\phi - 2m_1\cos\phi}{2\sigma_2^2(1-\rho^2)} - \frac{\sin^2\phi - 2m_2\sin\phi}{2\sigma_1^2(1-\rho^2)} \\ & - \rho\frac{\sin\phi\cos\phi - m_1\sin\phi - m_2\cos\phi}{2-2\rho^2} \Big\}, \end{aligned} \tag{1.17}$$

which can be simplified using trigonometric relations to yield

$$\begin{aligned} \mathcal{GvM}(\phi) \propto \exp\Big\{ & \left(\frac{m_2}{2-2\rho^2} - \frac{m_1}{2\sigma_2^2(1-\rho^2)} \right)\cos\phi + \left(\frac{\sigma_1^2 - \sigma_2^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)} \right)\cos 2\phi + \\ & \left(\frac{m_1}{2-2\rho^2} - \frac{m_2}{2\sigma_1^2(1-\rho^2)} \right)\sin\phi + \left(\frac{\rho}{2\rho^2-2} \right)\sin 2\phi \Big\}. \end{aligned} \tag{1.18}$$

Equation (1.18) implies that the Generalised von Mises parameters can be posed as a complex variable $\boldsymbol{z} \in \mathbb{C}^2$ such that

$$z_1 = \left(\frac{m_2}{2-2\rho^2} - \frac{m_1}{2\sigma_2^2(1-\rho^2)} \right) + i\left(\frac{m_1}{2-2\rho^2} - \frac{m_2}{2\sigma_1^2(1-\rho^2)} \right) \tag{1.19}$$

$$z_2 = \left(\frac{\sigma_1^2 - \sigma_2^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)} \right) + i\left(\frac{\rho}{2\rho^2-2} \right) \tag{1.20}$$

and the Generalised von Mises parameters relate to the absolute values and phase of z_1 and z_2. In particular, if we denote the absolute value of a complex number by the function abs and its angle in the interval $[0, 2\pi)$ by the function ang,

$$\kappa_1 = \mathrm{abs}(z_1), \quad \mu_1 = \mathrm{ang}(z_1), \quad \kappa_2 = \mathrm{abs}(z_2), \text{ and } \mu_2 = \frac{\mathrm{ang}(z_2)}{2}. \tag{1.21}$$

As the name indicates, the Generalised von Mises distribution is closely related to the von Mises distribution ($\upsilon\mathcal{M}$) (von Mises, 1918),

$$\upsilon\mathcal{M}(\phi;\kappa,\mu) = \frac{1}{2\pi\mathcal{I}_0(\kappa)}\exp\Big\{\kappa\cos(\phi-\mu)\Big\}. \tag{1.22}$$

The von Mises distribution was the first circular distribution ever proposed and is arguably the most widely know and studied circular distribution. It is a unimodal distribution whose trigonometric moments are obtained as ratios of Bessel functions. A Generalised von Mises distribution reduces to a von Mises distribution when the covariance matrix of the underlying bivariate Gaussian is isotropic. An excellent overview of the von Mises distribution along with its properties and statistical tests is presented by Kanti Mardia and Peter Jupp in (Mardia, 1999).

1.1.3 Other constructions in the literature

Other constructions for distributions on the unit circle make use of functional maps. These maps, called *linking functions* (ℓ) assign $-\infty$ to 0 and $+\infty$ to 2π. One of the most common linking functions is the scaled probit link,

$$\ell(\mathsf{x}) = 2\pi\left(\frac{1}{1+\exp(\beta\cdot\mathsf{x})}\right). \tag{1.23}$$

Despite the simplicity of this approach, the choice of mapping $-\infty$ to 0 and $+\infty$ to 2π is arbitrarily defined and may have undesirable consequences over the obtained density. For example, using the scaled probit linking function on a Gaussian distribution imposes that the region around 0 will necessarily have low probability mass. More precisely, for any $\xi\in\mathbb{R}, \xi\to 0$, the probability mass of the interval $[0, 0+\xi]\cup[2\pi-\xi, 2\pi)$ will tend to zero a consequence of $p(\mathsf{x}\to-\infty) = p(\mathsf{x}\to+\infty) = 0$, regardless of the mean of the Gaussian distribution. Besides showing the unsuitability of this approach, this arbitrary singularity around 0 exemplifies the need for bespoke distributions on the unit circle and its generalisations.

Another approach in literature is termed Kernel Density Estimation (KDE). This approach expresses the a distribution implicitly through the average of a mapping evaluated at a finite number of inducing points $\varrho_1,\ldots,\varrho_{\mathrm{N}}$, i.e. a distribution p is defined as

$$p(\phi) = \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} k_\theta(\phi-\varrho_{\mathrm{n}}) \tag{1.24}$$

where k_θ is the mapping that implicitly defines the distribution with parameters θ. Di Marzio et al. (2009) detailed the mapping requirements for inducing the circular distributions. This mapping $k : [0, 2\pi) \to \mathbb{R}$, termed a circular probability kernel, needs to be defined such that

1. k admits a convergent Fourier expansion with the form $\frac{1}{2\pi}\left(1 + \sum_{\mathrm{j}}^{\infty} \gamma_{\mathrm{j}}(\theta)\cos(\mathrm{j}\phi)\right)$,
2. k is normalised in the unit circle, $\int_0^{2\pi} k_\theta(\phi)\mathrm{d}\phi = 1$,
3. for $\eta_{\mathrm{j}}(k_\theta) = \int_0^{2\pi} \sin^{\mathrm{j}}(\phi)k_\theta(\phi)\mathrm{d}\phi$, if $\eta_{\mathrm{i}}(k_\theta) \neq 0$, all $\eta_{\mathrm{j}}(k_\theta) = 0$ for $0 < \mathrm{j} < \mathrm{i}$,
4. as the smoothing parameter $\theta \to \infty$, $\int_{\psi-\epsilon}^{\psi+\epsilon} k_\theta(\phi)\mathrm{d}\phi = 1$ with $\epsilon \to 0$ and $\psi \in [0, 2\pi)$.

Di Marzio and coworkers also later expanded to encompass circular regression (Di Marzio et al., 2017) merging the linking function treatment with KDEs.

Kernel Density Estimation is a flexible approach that can handle multiple modes and provide asymmetric distributions. However, these advantages are hampered by the difficulty in selecting the smoothing parameter. A common approach to solve this problem is to minimise the a loss function of the KDE differences to a known, parametric distribution.

1.2 Multivariate extensions

There are multiple ways to generalise circular variables to higher dimensions. Each type of generalisation is associated with a different kind of topological manifold. Next, we will explore distributions over the surface of (hyper-)spheres in Section 1.2.1 and distributions over the surface of (hyper-)toruses in Section 1.2.2. These subsequent sections are quite technical to provide a comprehensive introduction of the topic. However, the reader should not lose sight that the central issue discussed is that there are two main classes of distributions over multi-dimensional distributions over angles: spheres and toruses. Other less prominent topologies are also discussed in Section 1.2.4.

1.2.1 Distributions over the surface of (hyper-)spheres

Distributions over the surface of (hyper-)spheres are often referred to as *directional distributions*, as they can be geometrically interpreted as distributions over unit vectors.

Following this geometrical intuition, directional distributions can be constructed from distribution over a multi-dimensional Euclidean space by applying the transformation

$$\mathbf{s} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \tag{1.25}$$

to obtain the unitary vector s. The unitary vector s can be then recast into a distribution over angles by parametrising s in (hyper-)spherical coordinates. For example, when $\mathbf{s}$ is defined on the usual sphere ($\mathbb{S}^2$), the usual spherical transformation applies, and the unit vector can be expressed as

$$\begin{bmatrix} \mathrm{s}_1 \\ \mathrm{s}_2 \\ \mathrm{s}_3 \end{bmatrix} = \begin{bmatrix} \sin\phi_1 \cos\phi_2 \\ \sin\phi_1 \cos\phi_2 \\ \cos\phi_1 \end{bmatrix}. \tag{1.26}$$

This transformation can be related to the conditioning approach to producing distributions on the unit circle, as spherical coordinates for the unit hypersphere ($\mathbb{S}^1$) reduces to the polar coordinate system, i.e. the unit circle, rather than parametrising the circle points by an angle ϕ, each point is represented as

$$\begin{bmatrix} \mathrm{s}_1 \\ \mathrm{s}_2 \end{bmatrix} = \begin{bmatrix} \sin\phi \\ \cos\phi \end{bmatrix}, \tag{1.27}$$

The Gaussian analogue for directional distributions using this construction is the Fisher-Bingham distribution ($\mathcal{FB}$) (Mardia, 1975a),

$$\mathcal{FB}(\mathbf{s}; \kappa, \boldsymbol{\eta}, \mathbf{A}) \propto \exp\left\{\kappa \boldsymbol{\eta}^\top \mathbf{s} - \frac{1}{2}\mathbf{s}^\top \mathbf{A}^{-1}\mathbf{s}\right\}. \tag{1.28}$$

where the parameters $\kappa \in \mathbb{R}$, and $\boldsymbol{\eta}$ is an N-dimensional unit vector, are related a N-dimensional Gaussian with $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ and $\kappa\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{m}$. The Fisher-Bingham distribution is equivalent to a Generalised von Mises distribution when $\mathrm{N} = 2$.

A large number of other directional distributions that can be related to a multivariate Gaussian and the Fisher-Bingham distribution exist. These distributions often refer to a specific covariance structure for the underlying multivariate Gaussian. For example, when the covariance matrix is assumed to be diagonal, the Fisher-Bingham distribution reduces to the von Mises-Fisher distribution ($\upsilon\mathcal{MF}$) (Fisher, 1953)

$$\upsilon\mathcal{MF}(\mathbf{s}; \kappa, \boldsymbol{\eta}) = \frac{1}{\Gamma(\frac{\mathrm{N}}{2})\mathcal{I}_{\frac{\mathrm{N}}{2}-1}(\kappa)} \left(\frac{\kappa}{2}\right)^{\frac{\mathrm{N}}{2}-1} \exp\left\{\kappa \boldsymbol{\eta}^\top \mathbf{s}\right\} \tag{1.29}$$

where N is the dimension of vector $\mathbf{s}$ and $\Gamma(\cdot)$ is the Gamma function.

The von Mises-Fisher distribution is also called the Langevin distribution and particular forms of the von Mises-Fisher distribution also bear special nomenclature. When $\mathrm{N} = 3$, the distribution is known as the Fisher distribution, while the case when $\mathrm{N} = 2$ reverts to a von Mises distribution.

Another distribution closely related to the Fisher-Bingham distribution is the Bingham distribution ($\mathcal{B}$) (Bingham, 1974)

$$\mathcal{B}(\mathbf{s}; \mathbf{\Sigma}) = \frac{1}{{}_0\mathcal{F}\left(\frac{1}{2}, \frac{\mathrm{N}}{2}, \mathbf{\Sigma}^{-1}\right)} \exp\left\{\mathbf{s}^\top \mathbf{\Sigma}^{-1} \mathbf{s}\right\} \tag{1.30}$$

where ${}_0\mathcal{F}$ is the confluent hypergeometric function of matrix argument, a special function of described in further details in (Abramowitz and Stegun, 1972).

The Bingham distribution can be viewed as a zero-mean N-dimensional Gaussian with covariance $\mathbf{\Sigma}$ that has been conditioned to the surface of the $(\mathrm{N}-1)$-sphere.

Other directional distributions based on a multivariate Gaussian distribution assuming a special structure for the covariance matrix include the Watson, Fisher-Watson, Bingham-Mardia and Kent distributions. A brief summary of each of these distributions and how they relate to the Fisher-Bingham distribution is provided diagrammatically in Figure 1.5. For a review of the properties of each of these distributions, see Kanti Mardia and Peter Jupp's cornerstone reference in directional statistics (Mardia, 1999).

Distributions that capture the correlations between multiple unit vectors are analysed within the framework of distributions over orthonormal matrices $\mathbf{S} \in \mathbb{R}^{\mathrm{D}\times\mathrm{N}}$, where N is the dimension of the hypersphere and D is the number of individual unit vectors. These 'multivariate' directional distributions belong to a special topological space known as Stiefel manifolds. More formally, a Stiefel manifold is defined as the space composed of the Cartesian product of N-dimensional spheres, i.e. $\mathbb{S}^{\mathrm{N}} \times \cdots \times \mathbb{S}^{\mathrm{N}}$.

Examples of distributions over Stiefel manifolds include the Matrix von Mises-Fisher Distribution ($\mathcal{M}v\mathcal{M}\mathcal{F}$) by Downs (1972), Khatri and Mardia (1977),

$$\mathcal{M}v\mathcal{M}\mathcal{F}(\mathbf{S}; \boldsymbol{m}, \mathbf{V}, \mathbf{K}) \propto \exp\left\{(\mathrm{vec}(\mathbf{S}) - \boldsymbol{m})^\top (\mathbf{V} \otimes \mathbf{K})(\mathrm{vec}(\mathbf{S}) - \boldsymbol{m})\right\}, \tag{1.31}$$

where $\mathrm{vec}(\mathbf{S})$ is the function that constructs single column vector by stacking the columns of the matrix $\mathbf{S}$, with $\otimes$ denoting the Kronecker product between the positive definite matrices $\mathbf{V}$ and $\mathbf{K}$.

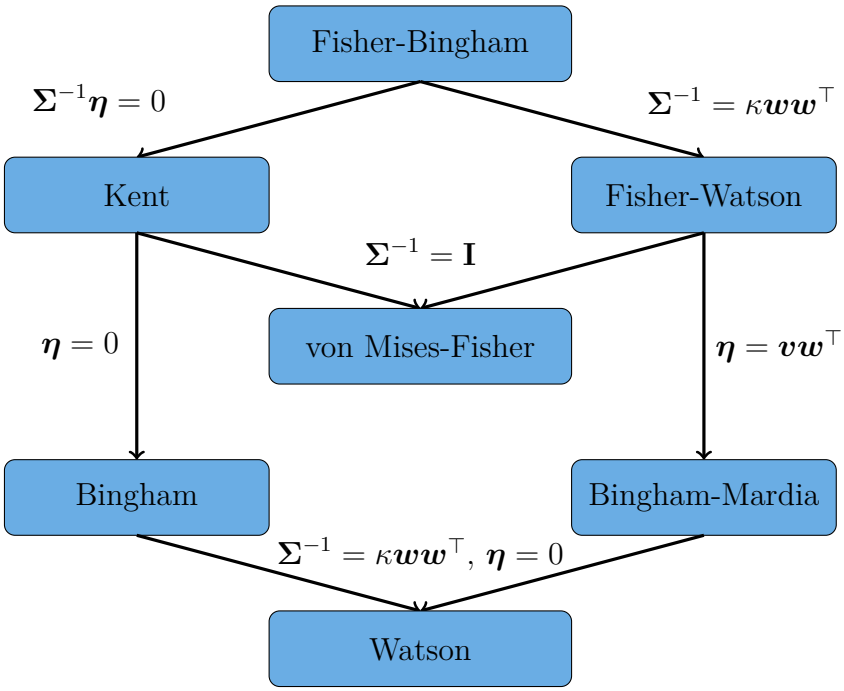


Fig. 1.5 Relationship between directional distributions based on conditioning a multivariate Gaussian to the unit circle: arrows show additional assumptions over the covariance matrix of the base Gaussian, extended from (Mardia, 1999).

The most general matrix distribution is the matrix Fisher-Bingham distribution ($\mathcal{MFB}$) (Kume et al., 2013),

$$\mathcal{MFB}(\mathbf{S}; \boldsymbol{\eta}, \boldsymbol{\Sigma}) \propto \exp\left\{\boldsymbol{\eta}^\top \text{vec}(\mathbf{S}) - \frac{1}{2}\text{vec}(\mathbf{S})^\top \boldsymbol{\Sigma}^{-1} \text{vec}(\mathbf{S})\right\} \tag{1.32}$$

where $\boldsymbol{\eta}$ is a location and concentration parameter, while $\boldsymbol{\Sigma}$ captures the covariance between unit vectors.

1.2.2 Distributions over the surface of (hyper-)toruses

Unlike distributions over the surface of (hyper-)spheres, distributions over the surface of (hyper-)toruses are described as the Cartesian product of random angles rather than two-dimensional unit vectors. The applications for distributions over (hyper-)spheres are also distinct from the applications for distributions over (hyper-)toruses. While the former is suitable for representing directions and orientations, the latter excels at capturing the phase component of complex-valued signals and sequences of correlated angular measurements.

All construction methods outlined in Section 1.1 to generate distributions on the unit circle can be used to produce distributions on the (hyper-)torus. For example, to generate a N-dimensional wrapped distribution, each variable is wrapped independently using the wrapping transformation of Equation (1.1). The same procedure applies to N-dimensional marginalised and conditioned distributions using the polar transformation.

The known (hyper-)toroidal distributions analogue to multivariate Gaussians with arbitrary covariance structure are the general Wrapped Gaussian ($g\mathcal{WG}$) (Ferrari, 2009), and the general Projected Gaussian ($g\mathcal{PG}$) (Hernandez-Stumpfhauser et al., 2017). The general Wrapped Gaussian can be analytically given as

$$\begin{aligned} g\mathcal{WG}(\boldsymbol{\phi}; \boldsymbol{m}, \boldsymbol{\Sigma}) = & \frac{1}{(2\pi)^{\frac{\mathrm{N}}{2}} (\det(\boldsymbol{\Sigma}))^{\frac{1}{2}}} \times \\ & \sum_{\mathrm{k}_1=-\infty}^{\infty} \cdots \sum_{\mathrm{k}_\mathrm{N}=-\infty}^{\infty} \exp\left\{ -\frac{1}{2} (\boldsymbol{\phi} - 2\mathbf{k}\pi)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\phi} - 2\mathbf{k}\pi) \right\} \end{aligned} \tag{1.33}$$

while the general Projected Gaussian bears no closed form expression for higher dimensions.

The conditioned analogue for a multivariate Gaussian is only known for cases where an additional structure is imposed over the covariance matrix. Mardia's multivariate von Mises distribution ($m\upsilon\mathcal{M}$) is one of such distributions (Mardia et al., 2008),

$$m\upsilon\mathcal{M}(\boldsymbol{\phi}; \boldsymbol{\kappa}, \mathbf{G}) \propto \exp\left\{ \boldsymbol{\kappa}^\top \cos \boldsymbol{\phi} - \frac{1}{2} \sin \boldsymbol{\phi}^\top \mathbf{G} \sin \boldsymbol{\phi} \right\}, \tag{1.34}$$

where $\mathbf{G}$ is a symmetric matrix with zeros on its main diagonal[4].

Mardia's multivariate von Mises assumes that the underlying 2N-dimensional Gaussian has zero mean and a covariance matrix that admits an inverse with the structure

$$\boldsymbol{\Sigma}_{\mathrm{ord}}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda} & \mathbf{A} \\ \mathbf{A} & \boldsymbol{\Lambda} \end{bmatrix} \tag{1.35}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix and $\mathbf{A}$ is an anti-symmetric matrix.

[4]The Multivariate von Mises can be also parametrised with an additional location parameter with a trivial shift $\boldsymbol{\phi}' = \boldsymbol{\phi} - \boldsymbol{\mu}$.

1.2.3 A short note on other multivariate extensions of circular distributions

Alternatives to deriving multivariate circular distributions analytically also exist. For example, Di Marzio and collaborators extended the Kernel Density Estimators for cicular models for to multiple dimension, both toroidal (Di Marzio et al., 2011) and spherical (di Marzio et al., 2012) generalisations. The treatment follows that directly from the univariate case described in Section 1.1.3.

Another important way to form multivariate distributions from univariate distributions is through the use of copulas. Copulas can be used to define dependencies in multivariate distributions through the cumulative density function of the marginal distributions. For a complete treatment, see the excellent introductions from Nelsen (2007) and Joe (1997) for an overview of the subject. In the field of circular statistics, Perlman and Wellner (2011) introduced and proved multiple results on circular copulas, and coined the term *circula* for a circular copula. Further analysis on circulas on their theoretical properties were developed later by Jones (2013) and Jones et al. (2015).

Perlman and Wellner (2011) provided a result particularly important for constructing multivariate circular distributions. They showed that only two and three dimensional circulas exist. This result has important consequences for higher dimensional generalisations of circular distributions as it implies that D-dimensional distributions on the hyper-torus or hyper-sphere can be constructed using a single circula. Instead, a *vine* of copulas must be used. Copula vines use a hierarchical structure of $\mathrm{D}(\mathrm{D}-1)/2$ bivariate conditional copulas to express a D-dimensional distribution. An example of how vines can be used for building complex dependency structures is given by Lopez-Paz et al. (2013) for Gaussian Processes.

1.2.4 Comment on the relationships between distributions on spheres, toruses and other manifolds

The relationship between distributions on the unit circle, (hyper-)spheres, (hyper-)toruses and Stiefel manifolds can be succinctly explained for distributions arising from polar transformation and conditioning. Figure 1.6 presents these relationships diagrammatically, showing how the assumptions about the underlying Gaussian distribution yield different distributions on these manifolds.

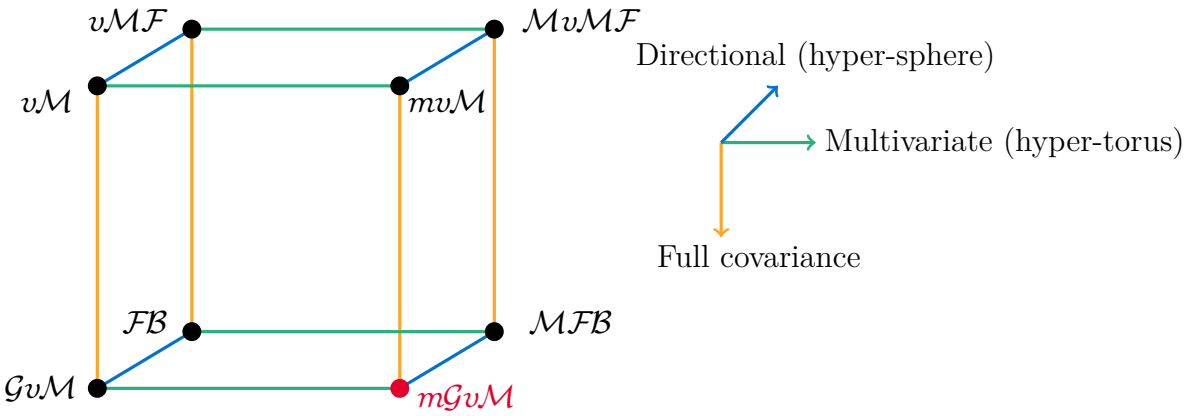


Fig. 1.6 A diagram outlining the relationship between circular and directional distributions, their multivariate extensions and the underlying Gaussian's covariance structure. A new distribution introduced in this thesis is highlighted in red.

For example, starting from the von Mises distribution, one can derive a von Mises-Fisher by considering the polar transformation a sub-case of the spherical transformation. If a toroidal construction is chosen instead, one obtains Mardia's multivariate von Mises distribution. By letting the covariance matrix of the underlying multivariate Gaussian for these distributions to have any form, one obtains the Fisher-Bingham distribution. The Stiefel manifold generalisation of directional distributions collapses onto distributions over hypertoruses when the unit vectors are two-dimensional.

In particular, the density for a hyper-toroidal topology built from a multivariate Gaussian with general covariance matrix had not been fully explored before. This distribution would be particularly important for modelling the covariances between phase angles, for example, in an array of complex signals. Furthermore, there are currently no efficient inference and learning methods for the intractable distributions on the multi-dimensional torus, sphere and Stiefel manifolds as shown in Figure 1.7.

Other topological manifolds using circular variables are (hyper-)cylindrical models. In these settings, the cylindrical distribution has both Euclidean variables and circular variables. Such models, however, are not atomic in the sense that they can be decomposed as a simple graphical model that combines (atomic) circular distributions with (atomic) Euclidean distributions, i.e.

$$p(\boldsymbol{\phi}, \mathbf{x}) = p(\boldsymbol{\phi}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\phi})p(\boldsymbol{\phi}). \tag{1.36}$$

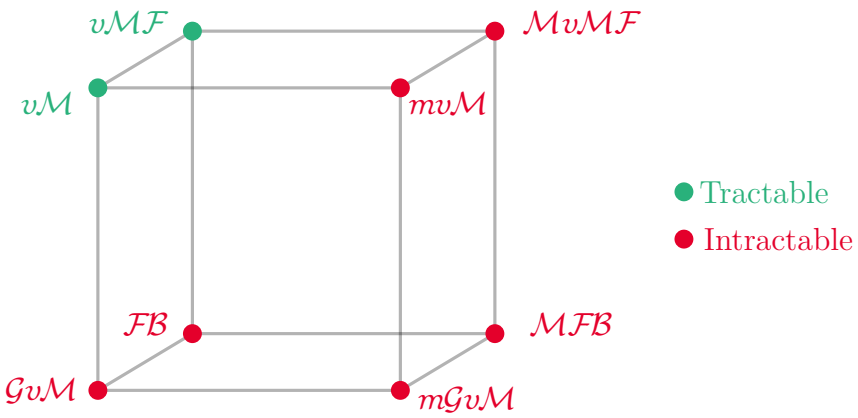


Fig. 1.7 A diagram indicating the tractable and intractable distributions on the (hyper-)torus, (hyper-)spheres and Stiefel manifolds. By tractable we imply that numerical stable and efficient approximations exist for calculating the moments of the underlying distribution (this criterion excludes the infinite series approximation for the $\mathcal{G}\upsilon\mathcal{M}$, which are only accurate for low concentration values). The methods presented in this thesis provide ways to perform inference and learning in the intractable distributions by using the tractable ones.

1.3 Aims and contributions of the thesis

There are two major contributions of this thesis: (i) introducing new bespoke multivariate models for circular variables, and (ii) providing algorithms for performing tractable inference and learning in these models. The intent behind such contributions is to redress the gap between the Machine Learning and the Circular Statistics communities so that models and algorithms for circular data are incorporated into mainstream probabilistic modelling. Moreover, the thesis uses the circular case to exemplify how intractable directional distributions on Stiefel manifolds can be inferred or approximated. The thesis is outlined as follows.

- **Chapter 2** presents and discusses the Multivariate Generalised von Mises ($m\mathcal{G}\upsilon\mathcal{M}$), a new multivariate model for circular variables that correctly represents the covariance structure between any two random angles in a toroidal topology. This distribution is also a high-dimensional extension of an unduly forgotten distribution proposed by Mardia (1975b). The Multivariate Generalised von Mises also allows us to empower circular statisticians with kernels. Originally developed for Gaussian Processes, kernels provide a flexible yet interpretable way to analyse and represent a multitude of structures in the data through the covariance of a multivariate Gaussian.

The use of kernels allows this thesis to make significant contributions to two major areas of circular statistics: circular regression and latent variable modelling. Before this thesis, regression problems between circular responses and circular inputs was treated as a different problem from regressing circular responses with Euclidean or other types of inputs. The use of kernels proposed in this thesis unifies all these problems into a single regression framework. In the latent variable modelling front, this thesis introduces a Principal Component Analysis analogue for circular variables, a previously unsolved problem in circular statistics.

- **Chapter 3** introduces model augmentations based on the Hubbard-Stratonovich transformation for arbitrary-dimensional distributions on hyper-toruses, hyper-spheres and other Stiefel manifolds. The Hubbard-Stratonovich transformation is rooted in statistical physics and is used here to eliminate computational difficulties associated with the normalising constants of these distributions. The transformation augments the $m\mathcal{G}v\mathcal{M}$, $\mathcal{FB}$ and $\mathcal{MFB}$ models with Euclidean variables that lift the distribution to (hyper-)cylindrical manifolds where inference can be efficiently computed.

- **Chapter 4** examines Variational Inference (VI) methods for performing approximate inference and learning for the multivariate Generalised von Mises models. The variational free energy framework's reliance on optimisation methods makes it a promising candidate for learning and inference in high dimensions. Here, a standard fully-factored mean-field approach is presented as a an efficient way to perform inference and a stardard to which other inference methods could be compared to.

- **Chapter 5** evaluates Expectation Propagation (EP) approaches for the $m\mathcal{G}v\mathcal{M}$. Here we introduce a novel approach based on a structured approximation following Manfred Opper and Ole Winther's approximation for Ising models (Opper and Winther, 2005). Such algorithm is central for the use of message passing methods with the $m\mathcal{G}v\mathcal{M}$ as a naïve application of the standard Expectation Propagation could not converge in our experiments.

- **Chapter 6** explores two Markov chain Monte Carlo techniques for the proposed models: Gibbs sampling and Hamiltonian Monte Carlo. Markov chain Monte Carlo methods have been widely used as the main approach to simulating circular variables. Gibbs sampling is one of the standard methods to perform inference with circular distributions and hence serves as a baseline to which all

other methods can be compared. Hamiltonian Monte Carlo represents a more sophisticated approach for Markov chain Monte Carlo. The use of contrastive divergence for learning a Multivariate Generalised von Mises is also assessed in this analysis.

- **Chapter 7** concludes the thesis reviewing the contributions of each chapter, and outlines new research avenues opened by the work contained in this thesis.

Part II

Models

Chapter 2

The multivariate Generalised von Mises distribution

This chapter introduces the multivariate Generalised von Mises distribution ($m\mathcal{G}v\mathcal{M}$), a multivariate circular density on the surface of hyper-toruses, that is the Cartesian product of N unit circles ($\mathbb{S}^1 \times \cdots \times \mathbb{S}^1$). This distribution receives its name from its one-dimensional conditionals, which are Generalised von Mises distributed. We demonstrate how to construct the $m\mathcal{G}v\mathcal{M}$, prove its central properties and introduce probabilistic models that employ the $m\mathcal{G}v\mathcal{M}$ for regression and latent variable modelling.

Besides introducing the $m\mathcal{G}v\mathcal{M}$, the chapter contains another original contribution: drawing on the machinery developed for Gaussian Process regression, we port the notion of kernels as covariance functions to the context of circular statistics through the $m\mathcal{G}v\mathcal{M}$. To best our knowledge, this is the first use of kernel functions within Circular Statistics. Covariance functions play a foundational role in Probabilistic Machine Learning, and the Gaussian Process community has developed them into a mature modelling framework. For completeness, we provide a brief overview of covariance functions and how they unify regression problems under a single framework.

We remark that apart from univariate special cases and the genesis of the $m\mathcal{G}v\mathcal{M}$ distribution, the remaining properties presented in this chapter comprise novel contributions.

2.1 The $m\mathcal{G}v\mathcal{M}$ model

Chapter 1 introduced methods to generate distributions on the hyper-torus, that is, on the Cartesian product of N unit circles $\mathbb{S}^1 \times \cdots \times \mathbb{S}^1$. One of such methods re-expressed a distribution over a Euclidean spaces in polar coordinates then conditioned

the resulting radial components to unity. The model we present in this section leverages that construction to produce the most general distribution of this kind based on the multivariate Gaussian distribution.

More precisely, starting from a 2N-dimensional random vector $\mathbf{x}$ following a multivariate Gaussian with arbitraty mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, i.e.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{m}, \boldsymbol{\Sigma}), \tag{2.1}$$

the construction in Section 1.1.2 stipulates that each of the N pairs of variables x_j and $\mathsf{x}_{\mathrm{N+j}}$ from the Gaussian must be transformed into polar coordinates and their radial components conditioned to unity, i.e.

$$\begin{bmatrix} \mathsf{x}_1 \\ \vdots \\ \mathsf{x}_\mathrm{N} \\ \mathsf{x}_{\mathrm{N}+1} \\ \vdots \\ \mathsf{x}_{2\mathrm{N}} \end{bmatrix} \xrightarrow{\text{Polar transformation}} \begin{bmatrix} r_1 \cos\phi_1 \\ \vdots \\ r_\mathrm{N} \cos\phi_\mathrm{N} \\ r_1 \sin\phi_1 \\ \vdots \\ r_\mathrm{N} \sin\phi_\mathrm{N} \end{bmatrix} \xrightarrow{r_\mathrm{i}=1 \text{ for i=1,...,N}} \begin{bmatrix} \cos\phi_1 \\ \vdots \\ \cos\phi_\mathrm{N} \\ \sin\phi_1 \\ \vdots \\ \sin\phi_\mathrm{N} \end{bmatrix}. \tag{2.2}$$

This procedure yields a log unnormalized density for the angles given as

$$\log p^*(\boldsymbol{\phi}|\boldsymbol{m}, \boldsymbol{\Sigma}) = -\frac{1}{2} \left(\begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} - \boldsymbol{m} \right)^\top \boldsymbol{\Sigma}^{-1} \left(\begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} - \boldsymbol{m} \right). \tag{2.3}$$

Equation (2.3) admits multiple simplifications. For example, if we define $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1}\boldsymbol{m}$ and expand the quadratic term, Equation (2.3) becomes

$$\log p^*(\boldsymbol{\phi}|\boldsymbol{v}, \boldsymbol{\Sigma}) = \boldsymbol{v}^\top \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} - \frac{1}{2} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}. \tag{2.4}$$

By expressing $\boldsymbol{v}$ in polar coordinates, the linear term of the density in Equation (2.4) attains a von Mises form, i.e.

$$\begin{aligned} \boldsymbol{v}^\top \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} &= \left(\begin{bmatrix} \cos\boldsymbol{\mu} \\ \sin\boldsymbol{\mu} \end{bmatrix}^\top \begin{bmatrix} \operatorname{diag}(\boldsymbol{\kappa}) & 0 \\ 0 & \operatorname{diag}(\boldsymbol{\kappa}) \end{bmatrix} \right)^\top \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \\ &= \sum_{\mathrm{n}=1}^{\mathrm{n}} \kappa_\mathrm{n}(\cos\mu_\mathrm{n}\cos\phi_\mathrm{n} + \sin\mu_\mathrm{n}\sin\phi_\mathrm{n}) \\ &= \boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi} - \boldsymbol{\mu}). \end{aligned} \tag{2.5}$$

As in the univariate von Mises case, the parameters $\boldsymbol{\kappa}$ will represent the concentration parameters for each $\boldsymbol{\phi}$. The $\boldsymbol{\mu}$ are also interpreted as in the von Mises case, that is, they are location parameters on which the $\boldsymbol{\phi}$ will be concentrated, dubbing as a mean angle vector.

Using this result into Equation (2.4) gives rise to the model

$$p(\boldsymbol{\phi}|\boldsymbol{\mu},\boldsymbol{\kappa},\boldsymbol{\Sigma}) \propto \exp\left\{ \boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi}-\boldsymbol{\mu}) - \frac{1}{2} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \right\} \tag{2.6}$$

which we term the multivariate Generalised von Mises and represent by $m\mathcal{G}v\mathcal{M}$. This density's name arises from the property that all of its one-dimensional conditionals are Generalised von Mises distributed, as it will be shown later in Section 2.2. An example of this distribution is shown in Figure 2.1.

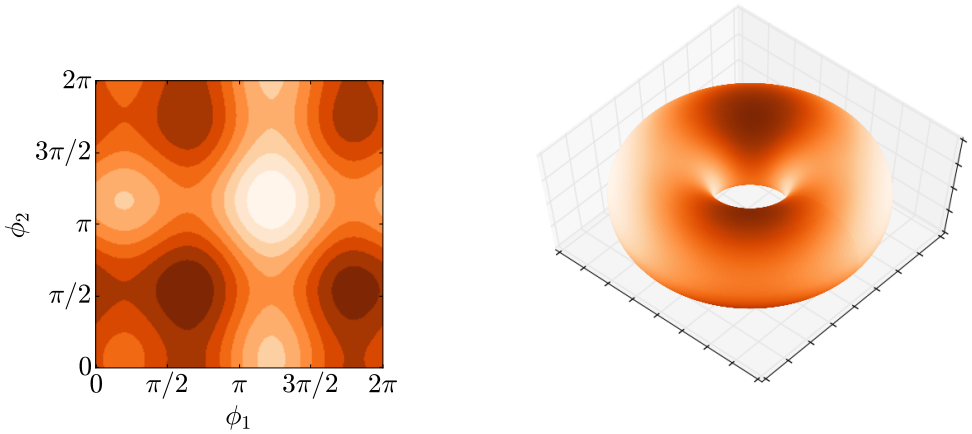


Fig. 2.1 Example of a two-dimensional $m\mathcal{G}v\mathcal{M}$ distribution with four modes plotted in the $[0, 2\pi) \times [0, 2\pi)$ plane and as a torus. Darker tones denote high probability zones, while lighter tones indicate low-probability regions.

The $m\mathcal{G}v\mathcal{M}$ in Equation (2.6) is over-parametrised; it possesses $3\mathrm{N}+\mathrm{N}^2$ parameters while drawing on trigonometric identities it can be expressed with fewer parameters. To show this, we express the quadratic term of Equation (2.6) as a block matrix of N by N matrices $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{H}$ such that

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{F} & \mathbf{G} \\ \mathbf{G}^\top & \mathbf{H} \end{bmatrix}. \tag{2.7}$$

Using the inverse-covariance form in Equation (2.7), the quadratic term can be expanded as

$$\begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} = \cos\boldsymbol{\phi}^\top \mathbf{F} \cos\boldsymbol{\phi} + 2\cos\boldsymbol{\phi}^\top \mathbf{G} \sin\boldsymbol{\phi} + \sin\boldsymbol{\phi}^\top \mathbf{H} \sin\boldsymbol{\phi}. \tag{2.8}$$

Writing the quadratic form of Equation (2.6) as Equation (2.8) allows the application of product-to-sum trigonometric identities. These identities can pose the RHS of Equation (2.8) as the sums

$$\begin{aligned} \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \Bigg[& \frac{\mathbf{F}_{\mathrm{n,j}} + \mathbf{H}_{\mathrm{n,j}}}{2} \cos(\phi_{\mathrm{n}} - \phi_{\mathrm{j}}) + \frac{\mathbf{F}_{\mathrm{n,j}} - \mathbf{H}_{\mathrm{n,j}}}{2} \cos(\phi_{\mathrm{n}} + \phi_{\mathrm{j}}) \\ & - \mathbf{G}_{\mathrm{n,j}} \sin(\phi_{\mathrm{n}} - \phi_{\mathrm{j}}) + \mathbf{G}_{\mathrm{n,j}} \sin(\phi_{\mathrm{n}} + \phi_{\mathrm{j}}) \Bigg]. \end{aligned} \tag{2.9}$$

If we further define the quantities $\mathbf{A}$, $\boldsymbol{\Omega}$, $\mathbf{B}$ and $\boldsymbol{\Upsilon}$ from $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{H}$ as

$$\mathbf{A}_{\mathrm{n,j}} = \mathrm{abs}\left({}^1\!/_2(\mathbf{F}_{\mathrm{n,j}} + \mathbf{H}_{\mathrm{n,j}}) - i\mathbf{G}_{\mathrm{n,j}} \right), \tag{2.10}$$

$$\boldsymbol{\Omega}_{\mathrm{n,j}} = \mathrm{ang}\left({}^1\!/_2(\mathbf{F}_{\mathrm{n,j}} + \mathbf{H}_{\mathrm{n,j}}) - i\mathbf{G}_{\mathrm{n,j}} \right), \tag{2.11}$$

$$\mathbf{B}_{\mathrm{n,j}} = \mathrm{abs}\left({}^1\!/_2(\mathbf{F}_{\mathrm{n,j}} - \mathbf{H}_{\mathrm{n,j}}) + i\mathbf{G}_{\mathrm{n,j}} \right), \text{ and} \tag{2.12}$$

$$\boldsymbol{\Upsilon}_{\mathrm{n,j}} = \mathrm{ang}\left({}^1\!/_2(\mathbf{F}_{\mathrm{n,j}} - \mathbf{H}_{\mathrm{n,j}}) + i\mathbf{G}_{\mathrm{n,j}} \right) \tag{2.13}$$

the RHS of Equation (2.8) attains its minimal form

$$\Big[\mathbf{A}_{\mathrm{n,j}} \cos(\phi_{\mathrm{n}} - \phi_{\mathrm{j}} - \boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}} \cos(\phi_{\mathrm{n}} + \phi_{\mathrm{j}} - \boldsymbol{\Upsilon}_{\mathrm{n,j}}) \Big]. \tag{2.14}$$

Drawing on the definitions presented in Equation (2.7), and from Equation (2.10) to Equation (2.13), the multivariate Generalised von Mises as presented in Equation (2.6)

can be written in its minimal form

$$m\mathcal{G}v\mathcal{M}(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\kappa},\mathbf{A},\mathbf{B},\boldsymbol{\Omega},\boldsymbol{\Upsilon}) \propto \exp\Big\{\boldsymbol{\kappa}^{\top}\cos(\boldsymbol{\phi}-\boldsymbol{\mu}) - \frac{1}{2}\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\Big[\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Big]\Big\} \tag{2.15}$$

with only $2\mathrm{N}^2$ parameters.

2.1.1 Generative view of the $m\mathcal{G}v\mathcal{M}$

Another way of expressing the $m\mathcal{G}v\mathcal{M}$ is through defining a generative model, that is, defining a series of conditional sampling procedures and operations on the samples attained at each step such that the obtained samples follow the distribution of interest. Under this view, the $m\mathcal{G}v\mathcal{M}$ construction outlined in this chapter can be explored by denoting the transformations through Dirac Delta functions as degenerate distributions.

To discuss this view in a concrete form, remember that the $m\mathcal{G}v\mathcal{M}$ arises from an arbitrary 2N-variate Gaussian distribution by both expressing the distribution in terms of polar components and conditioning all radial components to unity. This process can be encoded using Dirac delta distributions if we define

$$p(\mathsf{x}) = \mathcal{N}(\mathsf{x};\boldsymbol{m},\boldsymbol{\Sigma}) \tag{2.16}$$

$$p(\mathsf{z}|\mathsf{x}) = \prod_{\mathrm{n}=1}^{\mathrm{N}}\delta\left(\mathsf{z}_{\mathrm{n}}^2+\mathsf{z}_{\mathrm{N+n}}^2-1\right)\delta\left(\mathsf{z}_{\mathrm{n}}-\mathsf{x}_{\mathrm{n}}\right)\delta\left(\mathsf{z}_{\mathrm{n+N}}-\mathsf{x}_{\mathrm{N+n}}\right) \tag{2.17}$$

$$p(\boldsymbol{\phi}|\mathsf{z}) = \prod_{\mathrm{n}=1}^{\mathrm{N}}\delta(\cos\phi_{\mathrm{n}}-\mathsf{z}_{\mathrm{n}})\delta(\sin\phi_{\mathrm{n}}-\mathsf{z}_{\mathrm{N+n}}), \tag{2.18}$$

and recall that the marginal distribution for $\boldsymbol{\phi}$ can be recovered from the joint model based on Equation (2.16) to Equation (2.16) marginalising both x and z, i.e.

$$p(\boldsymbol{\phi}) = \int p(\boldsymbol{\phi}|\mathsf{z})p(\mathsf{z}|\mathsf{x})p(\mathsf{x})\mathrm{d}\mathsf{x}\mathrm{d}\mathsf{z}. \tag{2.19}$$

$$= \int p(\boldsymbol{\phi}|\mathsf{z})p(\mathsf{z})\mathrm{d}\mathsf{z}. \tag{2.20}$$

The reasoning behind the choice of functions for the model just presented benefits from further explanation. As with the standard $m\mathcal{G}v\mathcal{M}$ construction, we start with the Gaussian states in Equation (2.16). Then, Equation (2.17) introduces auxiliary variables z to constrain x to the unit circle. The conditioning step is performed using the sifting property when integrating Dirac Delta functions. Equation (2.18) then uses the same sifting property as a proxy for the polar transformation.

Posing the multivariate Generalised von Mises as a generative process can be useful for deriving approximate inference methods as described in Chapter 5.

2.1.2 Related models

The multivariate Generalised von Mises relates to several other distributions. For example, relationships between circular and directional distributions are not unusual and the $m\mathcal{G}v\mathcal{M}$ finds in the Matrix Fisher-Bingham distribution its generalisation to the directional case. In particular, the form presented in Equation (2.4) showcases the $m\mathcal{G}v\mathcal{M}$ as a Matrix Fisher-Bingham concentration vector $\boldsymbol{v}$ and matrix $\boldsymbol{\Sigma}^{-1}$. Another distribution which the $m\mathcal{G}v\mathcal{M}$ is directly related to is a bivariate distribution that generalised the von Mises-Fisher distribution by Mardia (1975b),

$$p(\phi_1, \phi_2) \propto \exp\Big\{\kappa_1 \cos(\phi_1 - \mu_1) + \kappa_2 \cos(\phi_2 - \mu_2) + a \cos\phi_1 \cos\phi_2 + b \sin\phi_1 \cos\phi_2 + c \cos\phi_1 \sin\phi_2 + d \sin\phi_1 \sin\phi_2\Big\} \tag{2.21}$$

This distribution is exactly the multivariate Generalised von Mises for two variables.

Assuming additional structures over the $m\mathcal{G}v\mathcal{M}$ parameters also leads to other distributions. For example, a special case defines the covariance in Equation (2.7) as a block-diagonal and obtained using the Kronecker product of the identity matrix in $\mathbb{R}^2$ with a N $\times$ N symmetric matrix $\mathbf{W}$, that is

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I}_{2\times 2} \otimes \mathbf{W} = \begin{bmatrix} \mathbf{W} & 0 \\ 0 & \mathbf{W} \end{bmatrix}. \tag{2.22}$$

The symmetry that Equation (2.22) establishes for sine and cosine quadratic terms induces multiple simplifications. For example, the fundamental trigonometric identity, $\cos^2\phi + \sin^2\phi = 1$, dictates that the diagonal of the matrix $\mathbf{W}$ should not impact the density shape. These diagonal values collapse to a multiplying constant which could be removed because of the density's normalisation. Given that the diagonal values of $\mathbf{W}$ do not alter the distribution, we can additionally require the trace or diagonal of $\mathbf{W}$ to be 0 as they are not identifiable.

The diagonal symmetric structure in Equation (2.22) also induces multiple simplifications in the parameter definitions from Equation (2.10) to Equation (2.12). More

precisely, the $m\mathcal{G}v\mathcal{M}$ parameters $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Upsilon}$ become

$$\mathbf{A}_{\mathrm{n,j}} = \mathrm{abs}\left(\frac{\mathbf{W}_{\mathrm{n,j}} + \mathbf{W}_{\mathrm{n,j}}}{2} - 0i\right) = \mathbf{W}, \tag{2.23}$$

$$\mathbf{B}_{\mathrm{n,j}} = \mathrm{abs}\left(\frac{\mathbf{W}_{\mathrm{n,j}} - \mathbf{W}_{\mathrm{n,j}}}{2} + 0i\right) = 0. \tag{2.24}$$

$$\boldsymbol{\Omega}_{\mathrm{n,j}} = \mathrm{ang}\left(\frac{\mathbf{W}_{\mathrm{n,j}} - \mathbf{W}_{\mathrm{n,j}}}{2} + 0i\right) = 0, \text{ and} \tag{2.25}$$

$$\boldsymbol{\Upsilon}_{\mathrm{n,j}} = \mathrm{ang}\left(\frac{\mathbf{W}_{\mathrm{n,j}} + \mathbf{W}_{\mathrm{n,j}}}{2} - 0i\right) = 0. \tag{2.26}$$

Combining these simplifications with Equation (2.15) form a density we introduce under the name of Toroidal Normal ($\mathcal{TN}$). This distribution is defined as

$$\mathcal{TN}(\boldsymbol{\phi}; \boldsymbol{\kappa}, \boldsymbol{\mu}, \mathbf{W}) \propto \exp\left\{\boldsymbol{\kappa}^{\top}\cos(\boldsymbol{\phi} - \boldsymbol{\mu}) - \frac{1}{2}\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left[\mathbf{W}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}} - \phi_{\mathrm{j}})\right]\right\} \tag{2.27}$$

and an example of which is shown in Figure 2.2.

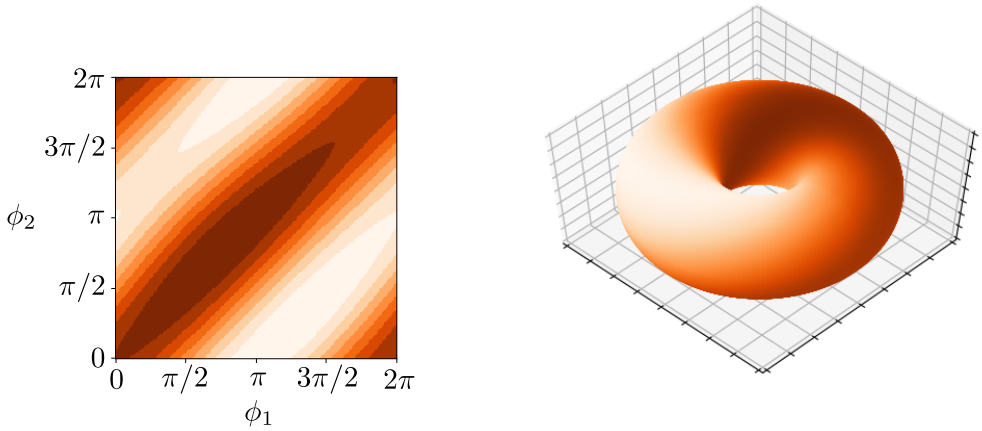


Fig. 2.2 Example of a two-dimensional $\mathcal{TN}$ plotted in the $[0, 2\pi) \times [0, 2\pi)$ plane and a torus, a unimodal distribution which is a special case of the $m\mathcal{G}v\mathcal{M}$. Darker tones denote high probability zones, while lighter tones indicate low-probability regions.

Here we remark that the $\mathcal{TN}$ corresponds to a multivariate Generalised von Mises distribution whose conditionals are all von Mises-distributed as shown in Section 2.2, but it differs from the Mardia's multivariate von Mises distribution (Mardia et al., 2008). To see this difference, recall that the $mv\mathcal{M}$ is given as

$$mv\mathcal{M}(\boldsymbol{\phi}; \boldsymbol{\kappa}, \mathbf{W}) \propto \exp\left\{\boldsymbol{\kappa}^{\top}\cos(\boldsymbol{\phi}) - \frac{1}{2}\sin\boldsymbol{\phi}^{\top}\mathbf{W}\sin\boldsymbol{\phi}\right\}, \tag{2.28}$$

while re-expressing the cosine differences in Equation (2.27) produces the distribution

$$\mathcal{TN}(\boldsymbol{\phi}) \propto \exp\Big\{\boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi}-\boldsymbol{\mu}) - \frac{1}{2}\Big[\cos\boldsymbol{\phi}^\top \mathbf{W}\cos\boldsymbol{\phi} + \sin\boldsymbol{\phi}^\top \mathbf{W}\sin\boldsymbol{\phi}\Big]\Big\}. \tag{2.29}$$

The differences between the $m\upsilon\mathcal{M}$ and the $\mathcal{TN}$ showcase that the Toroidal Normal is the most-general distribution derived from the $m\mathcal{G}\upsilon\mathcal{M}$ whose conditionals are all von Mises distributed. Consequently, the $\mathcal{TN}$ is the most general symmetric, unimodal distribution on the hyper-torus that stems from a multivariate Gaussian distribution.

2.1.3 Higher order extensions

As with the higher order Generalised von Mises, the $m\mathcal{G}\upsilon\mathcal{M}$ can also be expanded to include D cosine harmonics if its genesis from a Gaussian distribution is disregarded. In this case, an $m\mathcal{G}\upsilon\mathcal{M}$ of order D can be defined as

$$\begin{aligned} m\mathcal{G}\upsilon\mathcal{M}_{\mathrm{D}}(\boldsymbol{\phi};\mathbf{M},\mathbf{K},\mathbf{A},\mathbf{B},\boldsymbol{\Omega},\boldsymbol{\Upsilon}) \propto \exp\Bigg\{ \sum_{\mathrm{n}=1}^{\mathrm{N}}\Bigg[\sum_{\mathrm{d}=1}^{\mathrm{D}} \mathbf{K}_{\mathrm{n,d}}\cos(\mathrm{d}(\phi_{\mathrm{d}}-\mathbf{M}_{\mathrm{d,n}})) \\ + \frac{1}{2}\sum_{\mathrm{j}=1}^{\mathrm{N}} \mathbf{A}_{i,\mathrm{j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Bigg]\Bigg\} \end{aligned} \tag{2.30}$$

As the further harmonics are added to the $m\mathcal{G}\upsilon\mathcal{M}$, the distribution can represent more modes for each variable. Hence, it becomes more flexible and can express behaviours on the hyper-torus and can be linked to Fourier-based shape modelling. An example of how Fourier-based modelling can be applied is given by Li et al. (2008) in a biological context.

However, the distribution loses its link to the multivariate Gaussian, since higher order moments are also constrained. This higher order moment dependency can be evidenced by invoking the general multi-angle formulae,

$$\cos\mathrm{D}\phi = \sum_{\mathrm{d}=0}^{\mathrm{D}/2}(-1)^{\mathrm{d}}\binom{\mathrm{D}}{2\mathrm{d}}\cos^{\mathrm{D}-2\mathrm{d}}\phi\cdot\sin^{2\mathrm{d}}\phi \tag{2.31}$$

$$\sin\mathrm{D}\phi = \sum_{\mathrm{d}=0}^{\mathrm{D}/2}(-1)^{\mathrm{d}}\binom{\mathrm{D}}{2\mathrm{d}+1}\cos^{\mathrm{D}-2\mathrm{d}-1}\phi\cdot\sin^{2\mathrm{d}}\phi, \tag{2.32}$$

which establishes the correspondence between a higher harmonic and a polynomial in cosine and sine functions.

2.2 Relevant properties of the $m\mathcal{G}v\mathcal{M}$ and $\mathcal{TN}$ for inference and modelling

The previous section introduced the multivariate Generalised von Mises and a particular case of the $m\mathcal{G}v\mathcal{M}$, the Toroidal Gaussian. In this section, we present properties of these models that are useful for modelling in applications and deriving special models. In particular, this section explores some information theoretical properties, the distribution of the model's conditionals and its modes. We stress that unless explicitly noted, all the results outlined in this section were not previously known for the arbitrary dimensional case.

2.2.1 The $m\mathcal{G}v\mathcal{M}$ and $\mathcal{TN}$ are maximum entropy distributions

In both applied or theoretical contexts, it is important to make as few assumptions as possible about the modelled phenomena or the derived result. In inference, this translates to choosing a prior distribution that embeds the fewest possible assumptions, i.e. it is non-informative.

Drawing on previous connections between information theory and statistical mechanics (Jaynes, 1957a,b), Jaynes (1968) argued that distributions which maximise information entropy correspond to the ones that incorporate the fewest prior assumptions on a given problem. Hence, such distributions should be chosen as priors for Bayesian inference. Examples of distributions that maximise information entropy feature many members of the exponential family, including the multivariate Gaussian.

The multivariate Generalised von Mises maximises the information entropy if the data lives in a N-dimensional torus with defined first and second moments. Likewise, the Toroidal Gaussian also maximises the entropy in the N-dimensional torus provided we assume the particular covariance structure of the $m\mathcal{G}v\mathcal{M}$ that induces a $\mathcal{TN}$. Therefore, these distribution should be considered when selecting priors for inference in the hyper-torus.

To verify this property, we formalise the problem of finding the distribution that maximises the entropy subject to the moment and topological constraints (i.e. the

distribution must live on the surface of a hyper-torus) as the problem

$$\begin{aligned} \underset{p}{\mathsf{maximise}} \; & -\int p(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathsf{x} \\ \mathsf{subject\ to} \; & \langle \mathbf{x} \rangle_{p(\mathbf{x})} = \boldsymbol{m}, \\ & \left\langle \mathbf{x}\mathbf{x}^\top \right\rangle_{p(\mathbf{x})} = \boldsymbol{\Sigma}, \\ & \mathsf{x}_{\mathrm{n}}^2 + \mathsf{x}_{\mathrm{N+n}}^2 = 1, \quad \mathrm{n} = 1, \ldots, \mathrm{N}, \\ & \int p(\mathbf{x}) \mathrm{d}\mathbf{x} = 1 \end{aligned} \tag{2.33}$$

where $\langle \cdot \rangle_{p(\mathsf{x})}$ denotes the expectation of the quantities within the bracket taken with respect to $p(\mathsf{x})$.

The problem in Equation (2.33) can be recast in terms of polar coordinates with unit radial component. Under this representation, Equation (2.33) becomes

$$\begin{aligned} \underset{p}{\mathsf{minimise}} \; & \int p(\boldsymbol{\phi}) \log p(\boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi} \\ \mathsf{subject\ to} \; & \left\langle \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix} \right\rangle_{p(\mathbf{x})} = \boldsymbol{m}, \\ & \left\langle \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix} \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix}^\top \right\rangle_{p(\mathbf{x})} = \boldsymbol{\Sigma}. \\ & \int p(\boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi} = 1. \end{aligned} \tag{2.34}$$

To obtain the optimal distribution p, we write the Lagrangian of Equation (2.34) and utilise calculus of variations. In particular, the first order optimality criterion imposes that an optimal p should satisfy

$$\begin{aligned} \frac{\delta \mathcal{L}}{\delta p} = & -\log p(\boldsymbol{\phi}) + \boldsymbol{\lambda}^\top \left(\left\langle \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix} \right\rangle_{p(\mathbf{x})} - \boldsymbol{m} \right) \\ & - \boldsymbol{\eta}^\top \mathrm{vec} \left(\left\langle \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix} \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix}^\top \right\rangle_{p(\mathbf{x})} - \boldsymbol{\Sigma} \right) \\ & - \boldsymbol{\nu}^\top \left(\int p(\boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi} - 1 \right) = 0. \end{aligned} \tag{2.35}$$

Equation (2.35) result implies that the optimal p will have the form

$$p(\boldsymbol{\phi}) \propto \exp\left\{ \boldsymbol{\lambda}^\top \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} - \frac{1}{2} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}^\top \text{mat}(\boldsymbol{\eta}) \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \right\} \tag{2.36}$$

where $\text{mat}(\boldsymbol{\eta})$ is the representation of $\boldsymbol{\eta}$ as $2\text{N} \times 2\text{N}$ matrix.

Equation (2.36) can be identified with the over-parametrised form of the $m\mathcal{G}v\mathcal{M}$ distribution and is the maximum entropy distribution with moments constrained to the unit circle. Furthermore, it is trivial to verify that the Toroidal Normal is also a maximum entropy distribution once $\boldsymbol{\Sigma}$ is constrained to have the structure that induces a $\mathcal{TN}$ from a $m\mathcal{G}v\mathcal{M}$.

2.2.2 Conditional and marginal distributions

Other properties of interest for modelling and inference refer to the distributions that can be obtained from subsets of a random vector. A concrete example comes from time series modelling, where it is often useful to understand what is the distribution of a variable at a particular time given the values it assumed in previous time steps. Another quantity of interest is to identify the distribution of a variable irrespective of the other variables. This section analyses such properties for $m\mathcal{G}v\mathcal{M}$ and $\mathcal{TN}$ distributions by investigating their conditional and marginal distributions.

Conditional distributions

To analyse what is the distribution of a subset of variables as in the time series example in a $m\mathcal{G}v\mathcal{M}$, we need to analyse the structure of conditional distributions. This property can be examined by we partition the indexes of $\boldsymbol{\phi}$ into two disjoint sets $\mathcal{A} = \{1, \ldots, \text{D}\}$ and $\mathcal{B} = \{\text{D}+1, \ldots, \text{N}\}$.

With the subsets $\phi_\mathcal{A}$ and $\phi_\mathcal{B}$, we can fix one of the subsets and analyse the distribution of the remaining variables. For example, if we fix the variables in $\mathcal{B}$, the

unnormalised log distribution over the distributions $\mathcal{A}$ is given as

$$\begin{aligned}\log p^*(\boldsymbol{\phi}_{\mathcal{A}}|\boldsymbol{\phi}_{\mathcal{B}}) &= \boldsymbol{\kappa}_{\mathcal{A}}^{\top}\cos(\boldsymbol{\phi}_{\mathcal{A}}-\boldsymbol{\mu}_{\mathcal{A}}) + \boldsymbol{\kappa}_{\mathcal{B}}^{\top}\cos(\boldsymbol{\phi}_{\mathcal{B}}-\boldsymbol{\mu}_{\mathcal{B}}) \\ &- \frac{1}{2}\sum_{\substack{\mathrm{n}=1\\ \mathrm{n}\in\mathcal{A}}}\sum_{\substack{\mathrm{j}=1\\ \mathrm{j}\in\mathcal{A}}}\Big[\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Big] \\ &- \frac{1}{2}\sum_{\substack{\mathrm{n}=1\\ \mathrm{n}\in\mathcal{A}}}\sum_{\substack{\mathrm{j}=1\\ \mathrm{j}\in\mathcal{B}}}\Big[\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Big] \\ &- \frac{1}{2}\sum_{\substack{\mathrm{n}=1\\ \mathrm{n}\in\mathcal{B}}}\sum_{\substack{\mathrm{j}=1\\ \mathrm{j}\in\mathcal{A}}}\Big[\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Big] \\ &- \frac{1}{2}\sum_{\substack{\mathrm{n}=1\\ \mathrm{n}\in\mathcal{B}}}\sum_{\substack{\mathrm{j}=1\\ \mathrm{j}\in\mathcal{B}}}\Big[\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Big].\end{aligned} \tag{2.37}$$

The $m\mathcal{G}v\mathcal{M}$ in Equation (2.37) admits further simplification since the variables in $\mathcal{B}$ are constant. Hence, terms which feature difference and sums of angles whose indexes lie in the set $\mathcal{B}$ could be re-expressed through a scaled cosine difference, that is, for $\mathrm{n}\in\mathcal{A}$ and $\mathrm{j}\in\mathcal{B}$,

$$\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}}) = \mathrm{r}\cos(\phi_{\mathrm{n}}-\varphi), \tag{2.38}$$

where $\mathrm{r}\in\mathbb{R}_+$ and $\varphi\in[0,2\pi)$.

Applying Equation (2.38) to Equation (2.37) and simplifying results in the log unnormalised density

$$\begin{aligned}\log p^*(\boldsymbol{\phi}_{\mathcal{A}}|\boldsymbol{\phi}_{\mathcal{B}}) &= \tilde{\boldsymbol{\kappa}}_{\mathcal{A}}^{\top}\cos(\boldsymbol{\phi}_{\mathcal{A}}-\tilde{\boldsymbol{\mu}}_{\mathcal{A}}) \\ &- \frac{1}{2}\sum_{\substack{\mathrm{n}=1\\ \mathrm{n}\in\mathcal{A}}}\sum_{\substack{\mathrm{j}=1\\ \mathrm{j}\in\mathcal{A}}}\Big[\mathbf{A}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}-\phi_{\mathrm{j}}-\boldsymbol{\Omega}_{\mathrm{n,j}}) + \mathbf{B}_{\mathrm{n,j}}\cos(\phi_{\mathrm{n}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{n,j}})\Big],\end{aligned} \tag{2.39}$$

which corresponds to a $m\mathcal{G}v\mathcal{M}$ distribution, where the mean and concentration terms fully absorb the information of the conditioned variables.

A particular case of interest occurs when the set $\mathcal{A}$ represents only one variable. In this case, the corresponding distribution for $\mathcal{A}$ is a Generalised von Mises

$$\mathcal{G}v\mathcal{M}(\phi_{\mathrm{n}}|\phi_{\neq\mathrm{n}}) \propto \exp\Big\{\tilde{\kappa}_{1,\mathrm{n}}\cos(\phi_{\mathrm{n}}-\tilde{\mu}_{1,\mathrm{n}}) + \tilde{\kappa}_{2,\mathrm{n}}\cos(2\phi_{\mathrm{n}}-2\tilde{\mu}_{2,\mathrm{n}})\Big\} \tag{2.40}$$

whose parameters can be linked to the over-parametrised form as

$$\begin{aligned}
\tilde{\kappa}_{\mathrm{n},1}\cos\tilde{\mu}_{\mathrm{n},1} &= \kappa_{1,\mathrm{n}}\cos(\mu_{1,\mathrm{n}}) \\
&\quad - \frac{1}{2}\sum_{\mathrm{j}\neq\mathrm{n}}\Big[(\boldsymbol{\Sigma}^{-1})_{\mathrm{n},\mathrm{j}}\cos(\phi_{\mathrm{j}}) + (\boldsymbol{\Sigma}^{-1})_{\mathrm{n},\mathrm{j}+\mathrm{N}}\sin(\phi_{\mathrm{j}})\Big] \\
\tilde{\kappa}_{\mathrm{n},1}\sin\tilde{\mu}_{\mathrm{n},1} &= \kappa_{1,\mathrm{n}}\sin(\mu_{1,\mathrm{n}}) \\
&\quad - \frac{1}{2}\sum_{\mathrm{j}\neq\mathrm{n}}\Big[(\boldsymbol{\Sigma}^{-1})_{\mathrm{n}+\mathrm{N},\mathrm{j}}\cos(\phi_{\mathrm{j}}) + (\boldsymbol{\Sigma}^{-1})_{\mathrm{n}+\mathrm{N},\mathrm{j}+\mathrm{N}}\sin(\phi_{\mathrm{j}})\Big] \\
\tilde{\kappa}_{\mathrm{n},2}\cos 2\tilde{\mu}_{\mathrm{n},2} &= -\frac{1}{4}\Big[(\boldsymbol{\Sigma}^{-1})_{\mathrm{n},\mathrm{n}} + (\boldsymbol{\Sigma}^{-1})_{\mathrm{n}+\mathrm{N},\mathrm{n}+\mathrm{N}}\Big] \\
\tilde{\kappa}_{\mathrm{n},2}\sin 2\tilde{\mu}_{\mathrm{n},2} &= -\frac{1}{2}(\boldsymbol{\Sigma}^{-1})_{\mathrm{n},\mathrm{n}+\mathrm{N}}.
\end{aligned} \tag{2.41}$$

Using the assumptions that simplify the multivariate Generalised von Mises into a Toroidal Normal, it is possible to show that the conditionals of a Toroidal Normal are also Toroidal Normal. That is, for a $\mathcal{TN}$-distributed $\boldsymbol{\phi}$ split into two mutually exclusive sets $\mathcal{A}$ and $\mathcal{B}$, the conditionals will follow

$$\log p^*(\boldsymbol{\phi}_{\mathcal{A}}|\boldsymbol{\phi}_{\mathcal{B}}) = \tilde{\boldsymbol{\kappa}}_{\mathcal{A}}^{\top}\cos(\boldsymbol{\phi}_{\mathcal{A}} - \tilde{\boldsymbol{\mu}}_{\mathcal{A}}) - \frac{1}{2}\sum_{\substack{\mathrm{n}=1\\ \mathrm{n}\in\mathcal{A}}}\sum_{\substack{\mathrm{j}=1\\ \mathrm{j}\in\mathcal{A}}}\Big[(\mathbf{K}^{-1})_{\mathrm{n},\mathrm{j}}\cos(\phi_{\mathrm{n}} - \phi_{\mathrm{j}})\Big]. \tag{2.42}$$

In special, when $\mathcal{A}$ contains only one index, e.g. n, the one-dimensional conditional is a von Mises variable with parameters of which are given by

$$\tilde{\kappa}\cos\tilde{\mu} = \kappa_{\mathrm{n}}\cos\mu_{\mathrm{n}} - \frac{1}{2}\sum_{\mathrm{j}=1}^{\mathrm{N}}(\mathbf{K}^{-1})_{\mathrm{n},\mathrm{j}}\cos(\phi_{\mathrm{j}}) \tag{2.43}$$

$$\tilde{\kappa}\sin\tilde{\mu} = \kappa_{\mathrm{n}}\sin\mu_{\mathrm{n}} - \frac{1}{2}\sum_{\mathrm{j}=1}^{\mathrm{N}}(\mathbf{K}^{-1})_{\mathrm{n},\mathrm{j}}\sin(\phi_{\mathrm{j}}). \tag{2.44}$$

Marginal distributions

To analyse the behaviour of a variable in isolation of other set of $m\mathcal{G}v\mathcal{M}$ variables we turn to marginal distributions. A known result from Mardia (1975b) the two-dimensional $m\mathcal{G}v\mathcal{M}$

$$\begin{aligned}
p(\phi_1,\phi_2) \propto \exp\Big\{&\kappa_1\cos(\phi_1-\mu_1) + \kappa_2\cos(\phi_2-\mu_2) + a\cos\phi_1\cos\phi_2 + \\
&b\sin\phi_1\cos\phi_2 + c\cos\phi_1\sin\phi_2 + d\sin\phi_1\sin\phi_2\Big\}
\end{aligned}$$

showed that the one-dimensional marginal of a two-dimensional $m\mathcal{G}v\mathcal{M}$ is

$$p(\phi_1) \propto \exp\Big\{\kappa_1 \cos(\phi_1 - \mu_1)\Big\} \times \mathcal{I}_0(q^{1/2}(\phi_1)) \tag{2.45}$$

where $\mathcal{I}_0$ is the modified Bessel function of first kind and order 0 and

$$\begin{aligned} q(\phi_1) =& \kappa_1^2 + 2\kappa_1(a\cos\mu_1 + b\sin\mu_1)\cos\phi_1 \\ &+ 2\kappa_1(c\cos\mu_1 + d\sin\mu_1)\sin\phi_1 + (a^2+b^2)\cos^2\phi_1 + (c^2+d^2)\sin^2\phi_1 \\ &+ (ab+cd)\cos\phi_1\sin\phi_1. \end{aligned} \tag{2.46}$$

This suffices to show that the marginals of the $m\mathcal{G}v\mathcal{M}$ are not in general $m\mathcal{G}v\mathcal{M}$-distributed. This lack of marginalisation consistency implies that the Daniell-Kolmogorov Extension Theorem (Daniell, 1919; Kolmogoroff, 1933; Rogers and Williams, 2000), typically used to prove the existence of an underlying infinite-dimensional stochastic process associated with underlying distribution, cannot be applied. This result also suggests that an approach leveraging circulas, the copula analogue for circular variables (see Section 1.2.3), should be considered if $m\mathcal{G}v\mathcal{M}$ marginals are important for the application in question.

Following the same argument outlined in this section, it is possible to show that a bivariate Toroidal Normal given as

$$p(\phi_1, \phi_2) \propto \exp\Big\{\kappa_1\cos(\phi_1 - \mu_1) + \kappa_2\cos(\phi_2 - \mu_2) + w\cos(\phi_1 - \phi_2)\Big\} \tag{2.47}$$

has one dimensional marginal

$$p(\phi_1) \propto \exp\Big\{\kappa_1 \cos(\phi_1 - \mu_1)\Big\} \times \mathcal{I}_0(q^{1/2}(\phi_1)) \tag{2.48}$$

where $\mathcal{I}_0$ is the modified Bessel function of first kind and order 0 with

$$q(\phi_1) = \kappa_1^2 + w^2 + w(2\kappa_1\cos(\cos\phi_1 - \mu_1) + \sin 2\phi_1). \tag{2.49}$$

and, therefore, is also not closed under marginalisation.

2.2.3 Modes

It is not trivial to assess the number of modes for the multivariate Generalised von Mises and the Toroidal Normal models for arbitrary parametrisations.

Since both distributions are exponential family, to establish modality, it suffices to analyse the critical points of their log-unnormalised densities and the Hessian at such points. We obtain such locations by solving

$$\frac{\partial}{\partial \phi_{\mathrm{i}}} \log p^*(\boldsymbol{\phi}) = 0, \tag{2.50}$$

where the partial derivative for the $m\mathcal{G}v\mathcal{M}$ case corresponds to

$$\begin{aligned}\frac{\partial}{\partial \phi_{\mathrm{i}}} \log p^*(\boldsymbol{\phi}) = &-\kappa_{\mathrm{i}} \sin(\phi_{\mathrm{i}} - \mu_{\mathrm{i}}) \\ &- \frac{1}{2} \sum_{\mathrm{j} \neq \mathrm{i}} [-\mathbf{A}_{\mathrm{i,j}} \sin(\phi_{\mathrm{i}} - \phi_{\mathrm{j}} - \boldsymbol{\Omega}_{\mathrm{i,j}}) - \mathbf{B}_{\mathrm{i,j}} \sin(\phi_{\mathrm{i}} + \phi_{\mathrm{j}} - \boldsymbol{\Upsilon}_{\mathrm{i,j}})] \\ &- \frac{1}{2} \sum_{\mathrm{j} \neq \mathrm{i}} [+\mathbf{A}_{\mathrm{i,j}} \sin(\phi_{\mathrm{j}} - \phi_{\mathrm{i}} - \boldsymbol{\Omega}_{\mathrm{i,j}}) - \mathbf{B}_{\mathrm{i,j}} \sin(\phi_{\mathrm{i}} + \phi_{\mathrm{j}} - \boldsymbol{\Upsilon}_{\mathrm{i,j}})] \\ = &-\kappa_{\mathrm{i}} \sin(\phi_{\mathrm{i}} - \mu_{\mathrm{i}}) + \sum_{\mathrm{j} \neq \mathrm{i}} [\mathbf{B}_{\mathrm{i,j}} \sin(\phi_{\mathrm{i}} + \phi_{\mathrm{j}} - \boldsymbol{\Upsilon}_{\mathrm{i,j}})]].\end{aligned} \tag{2.51}$$

It is possible to re-express the sums Equation (2.51) through phasor addition. This leads to the realisation that Equation (2.51) is equivalent to the imaginary part of a phasor

$$\frac{\partial}{\partial \phi_{\mathrm{i}}} \log p^*(\boldsymbol{\phi}) = \mathfrak{Im} \left\{ \eta_{\mathrm{i}} \exp\{i(\phi_{\mathrm{i}} - \nu_{\mathrm{i}})\} \right\} = 0 \tag{2.52}$$

the phasor components η_{i} and ν_{i} are given as

$$\begin{aligned}\eta_{\mathrm{i}}^2 &= \left(\kappa_{\mathrm{i}} \cos \mu_{\mathrm{i}} - \sum_{\mathrm{j} \neq \mathrm{i}} \mathbf{B}_{\mathrm{j,i}} \cos(\phi_{\mathrm{j}} + \boldsymbol{\Upsilon}_{\mathrm{j,i}}) \right)^2 + \left(\kappa_{\mathrm{i}} \sin \mu_{\mathrm{i}} - \sum_{\mathrm{j} \neq \mathrm{i}} \mathbf{B}_{\mathrm{j,i}} \sin(\phi_{\mathrm{j}} + \boldsymbol{\Upsilon}_{\mathrm{j,i}}) \right)^2 \\ \nu_{\mathrm{i}} &= \arctan \left(\frac{\kappa_{\mathrm{i}} \sin \mu_{\mathrm{i}} - \sum_{\mathrm{j} \neq \mathrm{i}} \mathbf{B}_{\mathrm{j,i}} \sin(\phi_{\mathrm{j}} + \boldsymbol{\Upsilon}_{\mathrm{j,i}})}{\kappa_{\mathrm{i}} \cos \mu_{\mathrm{i}} - \sum_{\mathrm{j} \neq \mathrm{i}} \mathbf{B}_{\mathrm{j,i}} \cos(\phi_{\mathrm{j}} + \boldsymbol{\Upsilon}_{\mathrm{j,i}})} \right).\end{aligned}$$

Equation (2.51) implies that for each angle ϕ_{i} the optima is given by $\phi_{\mathrm{i}}^\star = \pm\pi - \nu_{\mathrm{i}}$. Next, the Hessian is evaluated at these locations to determine points are modes of the distribution. The Hessian at such locations need to be positive-definite in order for the location to be a mode.

By differentiating the gradient and using phasor arithmetic, the Hessian can be shown to be defined as

$$\frac{\partial^2}{\partial\phi_{\mathrm{i}}^2}\log p^*(\boldsymbol{\phi}) = -\kappa_{\mathrm{i}}\cos(\phi_{\mathrm{i}}-\mu_{\mathrm{i}}) + \sum_{\mathrm{j}\neq\mathrm{i}}\mathbf{B}_{\mathrm{i,j}}\cos(\phi_{\mathrm{i}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{i,j}}) \tag{2.53}$$

$$= \mathfrak{Re}\left\{\tilde{\eta}_{\mathrm{i}}\exp\{\imath(\phi_{\mathrm{i}}-\tilde{\nu}_{\mathrm{i}})\}\right\} \tag{2.54}$$

$$\frac{\partial^2}{\partial\phi_{\mathrm{i}}\partial\phi_{\mathrm{j}}}\log p^*(\boldsymbol{\phi}) = \mathbf{B}_{\mathrm{i,j}}\cos(\phi_{\mathrm{i}}+\phi_{\mathrm{j}}-\boldsymbol{\Upsilon}_{\mathrm{i,j}}). \tag{2.55}$$

The analytic expressions for the Hessian entries do not suggest any special properties can be leveraged to analyse positive-definiteness at all 2^{N} $\phi_{\mathrm{i}}^\star$ locations. Therefore, the number of modes cannot be ascertain specifically and only a maximum of 2^{N} can be obtained.

For the $\mathcal{TN}$ distribution, the same reasoning can be conducted to arrive at the gradients and Hessian of the unnormalised log-$\mathcal{TN}$. These quantities are given analytically as

$$\frac{\partial}{\partial\phi_{\mathrm{i}}}\log p^*(\boldsymbol{\phi}) = -\kappa_{\mathrm{i}}\sin(\phi_{\mathrm{i}}-\mu_{\mathrm{i}}) + \sum_{\mathrm{j}=1}^{\mathrm{N}}\mathbf{K}^{-1}_{\mathrm{i,j}}\sin(\phi_{\mathrm{i}}-\phi_{\mathrm{j}})$$

$$= \mathfrak{Im}\left\{\eta_{\mathrm{i}}\exp\left\{\phi_{\mathrm{i}}-\nu_{\mathrm{i}}\right\}\right\}, \tag{2.56}$$

$$\frac{\partial^2}{\partial\phi_{\mathrm{i}}^2}\log p^*(\boldsymbol{\phi}) = -\kappa_{\mathrm{i}}\cos(\phi_{\mathrm{i}}-\mu_{\mathrm{i}}) + \sum_{\mathrm{j}=1}^{\mathrm{N}}\mathbf{K}^{-1}_{\mathrm{i,j}}\cos(\phi_{\mathrm{i}}-\phi_{\mathrm{j}})$$

$$= \mathfrak{Re}\left\{\eta_{\mathrm{i}}\exp\left\{\phi_{\mathrm{i}}-\nu_{\mathrm{i}}\right\}\right\}, \text{ and} \tag{2.57}$$

$$\frac{\partial^2}{\partial\phi_{\mathrm{i}}\partial\phi_{\mathrm{j}}}\log p^*(\boldsymbol{\phi}) = -\mathbf{K}^{-1}_{\mathrm{i,j}}\sin(\phi_{\mathrm{i}}-\phi_{\mathrm{j}}). \tag{2.58}$$

As with the $m\mathcal{G}v\mathcal{M}$ case, it is not trivial to establish positive-definiteness for the Hessian at the critical points. Hence, we can only assume that in general, the $\mathcal{TN}$ distribution will be multi-modal despite posessing unimodal conditionals.

2.3 Modelling with the multivariate Generalised von Mises

In this section we will explore the ways in which the multivariate Generalised von Mises can be used in practice for modelling. In Section 2.3.1 we discuss the role of the $m\mathcal{G}v\mathcal{M}$ as a posterior in a multitude of contexts. Then we analyse specific cases

that result in $m\mathcal{G}v\mathcal{M}$ posteriors: circular regression in Section 2.3.2 and latent variable modelling in Section 2.3.3.

2.3.1 Relationship to other distributions: Posterior and approximations

In Probabilistic Machine Learning, learning and inference in probabilistic models are often performed by direct application of Bayes rule,

$$p(\mathsf{x}|\mathsf{y}) = \frac{p(\mathsf{y}|\mathsf{x})p(\mathsf{x})}{p(\mathsf{y})} \tag{2.59}$$

where $p(\mathsf{x})$ is the prior distribution over x, $p(\mathsf{y}|\mathsf{x})$ is the likelihood of the data y and $p(\mathsf{x}|\mathsf{y})$ is the posterior distribution over x.

The multivariate Generalised von Mises arises as a posterior to many different distributions spanning both circular, directional and Euclidean spaces. Figure 2.3 exemplifies the range of models that give rise by the multivariate Generalised von Mises.

While the general Wrapped Gaussian and the general Projected Gaussian might arise as posterior densities in some cases, the models which produce such distributions are not as clear nor as broad as the ones encompassed by the $m\mathcal{G}v\mathcal{M}$. This fact can be traced to their intricate functional forms. Mardia's multivariate von Mises and the Toroidal Normal are sub-cases of the multivariate Generalised von Mises and will only be associated with a posterior distribution under exceptional cases when the likelihood terms do not induce cross-correlations between sine and cosine terms of the $m\mathcal{G}v\mathcal{M}$.

The $m\mathcal{G}v\mathcal{M}$ can also be used to approximate unimodal symmetric multivariate distributions, as the Wrapped Gaussian, and multimodal asymmetric multivariate circular distributions, as is the case for the general Projected Gaussian.

For illustrative purposes, we numerically compared approximations the general Wrapped Gaussian and multivariate von Mises approximations to a base $m\mathcal{G}v\mathcal{M}$ and $m\mathcal{G}v\mathcal{M}$ approximations to these two distributions. The approximations were obtained by numerically minimising the KL divergence between the approximating distribution and the base distributions. These experiments were conducted in a two-dimensional setting and did not include the projected Gaussian to render the computation of the normalising constants tractable by numerical integration. The resulting distributions are shown in Table 2.1. In Table 2.1, the $mv\mathcal{M}$ and the multivariate wrapped Gaussian cannot capture the multimodality and asymmetry of the $m\mathcal{G}v\mathcal{M}$. Moreover, these

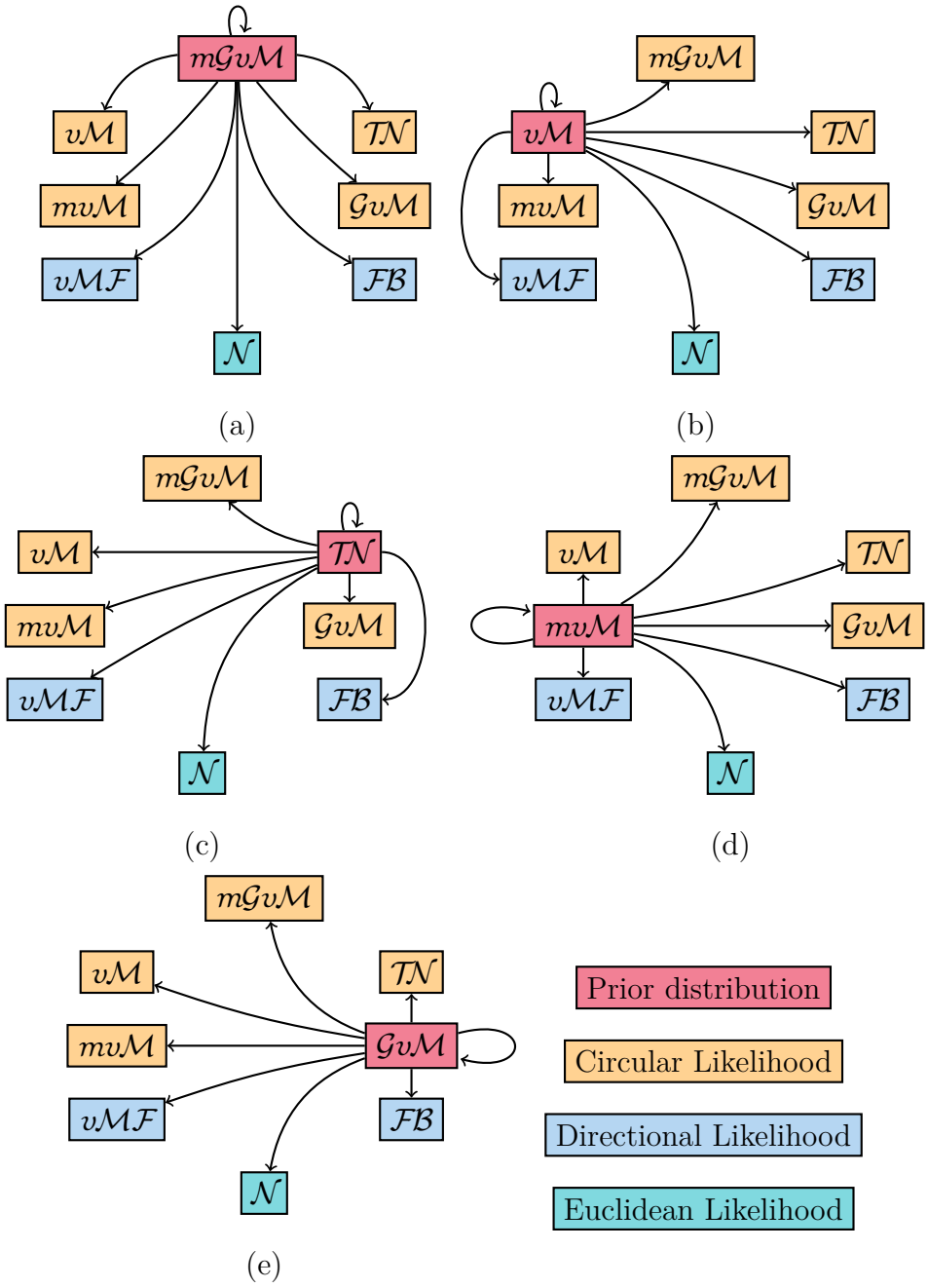


Fig. 2.3 Diagrams outlining how circular prior distributions be combined with different likelihoods yield a multivariate Generalised von Mises posterior. The source node in red corresponds to the prior, while the node the connecting arrow points to represents the likelihood. The priors under consideration are the $m\mathcal{G}v\mathcal{M}$ shown in (a), $v\mathcal{M}$ shown in (b), $\mathcal{TN}$ shown in (c), $mv\mathcal{M}$ shown in (d), and $\mathcal{G}v\mathcal{M}$ shown in (e). The conjugacy relationship shown in diagrams (a) to (e) is established through the mean of the distributions shown.

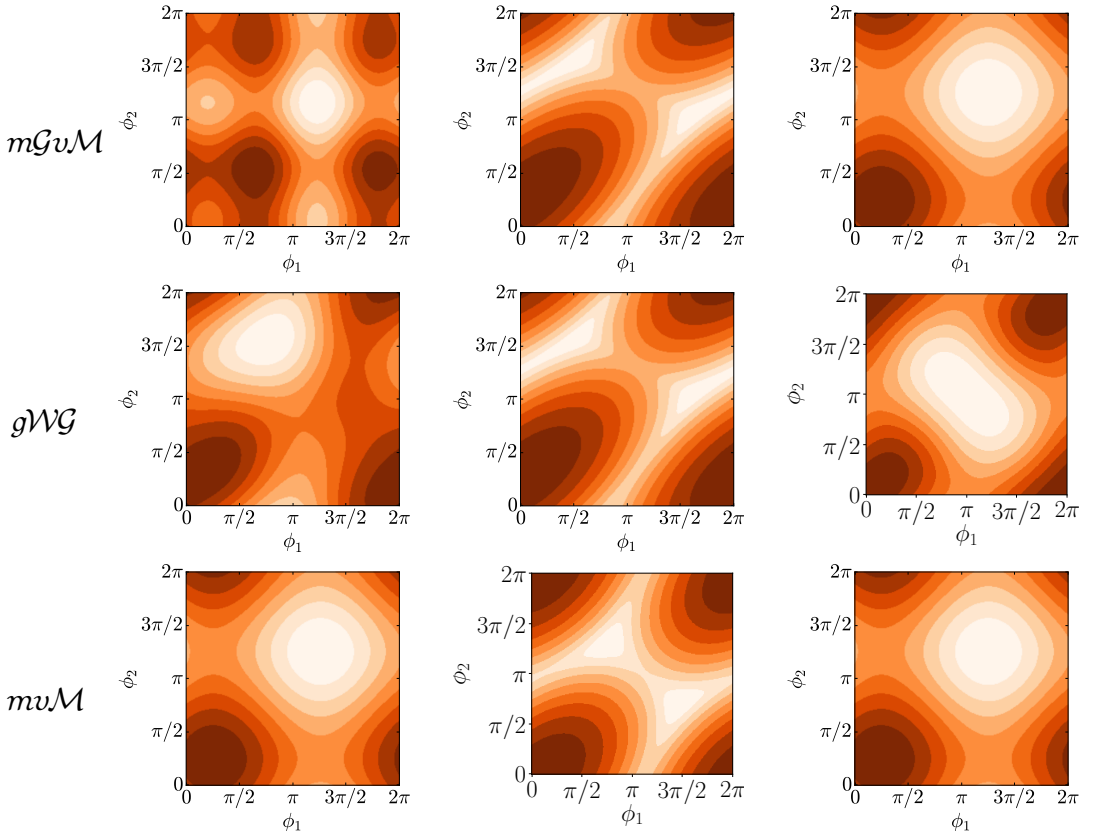


Table 2.1 Circular distributions and their approximations obtained by numerically minimising the KL divergence. Diagonal entries represent the true distribution of each row. Off-diagonal entries show the approximation of the diagonal entry using the column distribution. For example, the entry on row 1 ($m\mathcal{G}v\mathcal{M}$) and column 2 ($g\mathcal{W}\mathcal{G}$) denote the obtained $g\mathcal{W}\mathcal{G}$ approximation to the $m\mathcal{G}v\mathcal{M}$ shown in row 1, column 1.

distributions approximate the multiple modes by increasing their variance and assigning a high probability to the region of low-probability between the modes of the $m\mathcal{G}v\mathcal{M}$. On the other hand, when the $mv\mathcal{M}$ and the multivariate wrapped Gaussian are approximated by the $m\mathcal{G}v\mathcal{M}$, the $m\mathcal{G}v\mathcal{M}$ can approximate well the high-probability zones of the wrapped Gaussian and its unimodality and fully recover the $mv\mathcal{M}$.

2.3.2 Circular regression

Consider a regression problem in which a set of noisy output circular variables $\{\psi_n\}_{n=1}^N$ have been collected at a number of input locations $\{\mathbf{x}_n\}_{n=1}^N$. The treatment will apply

to inputs that can be multi-dimensional and lie in any space (e.g., they could be circular themselves). The goal is to predict circular variables $\{\psi^*_{\mathrm{m}}\}_{\mathrm{m}=1}^{\mathrm{M}}$ at unseen input points $\{\mathbf{x}^*_{\mathrm{m}}\}_{\mathrm{m}=1}^{\mathrm{M}}$.

Here we leverage the connection between the $m\mathcal{G}v\mathcal{M}$ distribution and the multivariate Gaussian to produce a powerful class of probabilistic models for this purpose based on Gaussian Processes. In what follows the outputs and inputs will be represented as vectors and matrices respectively, that is $\boldsymbol{\psi}$, $\mathbf{x}$, $\boldsymbol{\psi}^*$ and $\mathbf{x}^*$.

Therefore, to present regression with the multivariate Generalised von Mises, this section succinctly reviews the Gaussian Process regression, which informs the $m\mathcal{G}v\mathcal{M}$ regression framework. Then, the particularities of the $m\mathcal{G}v\mathcal{M}$ regression case such as the transductive characteristic of the model are analysed in detail Section 2.3.2. We remark that regardless of some limitations of the $m\mathcal{G}v\mathcal{M}$ regression when compared to Gaussian Processes, the advances promoted by the $m\mathcal{G}v\mathcal{M}$ is substantial for both Circular Statistics and Data Science as it provides an unifying regression framework for circular responses, adequate representation of angle correlations and uncertainty.

Gaussian Process regression

In standard Gaussian Process regression (Rasmussen and Williams, 2006), a multivariate Gaussian prior is placed over the underlying unknown function values at the input points

$$p(\mathbf{f}|\mathbf{x}) = \mathcal{GP}(\mathbf{f}; 0, \mathbf{K}(\mathbf{x}, \mathbf{x}')), \tag{2.60}$$

and a Gaussian noise model is assumed to produce the observations at each input location,

$$p(\mathsf{y}_{\mathrm{n}}|\mathsf{f}_{\mathrm{n}}, \mathbf{x}_{\mathrm{n}}) = \mathcal{N}(\mathsf{y}_{\mathrm{n}}; \mathsf{f}_{\mathrm{n}}, \sigma^2_{\mathsf{y}}). \tag{2.61}$$

The prior over the function values is specified using the Gaussian Process's covariance function $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ that encapsulates prior assumptions about the properties of the underlying regression function.

Prediction then involves forming the posterior predictive distribution, which also takes a Gaussian form due to conjugacy, i.e.

$$p(\mathbf{f}^*|\mathbf{y}, \mathbf{x}, \mathbf{x}^*) = \mathcal{GP}(\mathbf{f}^*; \boldsymbol{m}, \boldsymbol{\Sigma}), \tag{2.62}$$

where the mean and covariance parameters are given by

$$\boldsymbol{m} = \mathbf{K}(\mathbf{x}^*, \mathbf{x}) \left(\mathbf{K}(\mathbf{x}, \mathbf{x}) + \sigma_{\mathsf{y}}^2 \mathbf{I} \right)^{-1} \boldsymbol{y} \tag{2.63}$$

$$\boldsymbol{\Sigma} = \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}^*, \mathbf{x}) \left(\mathbf{K}(\mathbf{x}, \mathbf{x}) + \sigma_{\mathsf{y}}^2 \mathbf{I} \right)^{-1} \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*). \tag{2.64}$$

The covariance function $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ is a cornerstone in Gaussian Processes modelling. Covariance functions are positive semi-definite kernels, i.e. functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ such that

$$\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}')^{\mathsf{H}} \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') \geq 0 \tag{2.65}$$

for all square-integrable f under measure μ, for which the evaluation of k at every input pair in the set $\{(\mathbf{x}_{\mathsf{i}}, \mathbf{x}_{\mathsf{j}}) | \mathsf{i} = 1, \ldots, \mathrm{N};\ \mathsf{j} = 1, \ldots, \mathrm{N}\}$ forms a Gram matrix identified with the covariance of the input locations.

There is a large body of knowledge on modelling with covariance functions and how to unify the treatment of different inputs ranging from Euclidean and circular variables to more exotic inputs such as strings (Lodhi et al., 2001) or graphs (Gärtner et al., 2003). Covariance functions can also be used for performing model criticism (Lloyd and Ghahramani, 2015) and automatically construct interpretable models (Duvenaud et al., 2013, 2011). Here we refrain from providing an exhaustive analysis of modelling with kernels and direct the interested reader to the works of Rasmussen and Williams (2006), Murphy (2012), and Duvenaud (2014).

Interpretable models based on covariance functions can be constructed using elementary kernels and kernel operations. Elementary kernels embody the central assumptions over the underlying function, while the kernel operations allow compounding elementary kernels to provide more intricate function structures.

Examples of assumptions embodied by elementary kernels include function smoothness and locality through the order-ν Matérn kernel

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \|\mathbf{x} - \mathbf{x}'\| \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{\sqrt{2\nu}}{\ell} \|\mathbf{x} - \mathbf{x}'\| \right) \tag{2.66}$$

or the Squared Exponential kernel

$$\mathbf{K}(\mathsf{x}, \mathsf{x}') = \sigma \exp \left\{ -\frac{\|\mathsf{x} - \mathsf{x}'\|^2}{2\ell^2} \right\}. \tag{2.67}$$

The order-ν Matérn kernel implies the underlying function is less than ν-times Mean Square differentiable, whereas the Squared Exponential kernel implies in an infinitely

differentiable function. In both cases, the parameter ℓ determines the correlation strength between different points. This characteristic can be used to weight the importance of distinct inputs as proposed by Sir David Mackay (1995) by localising the parameter ℓ in his Automatic Relevance Determination kernel,

$$\mathbf{K}(\mathsf{x},\mathsf{x}') = \sigma \exp\left\{-\sum_{\mathsf{j}=1}^{\mathrm{N}} \frac{\left(\mathsf{x}_{\mathsf{j}} - \mathsf{x}'_{\mathsf{j}}\right)^2}{2\ell_{\mathsf{j}}^2}\right\}. \tag{2.68}$$

Other function behaviours such as periodicity and linear trends can be incorporated through the periodic kernel[1]

$$\mathbf{K}(\mathsf{x},\mathsf{x}') = \sigma \exp\left\{\alpha \cos\left(\beta(\mathsf{x}-\mathsf{x}')\right)\right\}, \tag{2.69}$$

and the linear kernel

$$\mathbf{K}(\mathsf{x},\mathsf{x}') = \sigma(\mathsf{x}-c)(\mathsf{x}'-c). \tag{2.70}$$

To form more complex functions, these simple kernels can be combined through kernel operations. Kernel operations are operations $\circ$ such that for $\mathbf{A}$ and $\mathbf{B}$ kernels, $\mathbf{K} = \mathbf{A} \circ \mathbf{B}$ is also a kernel. Two elementary operations include kernel (element-wise) addition

$$\mathbf{K}(\mathsf{x},\mathsf{x}') = \mathbf{A}(\mathsf{x},\mathsf{x}') \oplus \mathbf{B}(\mathsf{x},\mathsf{x}') \tag{2.71}$$

and kernel (element-wise) multiplication

$$\mathbf{K}(\mathsf{x},\mathsf{x}') = \mathbf{A}(\mathsf{x},\mathsf{x}') \odot \mathbf{B}(\mathsf{x},\mathsf{x}') \tag{2.72}$$

The function associated with the addition of kernels $\mathbf{A}$ and $\mathbf{B}$ will present a behaviour incorporating both structures of kernel $\mathbf{A}$ and $\mathbf{B}$. Multiplication of different kernels can weight the structures of kernel $\mathbf{A}$ with the structures of kernel $\mathbf{B}$, effectively convolving the two function behaviours.

[1]The periodic kernel proposed by MacKay (1998) is presented throughout the Probabilistic Machine Learning literature as a more convoluted expression with a squared sine relation.

Here, we adopt an equivalent form that can be obtained using double angle formulas. This alternative presentation emphasises three major points: (i) the similarities with the von Mises family of distributions, and (ii) remark that it is the analogue of a Squared Exponential kernel constrained to the unit circle and (iii) the cosine difference is the Euclidean distance analogue on the unit circle.

Operations are not restricted to multiplication or addition. More complex operators can be posed, such as the change point kernel,

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \sigma(\mathbf{x})\mathbf{A}(\mathbf{x}, \mathbf{x}') + (1 - \sigma(\mathbf{x}))\mathbf{B}(\mathbf{x}, \mathbf{x}'), \tag{2.73}$$

which alternates between the structures of kernels $\mathbf{A}$ and $\mathbf{B}$, or nesting kernels (Cho and Saul, 2009; Hermans and Schrauwen, 2012),

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \mathbf{A}\Big(\mathbf{B}(\mathbf{x}, \mathbf{x}'), \mathbf{C}(\mathbf{x}, \mathbf{x}')\Big), \tag{2.74}$$

whose behaviour extends that of atomic kernels into more complex structures when kernel $\mathbf{A}$ differs[2] from that of the kernels $\mathbf{B}$ and $\mathbf{C}$ (Duvenaud, 2014).

$m\mathcal{G}v\mathcal{M}$ regression

Regression using the $m\mathcal{G}v\mathcal{M}$ distribution draws upon the covariance function framework developed for Gaussian Process regression.

In $m\mathcal{G}v\mathcal{M}$ regression, an underlying function maps input locations $\mathbf{x}$ to variables $\boldsymbol{\phi}$ in the unit circle and their noisy realisations $\boldsymbol{\psi}$. A schematic overview of the similarities and differences between Gaussian Process regression and $m\mathcal{G}v\mathcal{M}$ regression is shown in Figure 2.4.

[2]Duvenaud (Duvenaud, 2014, Chapter 5) obtained the result that an infinite nesting of Squared Exponential covariance functions yields another Squared Exponential function. This result was part of a comparison of Gaussian Processes to Deep Gaussian Processes (Damianou and Lawrence, 2013), models with recursive latent GP structure,

$$p(\mathbf{y}) = \int \mathcal{GP}(\mathbf{y}|\mathbf{x}_1) \times \left[\prod_{\mathrm{j}=1}^{\mathrm{D}-1} \mathcal{GP}(\mathbf{x}_i|\mathbf{x}_{i+1})\right] \mathrm{d}\mathbf{x}_1 \times \cdots \times \mathrm{d}\mathbf{x}_{\mathrm{D}}. \tag{2.75}$$

However, we note that Duvenaud's comparison is limited only to the nesting of Squared Exponential kernels. While this analysis is a major step in establishing a comparison between 'deep' kernels and deep Gaussian Processes, we believe nested kernels deserve further analysis. In particular, a comparison of 'deep' kernel representations to those obtained from deep Gaussian Processes could be established by nesting Neural Network kernels by Williams (Williams, 1997),

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \frac{2}{\pi}\arcsin\left\{\frac{2\mathbf{x}^\top\boldsymbol{\Sigma}\mathbf{x}'}{\sqrt{\left(1 + 2\mathbf{x}^\top\boldsymbol{\Sigma}\mathbf{x}\right)\left(1 + 2\mathbf{x}'^\top\boldsymbol{\Sigma}\mathbf{x}'\right)}}\right\}. \tag{2.76}$$

Under this approach, the nonlinearity introduced by the arcsin function prevents the nested kernel from admitting simple reductions like the one found in the Squared Exponential case. Moreover, it accurately represents the behaviour of 'stacking' Neural Network layers, which better describes how deep Neural Networks are constructed. To the best of our knowledge, the aforementioned analysis has not yet been conducted.

The ability to use Gaussian Processes' covariance functions is a major motivation for using a $m\mathcal{G}v\mathcal{M}$ regression model. The introducing a function for the covariance of the over-complete $m\mathcal{G}v\mathcal{M}$ allows handling a myriad of different input variables, thus unifying the treatment of circular regression for different types of inputs. Currently, regression for circular variables relies on bespoke regression frameworks based on the input data type. Next, we motivate the simplest $m\mathcal{G}v\mathcal{M}$ regression model based

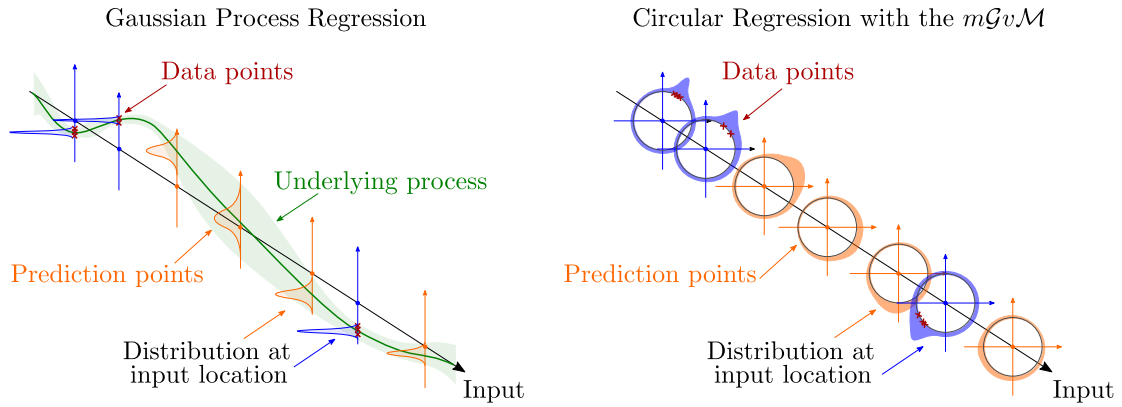


Fig. 2.4 Similarities and differences between Gaussian Process regression and Circular regression with the $m\mathcal{G}v\mathcal{M}$.

on Gaussian Process regression defining functions whose image coincides with the unit circle. The $m\mathcal{G}v\mathcal{M}$ would fit such description because it is connected with the multivariate Gaussian distribution by constraining it to live in the unit circle. This property was previously shown in Section 2.2. For example, we can assume that prior to observing data, the responses $\boldsymbol{\phi}$ at the points $\mathbf{x}$ are well represented by Toroidal Normal prior,

$$p(\boldsymbol{\phi}|\mathbf{x}) = \mathcal{TN}(\boldsymbol{\phi}; 0, 0, \mathbf{K}(\mathbf{x}, \mathbf{x}')), \tag{2.77}$$

if we do not want to impose correlations between the sine and cosine terms.

If, as in the outlined Gaussian Process model, isotropic noise is assumed for the data $\boldsymbol{\psi}$, the model obtained is a product of von Mises

$$p(\boldsymbol{\psi}|\boldsymbol{\phi}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} \upsilon\mathcal{M}(\psi_{\mathrm{n}}; \phi_{\mathrm{n}}, \kappa), \tag{2.78}$$

The posterior arising from combining Equation (2.77) and Equation (2.78) is also a Toroidal Normal

$$p(\boldsymbol{\phi}|\boldsymbol{\psi}, \mathbf{x}) = \mathcal{TN}(\boldsymbol{\phi}; \kappa_1 \times \mathbf{1}, \boldsymbol{\psi}, \mathbf{K}(\mathbf{x}, \mathbf{x}')) \tag{2.79}$$

where $\mathbf{1}$ is the vector whose all entries are one.

A more general model expanding on this particular case is consider all ranges of possible correlations between the sine and cosines of the response for the location modelled. This general construction case uses a $m\mathcal{G}v\mathcal{M}$ prior of the form

$$p(\boldsymbol{\phi}|\boldsymbol{\psi}, \mathbf{x}) = m\mathcal{G}v\mathcal{M}\left(\boldsymbol{\phi}; \boldsymbol{\eta}, \boldsymbol{\omega}(\mathbf{x}), \begin{bmatrix} \mathbf{F}(\mathbf{x}, \mathbf{x}') & \mathbf{G}(\mathbf{x}, \mathbf{x}') \\ \mathbf{G}^{\top}(\mathbf{x}, \mathbf{x}') & \mathbf{H}(\mathbf{x}, \mathbf{x}') \end{bmatrix}\right). \tag{2.80}$$

where $\boldsymbol{\omega}(\cdot)$ represents a mean-angle function and $\boldsymbol{\eta}$ is an associated concentration vector. The a more general likelihood than the von Mises distribution would be a circular distribution that allows for a correlated Normal behaviour. A distribution that is conjugate to the $m\mathcal{G}v\mathcal{M}$ and satisfies this criterion is the Generalised von Mises

$$p(\psi_{\mathrm{n}}|\phi_{\mathrm{n}}) = \mathcal{G}v\mathcal{M}(\psi_{\mathrm{n}}; \phi_{\mathrm{n}}, \kappa_1, \phi_{\mathrm{n}}, \kappa_2). \tag{2.81}$$

The model resulting from Equation (2.80) and Equation (2.81) is a $m\mathcal{G}v\mathcal{M}$ distribution,

$$p(\boldsymbol{\phi}|\boldsymbol{\psi}, \mathbf{x}) = m\mathcal{G}v\mathcal{M}\left(\boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}\right), \tag{2.82}$$

whose parameters can be defined succinctly using phasor arithmetic. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are implicitly given as

$$\alpha_i \mathrm{i}\cos(\phi_{\mathrm{i}} - \beta_{\mathbf{i}}) = \mathfrak{Re}\left\{\sum_{\mathrm{i}} \kappa_1 e^{i(\psi_{\mathrm{n}} - \phi_{\mathrm{n}})} + \eta_{\mathrm{n}} e^{i(\phi_{\mathrm{n}} - \omega(\mathsf{x}_{\mathrm{n}}))}\right\} \tag{2.83}$$

where $\mathfrak{Re}$ denotes the real part of a complex number and $i = \sqrt{-1}$. Noting that the second harmonic of the $\mathcal{G}v\mathcal{M}$ can be re-written in matrix form as

$$\begin{aligned}
\kappa_2 \cos(2\psi_{\mathrm{n}} - 2\phi_{\mathrm{n}}) &= \kappa_2 \left(\cos^2(\psi_{\mathrm{n}} - \phi_{\mathrm{n}}) - \sin^2(\psi_{\mathrm{n}} - \phi_{\mathrm{n}})\right) \\
&= \kappa_2 \Big((\cos\psi_{\mathrm{n}} \cos\phi_{\mathrm{n}} + \sin\psi_{\mathrm{n}} \sin\phi_{\mathrm{n}})^2 \\
&\qquad -(\sin\psi_{\mathrm{n}} \cos\phi_{\mathrm{n}} - \sin\phi_{\mathrm{n}} \cos\psi_{\mathrm{n}})^2\Big) \\
&= \kappa_2 \Big(\cos^2\phi_{\mathrm{n}} \cos 2\psi_{\mathrm{n}} + 2\sin\phi_{\mathrm{n}} \cos\phi_{\mathrm{n}} (1/2 \sin 2\psi_{\mathrm{n}}) \\
&\qquad + \sin^2\phi_{\mathrm{n}} (-\cos 2\psi_{\mathrm{n}})\Big) \\
&= \begin{bmatrix} \cos\phi_{\mathrm{n}} \\ \sin\phi_{\mathrm{n}} \end{bmatrix}^{\top} \begin{bmatrix} \kappa_2 \cos 2\psi_{\mathrm{n}} & \frac{\kappa_2}{2} \sin 2\psi_{\mathrm{n}} \\ \frac{\kappa_2}{2} \sin 2\psi_{\mathrm{n}} & -\kappa_2 \cos 2\psi_{\mathrm{n}} \end{bmatrix} \begin{bmatrix} \cos\phi_{\mathrm{n}} \\ \sin\phi_{\mathrm{n}} \end{bmatrix},
\end{aligned}$$

the matrix $\mathbf{W}$ in Equation (2.82) can be explicitly given as

$$\mathbf{W} = \left(\begin{bmatrix} \mathbf{F}(\mathbf{x}, \mathbf{x}') & \mathbf{G}(\mathbf{x}, \mathbf{x}') \\ \mathbf{G}^\top(\mathbf{x}, \mathbf{x}') & \mathbf{H}(\mathbf{x}, \mathbf{x}') \end{bmatrix}^{-1} + \begin{bmatrix} \operatorname{diag}(\kappa_2 \cos 2\boldsymbol{\psi}) & \operatorname{diag}(\frac{\kappa_2}{2} \sin 2\boldsymbol{\psi}) \\ \operatorname{diag}(\frac{\kappa_2}{2} \sin 2\boldsymbol{\psi}) & -\operatorname{diag}(\kappa_2 \cos 2\boldsymbol{\psi}) \end{bmatrix} \right)^{-1}.$$

In both the regression model with the Toroidal Normal and the more general one with the multivariate Generalised von Mises, inference proceeds subtly differently to that in a $\mathcal{GP}$. Both the Toroidal Gaussian and multivariate Generalised von Mises lack of consistency under marginalisation, hence the $m\mathcal{G}v\mathcal{M}$ regression should be treated as transductive model, that is, the locations for the predictions should be specified at inference time.

Another aspect where the $m\mathcal{G}v\mathcal{M}$ regression model differs from Gaussian Processes lies in the fact that the covariance function parameters—the model hyper-parameters—for Toroidal Normal priors cannot be easily learned. This issue arises because the $\mathcal{TN}$ is an intractable distribution, in the sense that its normalising constant is unknown. Therefore, it is not trivial to perform learning in these models and any approximate inference methods used for learning will need to additionally approximate the normaliser of the prior. This additional difficulty is termed double-intractability in probabilistic machine learning. This issue arises in prominent probabilistic machine learning models such as Ising models, Restricted Boltzmann Machines and Markov Random Fields, as well as in most circular and directional statistics models in higher dimensions.

2.3.3 Latent variable modelling

We motivate latent variable modelling with the $m\mathcal{G}v\mathcal{M}$ with the concrete problem of learning the motion of an articulated rigid body from noisy measurements in a Euclidean space. Articulated rigid bodies can represent a large class of physical problems including mechanical systems, human motion and molecular interactions. The dynamics of rigid bodies can also be fully described by rotations around a fixed point plus a translation and, therefore, can be succinctly represented using angles as described by Chirikjian and Kyatkin (2000).

For simplicity, we will restrict our treatment to a rigid body with N articulations on a 2-dimensional Euclidean space and rotations only, as the discussion trivially generalises to higher-dimensional spaces and translations can be incorporated through an additional linear term. Extensions for 3-dimensional models follow directly from the 2-dimensional case, which can be seen as the first step towards these more complex models.

The Euclidean components of any point on an articulated rigid body can be described using the angles between each articulation and their distances. More precisely, for an upright, counter-clockwise coordinate system, the horizontal and vertical components of a point in the N-th articulator can be written as

$$x_{\mathrm{N}} = \sum_{\mathrm{j}=1}^{\mathrm{n}} l_{\mathrm{j}} \sin(\boldsymbol{\varphi}_{\mathrm{j}}) \tag{2.84}$$

$$y_{\mathrm{N}} = -\sum_{\mathrm{j}=1}^{\mathrm{N}} l_{\mathrm{j}} \cos(\boldsymbol{\varphi}_{\mathrm{j}}), \tag{2.85}$$

where l_{j} is the length of a link j to the next link or the marker as displayed in Figure 2.5.

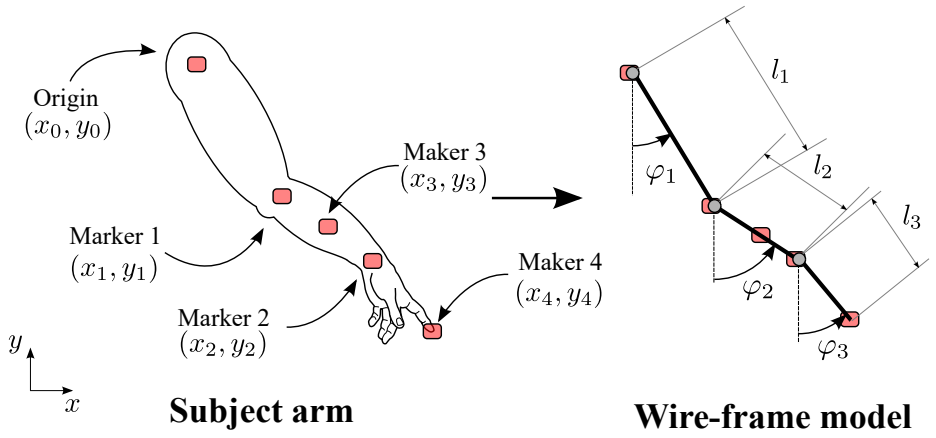


Fig. 2.5 Schematic view of the motion caption problem outlining the coordinate system and wire-frame diagram extraction.

Considering noise corruption on the measurements for each marker position, Equation (2.84) and Equation (2.85) become

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} -\mathbf{L} \\ \mathbf{L} \end{bmatrix} \begin{bmatrix} \cos \boldsymbol{\varphi} \\ \sin \boldsymbol{\varphi} \end{bmatrix} + \boldsymbol{\epsilon} \tag{2.86}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda}$ a diagonal matrix and $\mathbf{L}$ is the matrix that encodes the distances between each pair of joints (i, j) through the $l_{\mathrm{i,j}}$ such that

$$l_{\mathrm{i,j}} = \begin{cases} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} & \text{if node } i \text{ connected to node } j \\ 0 & \text{otherwise} \end{cases} \tag{2.87}$$

Without loss of generality, we can model only the variation around a fixed mean angle for each joint, i.e. $\varphi_{\mathrm{d}} = \phi_{\mathrm{d}} - \mu_{\mathrm{d}}$ which results in the general model for noisy measurements

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} -\mathbf{L} \\ \mathbf{L} \end{bmatrix} \begin{bmatrix} \cos\left(\boldsymbol{\mu} - \boldsymbol{\phi}\right) \\ \sin\left(\boldsymbol{\mu} - \boldsymbol{\phi}\right) \end{bmatrix} + \boldsymbol{\epsilon} \tag{2.88}$$

$$= \begin{bmatrix} -\mathbf{L} \\ \mathbf{L} \end{bmatrix} \mathbf{R}(\boldsymbol{\mu}) \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} + \boldsymbol{\epsilon} \tag{2.89}$$

$$= \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} + \boldsymbol{\epsilon} \tag{2.90}$$

where $\mathbf{R}(\boldsymbol{\mu})$ is a rotation matrix, $\mathbf{A}$ and $\mathbf{B}$ are distance matrices after the rotation.

The prior over the joint angles can be modelled by a multivariate Generalised von Mises. Here we take inspiration from Principal Component Analysis and use independent von Mises distributions

$$p(\boldsymbol{\phi}_{\mathrm{n,d}}) = \upsilon\mathcal{M}\left(\phi_{\mathrm{n,d}}; 0, \kappa_d\right), \tag{2.91}$$

although in some cases when performing motion capture, such as a flexed arm rotating around the shoulder joint, the limb angles may exhibit a dependence structure.

Due to conjugacy, the posterior distribution over the latent angles is a $m\mathcal{G}\upsilon\mathcal{M}$ distribution. This can be informally verified by noting that the log unnormalised priors on the latent angles $\boldsymbol{\phi}$ are linear functions of sines and cosines, while the log unnormalised likelihood is the exponential of a quadratic function in sine and cosines. This choice of distributions implies that the log unnormalised posterior being a quadratic function of sines and cosines,

$$\log p^*(\boldsymbol{\phi}|\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\kappa}(\boldsymbol{x}, \boldsymbol{y})^\top \cos\Big(\boldsymbol{\phi} - \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{y})\Big) - \frac{1}{2} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \boldsymbol{\Sigma}(\mathbf{A}, \mathbf{B}) \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \tag{2.92}$$

where

$$\boldsymbol{\Sigma}(\mathbf{A}, \mathbf{B}) = \left(\begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}^\top \boldsymbol{\Lambda}^{-1} \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \right)^{-1} \text{ and } \begin{bmatrix} \boldsymbol{\kappa}(\boldsymbol{x}, \boldsymbol{y}) \odot \sin\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{y}) \\ \boldsymbol{\kappa}(\boldsymbol{x}, \boldsymbol{y}) \odot \cos\boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix}, \tag{2.93}$$

hence, the posterior is a

$$p(\boldsymbol{\phi}|\boldsymbol{x}, \boldsymbol{y}) = m\mathcal{G}\upsilon\mathcal{M}\Big(\boldsymbol{\phi}; \boldsymbol{\kappa}(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{\mu}(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{\Sigma}(\mathbf{A}, \mathbf{B})\Big) \tag{2.94}$$

The model can be extended to treat the parameters in a Bayesian way by including sparse priors over the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ and the observation noise. A standard choice for this task is to define Automatic Relevance Detection priors (Mackay, 1995) over the columns of these matrices,

$$p(\mathbf{A}_{\mathrm{n,d}}) = \mathcal{N}(\mathbf{A}_{\mathrm{n,d}}; 0, \sigma^2_{\mathbf{A},\mathrm{d}}) \tag{2.95}$$
$$p(\mathbf{B}_{\mathrm{n,d}}) = \mathcal{N}(\mathbf{B}_{\mathrm{n,d}}; 0, \sigma^2_{\mathbf{B},\mathrm{d}}), \tag{2.96}$$

in order to perform automatic structure learning. Additional Inverse Gamma priors over $\sigma^2_{\mathbf{A},\mathrm{d}}$, $\sigma^2_{\mathbf{B},\mathrm{d}}$ and the entries of $\mathbf{\Lambda}$, i.e.

$$p(\sigma^2_{\mathbf{A},\mathrm{d}}) = \mathcal{IG}(\sigma^2_{\mathbf{A},\mathrm{d}}, \alpha_{\mathbf{A},\mathrm{d}}, \beta_{\mathbf{A},\mathrm{d}}) \tag{2.97}$$
$$p(\sigma^2_{\mathbf{B},\mathrm{d}}) = \mathcal{IG}(\sigma^2_{\mathbf{B},\mathrm{d}}, \alpha_{\mathbf{B},\mathrm{d}}, \beta_{\mathbf{B},\mathrm{d}}) \tag{2.98}$$
$$p(\mathbf{\Lambda}_{\mathrm{n,n}}) = \mathcal{IG}(\mathbf{\Lambda}_{\mathrm{n,n}}, \alpha_{\mathbf{\Lambda},\mathrm{n}}, \beta_{\mathbf{\Lambda},\mathrm{n}}), \tag{2.99}$$

complete the model as regularisers for measurement noise and the pruning of $\mathbf{A}$ and $\mathbf{B}$ entries. Unlike the regression model, the latent variable model is not doubly-intractable as all components of the model $p(\boldsymbol{\phi}, \boldsymbol{x}, \boldsymbol{y})$ have tractable normalising constants. Hence, its parameters can be learned using methods that only approximate the $m\mathcal{G}v\mathcal{M}$ posterior distribution.

The dimensionality reduction $m\mathcal{G}v\mathcal{M}$ model has constructed analogously to the Probabilistic Factor Analysis (PFA) model introduced by Tipping and Bishop (1997, 1999). The special case when the diagonal matrix of the noise is isotropic, i.e. $\mathbf{\Lambda} = \sigma^2\mathbf{I}$, coincides with Probabilistic Principal Component Analysis (PPCA).

Hence, we denote the dimensionality reduction models with diagonal $\mathbf{\Lambda}$ and isotropic $\mathbf{\Lambda}$, circular Factor Analysis (cFA) and circular Principal Component Analysis (cPCA). Further connections between these circular models and their Euclidean counterparts are explored next through limiting behaviour and geometrical analysis.

The cFA model is given by

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}; 0, \mathbf{I}) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}; \mathbf{W}\mathbf{x}, \mathbf{\Lambda}) \end{aligned} \tag{2.100}$$

where $\mathbf{W}$ is a matrix that encodes the linear mapping between hidden components $\mathbf{x} \in \mathbb{R}^{\mathrm{M}}$ and data $\mathbf{y} \in \mathbb{R}^{\mathrm{N}}$, with $\mathrm{N} > \mathrm{M}$.

If we impose that each of the latent components x_m is sinusoidal and may be parametrised by a hidden angle ϕ_m plus a phase shift φ_m, we obtain the model

$$\begin{aligned} p(\phi_\mathrm{m}) &= \mathcal{GvM}(\phi; \kappa_{1,\mathrm{m}}, \kappa_{2,\mathrm{m}}, \mu_{1,\mathrm{m}}, \mu_{2,\mathrm{m}}) \\ p(\mathsf{x}_\mathrm{m}|\phi_\mathrm{m}) &= \delta(\mathsf{x}_\mathrm{m} - \sin(\phi_\mathrm{m} - \varphi_\mathrm{m})) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}; \mathbf{W}\mathbf{x}, \mathbf{\Lambda}) \end{aligned} \tag{2.101}$$

To obtain the relation directly between the data and the hidden angle, we integrate out the latent components $\mathbf{x}$

$$p(\mathbf{y}|\boldsymbol{\phi}) = \int \delta(\mathbf{x} - \sin(\boldsymbol{\phi} - \boldsymbol{\varphi}))\mathcal{N}(\mathbf{y}; \mathbf{W}\mathbf{x}, \mathbf{\Lambda})\mathrm{d}\mathbf{x} \tag{2.102}$$

which results in the model used in the $m\mathcal{GvM}$ dimensionality reduction application.

Alternatively, it is also possible to show the limiting behaviour of the model arising from the $m\mathcal{GvM}$ dimensionality reduction application becomes the PFA model, for mean angles $\boldsymbol{\mu} \to 0$ and high concentration parameters. In this regime, the small angle approximation

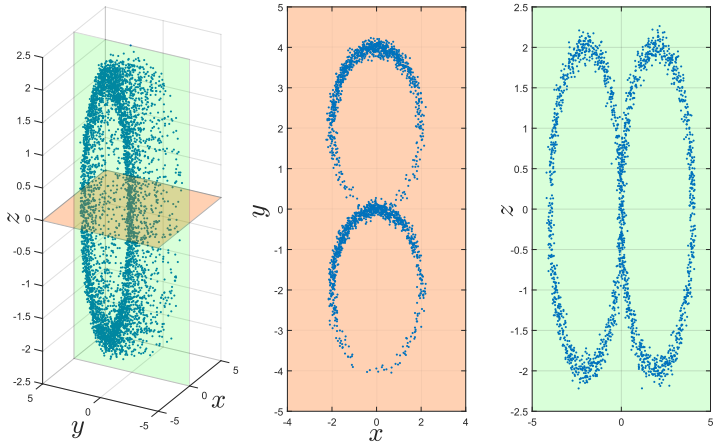
$$\sin\phi \approx \phi, \quad \cos\phi \approx 0 \tag{2.103}$$

is valid and leads to the Generalised von Mises priors simplification to

$$\begin{aligned} p(\phi) &\propto \exp\left\{\kappa_1\cos(\phi - \mu_1) + \kappa_2\cos(2(\phi - \mu_2))\right\} \\ &\propto \exp\left\{-\kappa_1\cos(\mu_2)\phi^2 + (\kappa_1\sin(\mu_1) + 2\kappa_2\sin(2\mu_2))\phi\right\} \\ &\propto \exp\left\{-\kappa_1\cos(\mu_2)\left[\phi - \frac{\kappa_1\sin(\mu_1) + 2\kappa_2\sin(2\mu_2)}{2\kappa_1\cos(\mu_2)}\right]^2\right\} \end{aligned} \tag{2.104}$$

which is proportional to a Gaussian distribution and shows that under the small angle regime, the coefficient matrix $\mathbf{A}$ is a good approximation for $\mathbf{W}$ and the model collapses to PFA.

Another connection between the dimensionality reduction with the $m\mathcal{GvM}$ and PFA may be established geometrically. While PFA describes the data through hidden hyperplanes, the lower dimensional description of the data with $m\mathcal{GvM}$ occurs through hidden toruses, as illustrated in Figure 2.6. The effect of priors in these systems is also highlighted by Figure 2.6. The mean angle and concentration of each prior impacts the distribution of mass along the direction of the angular component on the hyper-torus. High concentration values on the prior leads to dense regions around the mean angle, as presented in the middle graph of Figure 2.6 while low concentration leads to uniform mass distribution, shown in the right graph of Figure 2.6.



[tbp]

Fig. 2.6 Plots of the model $x = 2\cos\phi_1 + \epsilon$, $y = 2\sin\phi_1 + 2\cos\phi_2 + \epsilon$, $z = 2\sin\phi_2 + \epsilon$ where $\phi_1 \sim \upsilon\mathcal{M}(50, \pi/2)$ is a peaked von Mises distribution, $\phi_2 \sim \upsilon\mathcal{M}(0.1, 0)$ is an almost-uniform von Mises distribution and the noise is $\epsilon \sim \mathcal{N}(0, 0.01)$ to exemplify a 3-dimensional Cartesian data set as a function of a 2-dimensional angular space: plot of samples from the model (left), samples on the $z = 0$ plane, which is equivalent to fixing $\phi_2 = \pm\pi$ (middle), samples on the $x = 0$ plane, which is equivalent to fixing $\phi_1 = \pm\pi/2$ (right).

An analogy often used to describe this shape of the data in the PFA's hidden space is a "fuzzy pancake", as the Gaussian noise induces the shape irregularity ("fuzziness"), of the hidden plane ("pancake"). Likewise, for dimensionality reduction with the $m\mathcal{G}\upsilon\mathcal{M}$ the corresponding analogy would be a "fuzzy doughnut", as the Gaussian noise also incurs in irregularities over the surface of a "doughnut", which bears a similar shape to a torus.

2.4 Related work

The $m\mathcal{G}\upsilon\mathcal{M}$ is a generalisation of Zemel, Williams and Mozer's Directional-Unit Boltzmann Machine (DUBM) (Zemel et al., 1993), which only featured the quadratic term and its proposition stemmed purely from Boltzmann Machines, and not from conditioning a Gaussian to the unit circle. At the same time, the $m\mathcal{G}\upsilon\mathcal{M}$ is a special case of a matrix Fisher-Bingham distribution, as can be identified by inspecting Equation (2.4) and Equation (1.32).

Scholz (Scholz, 2007) proposed a circular PCA based on the circular neural network of (Kirby and Miranda, 1996). A more appropriate name such this model, however, would be a circular auto-encoder, as the model's architecture fundamentally resembles

that of two neural networks sharing a latent space with circular units shown in Figure 2.7. Furthermore, Scholz's circular PCA was not presented in a probabilistic

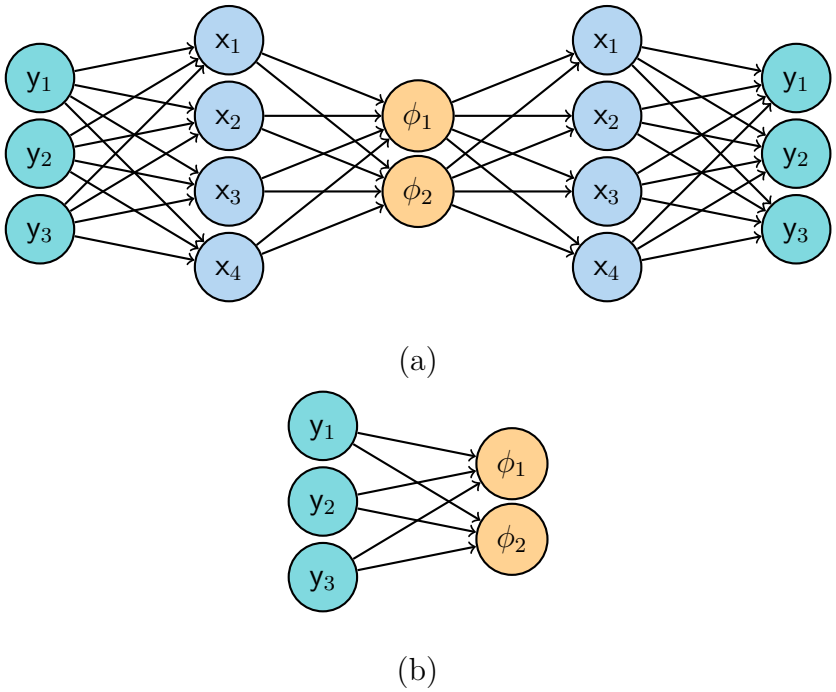


Fig. 2.7 Comparison of model architectures for dimensionality reduction with circular variables. The model by Scholz (2007) (a) uses intermediary latent Euclidean states and an auto-encoder structure to learn latent circular variables from Euclidean observations. The circular PCA outlined in this thesis (b) directly maps from observed space to latent circular variables.

framework, nor was it related to standard PCA. However, using the $m\mathcal{G}v\mathcal{M}$ it would be possible to build a probabilistic model based on this architecture, a probabilistic circular auto-encoder.

2.5 Summary

This chapter presents two major contributions: a new multivariate distribution for circular variables, and a modelling framework for circular variables based on covariance functions.

Unlike previously multivariate distributions for circular variables, the first contribution of this chapter, we introduced the Multivariate Generalised von Mises, a multivariate distribution that allows for the most general covariance structure between any two angles. We showed that besides generalising previously existing distributions,

the $m\mathcal{G}v\mathcal{M}$ has important theoretical properties including that it is a maximum entropy distribution, an exponential family member, and a posterior for a large number of models. Other properties demonstrated in this chapter for submodels of the $m\mathcal{G}v\mathcal{M}$ will be central to the algorithms outlined in Part III. More interestingly, we introduced properties regarding the marginal distributions and the process view of the distribution. These important results had not been derived previously and present novel contributions of this thesis.

The second contribution of this chapter relies on porting the covariance function modelling framework to circular variables. We outlined how this framework developed for Gaussian Processes can be incorporated into $m\mathcal{G}v\mathcal{M}$-based models for both regression and latent variable settings. To best our knowledge, the covariance function modelling framework reviewed in this chapter had not previously been applied to circular statistics problems. We also showed how regression and latent variable models with the multivariate Generalised von Mises relate to their Euclidean counterparts.

Although the multivariate Generalised von Mises lays the foundations for porting the Gaussian Process machinery to circular statistics, the $m\mathcal{G}v\mathcal{M}$ is an intractable distribution. Moreover, its reliance on covariance functions implies that the $m\mathcal{G}v\mathcal{M}$ requires careful, bespoke approximations to handle large data sets. In the next chapter, we introduce models based on transformations of the Generalised von Mises to solve these issues.

Chapter 3

Augmented representations of the $m\mathcal{G}v\mathcal{M}$

The previous chapter introduced the multivariate Generalised von Mises distribution, the maximum entropy distribution on the hyper-torus that arises as a posterior for a multitude of different probabilistic models. While this new distribution provides a basis for applying the covariance function modelling framework, the lack of a closed-form expression to evaluate the $m\mathcal{G}v\mathcal{M}$ normalising constant makes inference and learning difficult. Perhaps the most important example of where such difficulties arise is in circular regression with $m\mathcal{G}v\mathcal{M}$ priors.

These issues are not exclusive to the $m\mathcal{G}v\mathcal{M}$, similar problems are endemic to Ising models (Hubbard, 1959), Restricted Boltzmann Machines (Hinton, 1989) and Markov Random Fields (Pakman and Paninski, 2013). Inference and learning in these models is made tractable through an augmented representation cleverly designed to eliminate the functional terms responsible for generating computational difficulties. An example of such augmented representations is the Hubbard Stratonovich transformation for Ising models, which augments Ising models introducing Gaussian states to fully factorise the resulting model and eliminate the quadratic terms responsible for Ising model's intractabilities.

Drawing on these augmented representations, this chapter presents a significant contribution to help solve the difficulties associated to the $m\mathcal{G}v\mathcal{M}$ intractability: a family of augmented models for the $m\mathcal{G}v\mathcal{M}$ which are amenable to approximate inference. This family of models decouples the quadratic terms that produce intractable normalising constants for the $m\mathcal{G}v\mathcal{M}$. This characteristic is important when performing learning on a model that has a $m\mathcal{G}v\mathcal{M}$ prior, such as the circular regression model presented in Chapter 2, as the normalising constant is responsible for performing

adequate parameter regularisation. The proposed augmentation also allows building sparse models for the $m\mathcal{G}u\mathcal{M}$, which can be useful when dealing with large data sets.

3.1 Derivations for the $m\mathcal{G}u\mathcal{M}$

In this section, we explore an augmentation of the $m\mathcal{G}u\mathcal{M}$ that allows writing it as a joint model with factorised structure and known partition functions leveraging the concept of exchangeability. Then, we relate the variable augmentation to the Hubbard-Stratanovich transformation, a widely used transformation from statistical physics for calculating partition functions.

3.1.1 Exchangeablility and modelling

The notion of exchangeability can be informally explained as order independence, i.e. an array of variables $\mathsf{x}_1, \ldots, \mathsf{x}_\mathrm{N}$ can have its indices permuted without any loss of information. This concept is made precised by the celebrated fundamental result from Bernardo de Finetti (de Finetti, 1931) which was later extended by Czesław Ryll-Nardzewski (Ryll-Nardzewski, 1957). This result states that the distribution of any a fully correlated random vector $\mathbf{x}$ can be re-expressed as a latent conditional independence structure

$$p(\mathsf{x}_1, \ldots, \mathsf{x}_\mathrm{N}) = \int p(\mathsf{f}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathsf{f} = \int \left[\prod_{\mathrm{n}=1}^{\mathrm{N}} p(\mathsf{x}_\mathrm{n}|\mathsf{f})\right] p(\mathsf{f})\mathrm{d}\mathsf{f}. \tag{3.1}$$

where f is a latent variable.

Bernardo and Smith (Bernardo and Smith, 2008) argue that de Finetti's theorem on exchangeability is a central piece in hierarchical models. Indeed, a vast range of hierarchical probabilistic Machine Learning models hinge on the concept of exchangeability. Examples typically include space partitioning and clustering models such as the Chinese Restaurant Process (Aldous, 1985; Pitman, 2002), the Indian Buffet Process (Ghahramani and Griffiths, 2006), Mondrian processes and forests (Lakshminarayanan et al., 2014; Roy and Teh, 2009), Dirichlet trees (Neal, 2003) and topic models (Blei et al., 2003). However, the entire field of inducing input representations for sparse Gaussian Processes (Bui and Turner, 2014; Lee et al., 2017; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006a) can can also be posited as a direct consequence of exchangeability.

A singularly useful extension of the exchangeability property is partial exchangeability. The concept of partial exchangeability was independently derived by David Aldous (Aldous, 1981) and Douglas Hoover (Hoover, 1979), who showed that the indices of an array of variables $\mathsf{x}_1, \ldots, \mathsf{x}_\mathrm{N}$ can be compartmentalised into $\mathrm{M} \leq \mathrm{N}$ disjoint partitions $\mathcal{I}_1, \ldots, \mathcal{I}_\mathrm{M}$ where each partition has a private latent variable f_m associated with it, i.e.

$$p(\mathsf{x}_1, \ldots, \mathsf{x}_\mathrm{N}) = \int p(\mathbf{f}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{f} = \int \prod_{\mathrm{m}=1}^{\mathrm{M}} \left[p(\mathsf{f}_\mathrm{m}) \prod_{\mathrm{n}\in\mathcal{I}_\mathrm{m}} p(\mathsf{x}_\mathrm{n}|\mathsf{f}_\mathrm{m}) \right] \mathrm{d}\mathbf{f}. \tag{3.2}$$

A particular case where this notion of decoupling through augmenting an exchangeable distribution can be seen in the Hubbard-Stratonovich transformation from statistical physics. Ruslan Stratonovich (Stratonovich, 1957) proposed the transformation which was later popularised by John Hubbard (Hubbard, 1959) as a method to estimate the partition function in many–body systems, particularly for Ising models, i.e.,

$$p(\mathbf{x}) \propto \exp\left\{ \boldsymbol{a}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{W}\mathbf{x} \right\} \tag{3.3}$$

where $\mathbf{x}$ is random vector such that each entry x_i represents a particle state taking -1 or +1 as values, $\boldsymbol{a}$ is a real-valued vector and $\mathbf{W}$ is a real-valued matrix.

Stratonovich's insight can be interpreted as through the lens of exchangeability as realizing that is no ordering requirement on how the individual particle states x_i in $\mathbf{x}$ are represented in the Ising model. Hence, the states are exchangeable and it is possible to derive an expansion that decoupled the model states as in Equation (3.2). This property would allow estimating the normalising constant of the Ising model.

The the distribution over the augmentation states $\mathbf{f}$ has to be judiciously chosen to promote the decoupling described by Equation (3.2). Stratonovich noticed that the coupling in Equation (3.3) arose exclusively from the term $+\frac{1}{2}\mathbf{x}^\top\mathbf{W}\mathbf{x}$, which implied that the conditional distribution $p(\mathbf{f}|\mathbf{x})$ would need to be an exponential family distribution with so that the dependency with $\mathbf{x}$ would need to be quadratic in order to *cancel* the coupling term in the Ising model.

One natural choice for a distribution with such characteristics would be a multivariate Gaussian with a mean dependent on $\mathbf{x}$, that is,

$$\begin{aligned} p(\mathbf{f}|\mathbf{x}) &= \mathcal{N}(\mathbf{f}; \mathbf{x}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{2/\mathrm{N}} \det \boldsymbol{\Sigma}} \exp\left\{-\frac{1}{2}(\mathbf{f}-\mathbf{x})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{f}-\mathbf{x})\right\} \\ &= \frac{1}{(2\pi)^{2/\mathrm{N}} \det \boldsymbol{\Sigma}} \exp\left\{-\frac{1}{2}\mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{f} + \mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}\right\}. \end{aligned} \tag{3.4}$$

This conditional distribution of augmentation states $\mathbf{f}$ given $\mathbf{x}$ implies that the joint Ising model becomes

$$p(\mathbf{f}, \mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{f} + \mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{a}^\top \mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{W}\mathbf{x}\right\} \tag{3.5}$$

$$= \exp\left\{-\frac{1}{2}\mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{f} + (\mathbf{f}^\top \boldsymbol{\Sigma}^{-1} + \boldsymbol{a}^\top)\mathbf{x} + \frac{1}{2}\mathbf{x}^\top (\mathbf{W} - \boldsymbol{\Sigma}^{-1})\mathbf{x}\right\}. \tag{3.6}$$

From Equation (3.6) it becomes clear that to eliminate the quadratic dependency on $\mathbf{x}$, $\boldsymbol{\Sigma}^{-1}$ must be equal to $\mathbf{W}$. Assuming $\boldsymbol{\Sigma}^{-1} = \mathbf{W}$, the joint distribution over $\mathbf{x}$ and $\mathbf{f}$ reduces to

$$p(\mathbf{f}, \mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{f} + (\mathbf{f}^\top \boldsymbol{\Sigma}^{-1}\mathbf{V} + \boldsymbol{a}^\top)\mathbf{x}\right\}, \tag{3.7}$$

leading to the factorisation

$$p(\mathbf{f}, \mathbf{x}) = p(\mathbf{x}|\mathbf{f})p(\mathbf{f}) = \mathrm{Multinomial}(\mathbf{x}; \mathbf{f}^\top \boldsymbol{\Sigma}^{-1} + \boldsymbol{a}^\top) \times \mathcal{N}(\mathbf{f}; \mathbf{0}, \boldsymbol{\Sigma}) \tag{3.8}$$

for which all partition functions can be easily evaluated from closed-form expressions.

From Equation (3.2) and the Hubbard-Stratonovich transformation, it becomes clear that partial exchangeability can be employed to decouple correlated variables in the multivariate Generalised von Mises, and more generally the matrix Fisher-Bingham distribution over Stiefel manifolds. Specifically, Equation (3.1) implies that a N-dimensional multivariate Generalised von Mises $m\mathcal{G}v\mathcal{M}$ over base states $\boldsymbol{\phi}$,

$$\begin{aligned} p(\boldsymbol{\phi}) &= m\mathcal{G}v\mathcal{M}\left(\boldsymbol{\phi}; \boldsymbol{\kappa}, \boldsymbol{\mu}, \mathbf{K}\right) \\ &\propto \exp\left\{\boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi} - \boldsymbol{\mu}) - \frac{1}{2}\begin{bmatrix}\cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi}\end{bmatrix}^\top \mathbf{K}^{-1} \begin{bmatrix}\cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi}\end{bmatrix}\right\}, \end{aligned} \tag{3.9}$$

can be recast in terms of latent partitions through augmentation states $\mathbf{f}$ as

$$p(\boldsymbol{\phi}) = \int p(\boldsymbol{\phi}|\mathbf{f})p(\mathbf{f})\mathrm{d}\mathbf{f} = \int p(\mathbf{f}) \prod_{\mathrm{n}=1}^{\mathrm{N}} p(\phi_{\mathrm{n}}|\mathbf{f})\mathrm{d}\mathbf{f} \tag{3.10}$$

where we leveraged the compartmentalised structure promoted by Equation (3.2) to further induce independence relationships between the each of the base states ϕ_{n}.

Since the intractability of the $m\mathcal{G}v\mathcal{M}$ distribution stems from the quadratic relationship between the trigonometric vector $\mathbf{v}(\boldsymbol{\phi}) = [\cos\boldsymbol{\phi}^\top, \sin\boldsymbol{\phi}^\top]^\top$, a clear parallel between the Hubbard-Stratonovich Transformation for Ising models and the $m\mathcal{G}v\mathcal{M}$ by can be established. Next, drawing on the insights of the Hubbard-Stratonovich transformation, we will derive a family of augmented representations for the $m\mathcal{G}v\mathcal{M}$ that decouple the circular states.

As in the Hubbard-Stratonovich for Ising models, we want to eliminate the quadratic function in the trigonometric vector. Hence, once more a multivariate Gaussian with vector of interest as a mean is a suitable choice for the distribution of the augmentation conditional, that is

$$\begin{aligned} p(\mathbf{f}|\boldsymbol{\phi}) &= \mathcal{N}(\mathbf{f}; \mathbf{v}(\boldsymbol{\phi}), \boldsymbol{\Sigma}) && (3.11)\\ &= \frac{1}{(2\pi)^{2/\mathrm{N}}\det\boldsymbol{\Sigma}} \exp\left\{-\frac{1}{2}(\mathbf{f}-\mathbf{v}(\boldsymbol{\phi}))^\top\boldsymbol{\Sigma}^{-1}(\mathbf{f}-\mathbf{v}(\boldsymbol{\phi}))\right\} && (3.12)\\ &= \frac{1}{(2\pi)^{2/\mathrm{N}}\det\boldsymbol{\Sigma}} \exp\left\{-\frac{1}{2}\mathbf{f}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f} + \mathbf{f}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}(\boldsymbol{\phi}) - \frac{1}{2}\mathbf{v}(\boldsymbol{\phi})^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}(\boldsymbol{\phi})\right\}. && (3.13) \end{aligned}$$

Using the augmented state conditional from Equation (3.13), we can form the joint density for $\boldsymbol{\phi}$ and $\mathbf{f}$ by multiplying by the $m\mathcal{G}v\mathcal{M}$ in Equation (3.9). This operation results in the distribution

$$\begin{aligned} p(\boldsymbol{\phi}, \mathbf{f}) \propto \exp\Big\{&-\frac{1}{2}\mathbf{f}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f} + \mathbf{f}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}(\boldsymbol{\phi}) - \frac{1}{2}\mathbf{v}(\boldsymbol{\phi})^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}(\boldsymbol{\phi}) && (3.14)\\ &+ \boldsymbol{\kappa}^\top\cos(\boldsymbol{\phi}-\boldsymbol{\mu}) - \frac{1}{2}\mathbf{v}(\boldsymbol{\phi})^\top\mathbf{K}^{-1}\mathbf{v}(\boldsymbol{\phi})\Big\}\\ = \exp\Big\{&-\frac{1}{2}\mathbf{f}^\top\boldsymbol{\Sigma}^{-1}\mathbf{f} + \Big(\boldsymbol{\Sigma}^{-1}\mathbf{f} + \boldsymbol{m}(\boldsymbol{\kappa},\boldsymbol{\mu})\Big)^\top\mathbf{v}(\boldsymbol{\phi}) && (3.15)\\ &-\frac{1}{2}\mathbf{v}(\boldsymbol{\phi})^\top(\boldsymbol{\Sigma}^{-1}+\mathbf{K}^{-1})\mathbf{v}(\boldsymbol{\phi})\Big\} \end{aligned}$$

where we introduced the notation $\boldsymbol{m}(\boldsymbol{\kappa},\boldsymbol{\mu})$ as $[(\boldsymbol{\kappa}\odot\cos\boldsymbol{\mu})^\top, (\boldsymbol{\kappa}\odot\sin\boldsymbol{\mu})^\top]$ for $\odot$ denoting the Hadamard product.

Equation (3.15) highlights central difference between the Hubbard-Stratonovich transformation for Ising models and the augmentation: in the $m\mathcal{G}v\mathcal{M}$ case, the covariance cannot simply mirror the $m\mathcal{G}v\mathcal{M}$'s kernel matrix $\mathbf{K}$. To find a suitable covariance, we draw on the spectral decomposition of the matrix $\mathbf{K}^{-1}$, i.e.,

$$\mathbf{K}^{-1} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}, \tag{3.16}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues of $\mathbf{K}^{-1}$ and $\mathbf{P}$ are orthonormal matrices. If let $\lambda_{\max}$ denote the maximum eigenvalue of $\mathbf{K}^{-1}$, the precision of the augmentation states can be chosen to be

$$\mathbf{\Sigma}^{-1} = (1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{K}^{-1} \tag{3.17}$$

for some $\epsilon > 0$.

It is simple to show that Equation (3.17) results in a valid choice of a precision matrix. Drawing on the definition for this covariance and the spectral decomposition, we can express the precision as

$$\begin{aligned}
\mathbf{\Sigma}^{-1} &= ((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{K}^{-1} && (3.18)\\
&= (1+\epsilon)\lambda_{\max}\mathbf{P}\mathbf{P}^{-1} - \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} && (3.19)\\
&= \mathbf{P}((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{\Lambda})\mathbf{P}^{-1}. && (3.20)
\end{aligned}$$

Since $(1+\epsilon)\lambda_{\max}$ is strictly greater than all of the eigenvalues of $\mathbf{K}^{-1}$, Equation (3.17) is a valid covariance matrix. With the precision from Equation (3.17), the joint model from Equation (3.15) can be written as

$$\begin{aligned}
p(\boldsymbol{\phi}, \mathbf{f}) \propto \exp\Big\{ & -\frac{1}{2}\mathbf{f}^\top\mathbf{\Sigma}^{-1}\mathbf{f} + \left(\mathbf{\Sigma}^{-1}\mathbf{f} + \boldsymbol{m}(\boldsymbol{\kappa}, \boldsymbol{\mu})\right)^\top \mathbf{v}(\boldsymbol{\phi}) && (3.21)\\
& -\frac{1}{2}\mathbf{v}(\boldsymbol{\phi})^\top\left((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{K}^{-1} + \mathbf{K}^{-1}\right)\mathbf{v}(\boldsymbol{\phi})\Big\}\\
= \exp\Big\{ & -\frac{1}{2}\mathbf{f}^\top\mathbf{\Sigma}^{-1}\mathbf{f} + \left(\mathbf{\Sigma}^{-1}\mathbf{f} + \boldsymbol{m}(\boldsymbol{\kappa}, \boldsymbol{\mu})\right)^\top \mathbf{v}(\boldsymbol{\phi}) && (3.22)\\
& -\frac{(1+\epsilon)\lambda_{\max}}{2}\mathbf{v}(\boldsymbol{\phi})^\top\mathbf{v}(\boldsymbol{\phi})\Big\}.
\end{aligned}$$

Using the fundamental trigonometric identity $\sin^2\phi + \cos^2\phi = 1$, the residual quadratic term $\mathbf{v}(\boldsymbol{\phi})^\top\mathbf{v}(\boldsymbol{\phi})$ becomes the number of circular variables, that is, N. The joint can be then factorised to yield a product of independent von Mises distributions

over the circular variables and a correlated Gaussian augmentation state, i.e.,

$$p(\boldsymbol{\phi}, \mathbf{f}) = p(\boldsymbol{\phi}|\mathbf{f})p(\mathbf{f}) = \left[\prod_{\mathrm{n}=1}^{\mathrm{N}} \upsilon\mathcal{M}(\phi_{\mathrm{n}}, \alpha_{\mathrm{n}}, \beta_{\mathrm{n}})\right] \times \mathcal{N}(\mathbf{f}; 0, \boldsymbol{\Sigma}) \tag{3.23}$$

where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are defined such that

$$\boldsymbol{m}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\Sigma}^{-1}\mathbf{f} + \boldsymbol{m}(\boldsymbol{\kappa}, \boldsymbol{\mu}). \tag{3.24}$$

There are infinite possible augmented joint distributions for each $m\mathcal{G}\upsilon\mathcal{M}$. This is evident from the fact that Equation (3.23) depends on the precision matrix of the augmented states, which is parametrised by a $\epsilon > 0$. A way to define a valid augmented distribution is that avoids the numerically cumbersome eigenvalue computations, is to define ϵ in terms of an upper bound. Given that a direct consequence of the positive-definiteness of $\mathbf{K}^{-1}$ is that its trace will be greater than the maximum eigenvalue, we draw on Bai and Golub (1997)'s upper bound for the trace of the inverse a N by N positive definite matrix $\mathbf{K}$,

$$\mathrm{Tr}\left(\mathbf{K}^{-1}\right) \leq \begin{bmatrix} \mathrm{Tr}(\mathbf{K}) \\ \mathrm{N} \end{bmatrix}^{\top} \begin{bmatrix} ||\mathbf{K}||_F^2 & \mathrm{Tr}(\mathbf{K}) \\ \gamma^2 & \gamma \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{N} \\ 1 \end{bmatrix} \tag{3.25}$$

where $||\cdot||_F^2$ is the squared Frobenius norm and γ is a positive constant that lower-bounds the eigenvalues of $\mathbf{K}$. While for an arbitrary matrix it is difficult to define the constant γ, in practice this issue can be completely remedied by always adding a small perturbation matrix $\sigma^2\mathbf{I}$ to the kernel matrix, for some $\sigma^2 \to 0^+$. Hence, expanding Equation (3.25), a valid ϵ can always be found using

$$\epsilon = \frac{\sigma^2 \mathrm{N}\mathrm{Tr}(\mathbf{K} + \sigma^2\mathbf{I}) - \mathrm{Tr}(\mathbf{K} + \sigma^2\mathbf{I})^2 - \sigma^2\mathrm{N}^2 + \mathrm{N}||\mathbf{K} + \sigma^2\mathbf{I}||_F^2}{\sigma^2||\mathbf{K} + \sigma^2\mathbf{I}||_F^2 - \sigma^4\mathrm{Tr}(\mathbf{K} + \sigma^2\mathbf{I})}. \tag{3.26}$$

Another relationship worth analysing lies in the graphical model structure of the augmented model. Latent variables $\mathbf{f}$ induces a multi-view structure (Blum and Mitchell, 1998; de Sa, 1994; Rüping and Scheffer, 2005) where f_{n} and $\mathsf{f}_{\mathrm{N+n}}$ can be understood as the real and imaginary views that produce the phase of a complex variable. To simplify the notation in the remainder of the chapter and emphasise this multi-view interpretation, in what follows we will denote $\mathsf{f}_1, \ldots, \mathsf{f}_{\mathrm{N}}$ as $\mathsf{f}_{\mathfrak{Re},1}, \ldots, \mathsf{f}_{\mathfrak{Re},\mathrm{N}}$ as the real component tied to the cosines in the models in Figure 3.1 and $\mathsf{f}_{\mathrm{N+1}}, \ldots, \mathsf{f}_{\mathrm{2N}}$ as $\mathsf{f}_{\mathfrak{Im},1}, \ldots, \mathsf{f}_{\mathfrak{Im},\mathrm{N}}$ tied to the sines of the graphical models of Figure 3.1.

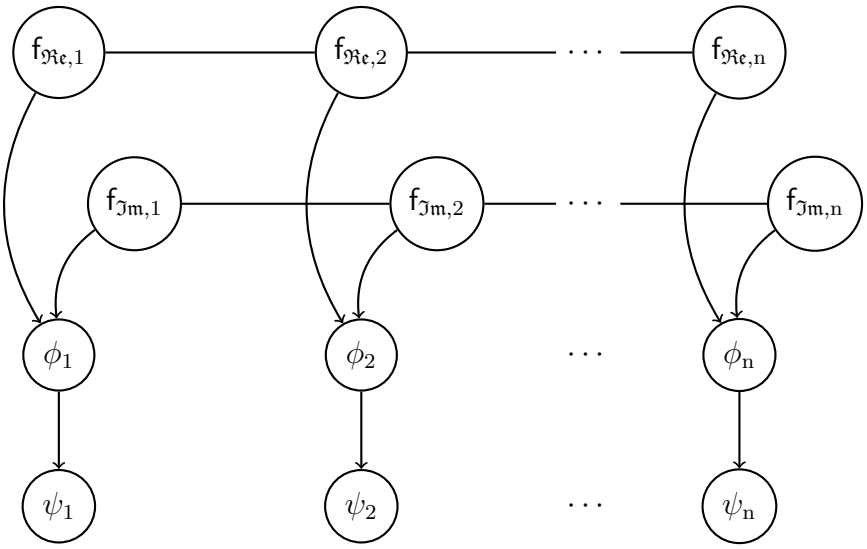


Fig. 3.1 The graphical model for different augmented *mGvM* representations denoting $\mathsf{f}_1, \ldots, \mathsf{f}_N$ as $\mathsf{f}_{\Re\mathfrak{e},1}, \ldots, \mathsf{f}_{\Re\mathfrak{e},N}$ and $\mathsf{f}_{N+1}, \ldots, \mathsf{f}_{2N}$ as $\mathsf{f}_{\Im\mathfrak{m},1}, \ldots, \mathsf{f}_{\Im\mathfrak{m},N}$: the graphical model of the augmented representation gives rise to a multi-view structure.

A few interesting properties arise from the functional form of the augmented representation. Perhaps the most important one lies in the analysing the augmented representation itself as a new probabilistic model. In particular, the resulting model will be consistent under marginalisation of the variable triple $(\phi_n, \mathsf{f}_{\Re\mathfrak{e},n}, \mathsf{f}_{\Im\mathfrak{m},n})$ by construction. Hence, we can lean on the Daniell–Kolmogorov Extension Theorem (Daniell, 1919; Kolmogoroff, 1933; Rogers and Williams, 2000) and assert that the augmented representation of the *mGvM* defines a stochastic process over $(\phi, \mathsf{f}_{\Re\mathfrak{e}}, \mathsf{f}_{\Im\mathfrak{m}})$.

A sample from the augmented representation are presented in Figure 3.2 for a *mGvM* with by viewing the augmented model as a generative process as presented in Algorithm 1. It can be noted that while the mean function is continuous and smooth, the individual samples of the process need not be. Since each sample is obtained from a von Mises distribution, for finite concentrations there is bound to be discontinuities between samples.

This noisy characteristic is different from the behaviour observed when using smooth covariance functions like the squared exponential kernel on standard Gaussian Process models, however it is also found in augmented representations of Gaussian Processes like sparse models. In particular, the structure resembles that of the SPGP of Snelson and Ghahramani (2006b) where noise is added to the diagonal entries of the covariance to generate heteroskedastic behaviour, as the net effect the augmentation

has on the individual samples of the model can be viewed as changing the value of the concentration of each ϕ locally.

Algorithm 1: Obtaining sample functions from the augmented representation of the $m\mathcal{G}v\mathcal{M}$

Sample $\mathbf{f}$ from $\mathcal{N}(\mathbf{f}; \boldsymbol{m}, (\epsilon\mathbf{I} - \mathbf{K}^{-1})^{-1})$
for $\mathrm{n} = 1 : \mathrm{N}$**,**
    Set $\kappa = \mathbf{abs}(\mathsf{f}_{\mathrm{n}} + \imath\mathsf{f}_{\mathrm{N+n}})$
    Set $\mu = \mathbf{ang}(\mathsf{f}_{\mathrm{n}} + \imath\mathsf{f}_{\mathrm{N+n}})$
    Sample ϕ_{n} **from** $v\mathcal{M}(\phi_{\mathrm{n}}; \kappa, \mu)$

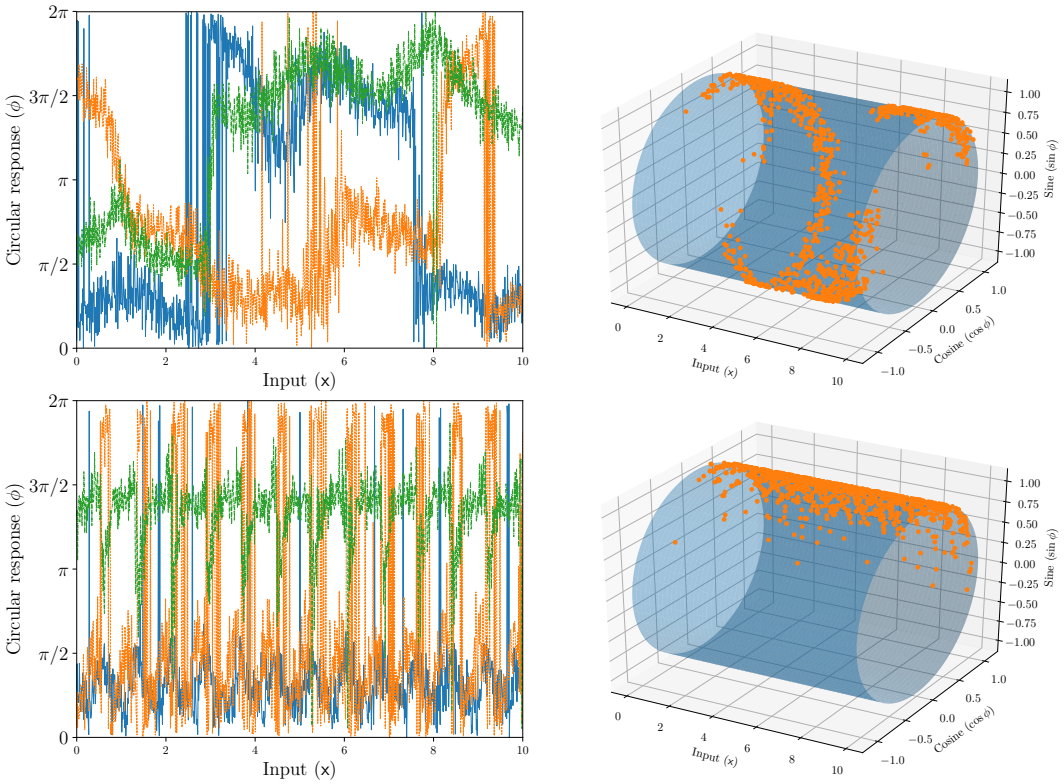


Fig. 3.2 Samples from the process defined by the augmented representation of the $m\mathcal{G}v\mathcal{M}$ model using squared exponential (top), periodic (bottom) kernels. Vertical lines spanning the entire interval $[0, 2\pi)$ indicate wrapping. The individual samples of the process are denoted by dashed lines and are not continuous.

3.2 Related work

Martens and Sutskever (2010) reintroduced the Hubbard-Stratonovich transformation in discrete Markov Random Fields as a way to parallelise Markov chain Monte Carlo

algorithms. Zhang and coworkers (Zhang et al., 2012) also leveraged this transformation as a method to apply the Hamiltonian Monte Carlo sampling method for discrete variables. Later, Pakman and Paninski (Pakman and Paninski, 2013) expanded on this work and analysed the transformation under a auxiliary-variable method for sampling binary states.

In the context of circular and directional statistics, series approximations to partition functions (Kume et al., 2013) and pseudo-likelihood models (Mardia, 2007) have been favoured over augmented representations when approximate partition functions for learning models. To our best knowledge, the topic of sparse covariances in augmented representations for circular distributions has not been explored previously in literature.

3.3 Augmented distributions for more general distributions on Stiefel Manifolds

Interestingly, the approach undertaken in Section 3.1 is widely applicable to problems of interest to circular statistics. In the context of circular and directional statistics, augmented representations based on the Hubbard-Stratonovich transformation can be interpreted as a geometrical transformation that changes the problem geometry from a unwieldy Stiefel manifold to a simpler distribution on a hyper-cylinder. This observation is important because this model augmentation is not only applicable to any $m\mathcal{G}v\mathcal{M}$-derived distribution, but also to the entire family of distributions derived from the matrix Fisher-Bingham as shown next.

Recall that the matrix Fisher-Bingham distribution ($\mathcal{MFB}$) (Kume et al., 2013) is

$$\mathcal{MFB}(\mathbf{S};\boldsymbol{\eta},\boldsymbol{\Sigma}) \propto \exp\left\{\boldsymbol{\eta}^{\top}\mathrm{vec}(\mathbf{S}) - \frac{1}{2}\mathrm{vec}(\mathbf{S})^{\top}\boldsymbol{\Sigma}^{-1}\mathrm{vec}(\mathbf{S})\right\} \tag{3.27}$$

where $\boldsymbol{\eta}$ is a location and concentration parameter, while $\boldsymbol{\Sigma}$ captures the covariance between unit vectors, and $\mathbf{S}$ is a $\mathrm{N} \times \mathrm{M}$ matrix where M is the dimension of the hyper-sphere of the Stiefel manifold and N is the number of Cartesian products between hyper-spheres manifolds.

To provide more intuition about the Matrix Fisher-Bingham distribution, we provide two concrete examples that tie to its applications. The first example a $m\mathcal{G}v\mathcal{M}$ of dimension D, since each Stiefel manifold is a unit circle, $\mathbb{S}^1$, $\mathrm{M} = 2$ and there are D Cartesian products between each hypersphere, $\mathrm{N} = \mathrm{D}$. The second example is a discrete time series consisting of T steps where each time point represents an orientation, i.e. a point the canonical sphere in $\mathbb{R}^3$ (i.e. $\mathbb{S}^2$). In this case, the matrix Fisher-Bingham

for the entire time series has N = T and M = 3. An augmented conditional suitable to the $\mathcal{MFB}$ can be formed using a Gaussian state as proposed to the $m\mathcal{G}v\mathcal{M}$ in Equation (3.13). Explicitly, the augmented conditional can be defined as

$$p(\mathbf{f}|\mathbf{S}) = \mathcal{N}\Big(\mathbf{f}; \text{vec}(\mathbf{S}), \Big((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{\Sigma}^{-1}\Big)^{-1}\Big) \tag{3.28}$$

where ϵ is a positive real constant and $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{\Sigma}^{-1}$).

Drawing on Equation (3.28), the joint distribution of base $\mathcal{MFB}$ states and the Gaussian augmentation states after simplifications analogous to those performed for the $m\mathcal{G}v\mathcal{M}$ case in Section 3.1 is

$$\begin{aligned} p(\mathbf{S}, \mathbf{f}) \propto \exp\Bigg\{ & \Big(\Big((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{\Sigma}^{-1}\Big)\mathbf{f} + \boldsymbol{\eta}\Big)^{\top} \text{vec}(\mathbf{S}) \\ & - \frac{1}{2}\mathbf{f}^{\top}\Big((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{\Sigma}^{-1}\Big)\mathbf{f}\Bigg\} \end{aligned} \tag{3.29}$$

which results in a model corresponding to a product of von Mises-Fisher conditional distributions and Gaussian states,

$$p(\mathbf{S}, \mathbf{f}) = \underbrace{\mathcal{N}\Big(\mathbf{f}; 0, \Big((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{\Sigma}^{-1}\Big)^{-1}\Big)}_{p(\mathbf{f})} \times \underbrace{\prod_{\text{n}=1}^{\text{N}} \upsilon\mathcal{MF}\Big(\mathbf{S}_{\text{n}}; \boldsymbol{\omega}_{\text{idx}(\mathbf{S}_{\text{n}})}\Big)}_{p(\mathbf{S}|\mathbf{f})}, \tag{3.30}$$

where $\mathbf{S}_{\text{n}}$ denotes the entire n-th row of $\mathbf{S}$, $\boldsymbol{\omega}$ is the von Mises-Fisher parameter given by $\boldsymbol{\omega} = \Big((1+\epsilon)\lambda_{\max}\mathbf{I} - \mathbf{\Sigma}^{-1}\Big)\mathbf{f} + \boldsymbol{\eta}$ and the notation $\boldsymbol{\omega}_{\text{idx}(\mathbf{S}_{\text{n}})}$ indicates the entries of $\boldsymbol{\omega}$ corresponding to $\mathbf{S}_{\text{n}}$.

3.4 Summary

In this chapter, we have introduced for the first time an augmented representation in the context of circular and directional variables motivated from a exchangeability perspective.

The augmented representation outlined in this chapter allows re-writing the $m\mathcal{G}v\mathcal{M}$ in a factored form for which all partition functions are known. This characteristic allows learning to be performed on models bearing the augmented representation of the $m\mathcal{G}v\mathcal{M}$ as a prior. When the partition functions for the prior are not known, learning has to be performed through techniques such as contrastive-divergence learning

explored in Chapter 6, which are computationally expensive and restricted to small scale models.

Combining the augmented representations with the sparse constructions outlined in this chapter allows models of considerable size to be learned. While not restricted to density modelling or circular regression, the sparse models defined in this chapter can be particularly useful when performing regression and prediction on large data sets.

Part III

Inference and learning

Chapter 4

Variational free energy methods

In Part II, we introduced the multivariate Generalised von Mises distribution along with an associated augmented representation related to the Hubbard-Stratonovich transformation. In Part III our focus shifts to providing methods to perform inference on these models. Therefore, from Chapter 4 to Chapter 6, we successively introduce, compare and explore methods for performing inference with the $m\mathcal{G}v\mathcal{M}$ and its augmented representations where applicable.

In this chapter, we introduce the Variational Free Energy framework for approximate inference with the $m\mathcal{G}v\mathcal{M}$. In particular, we focus on developing a mean field approximation for the $m\mathcal{G}v\mathcal{M}$ and test its performance against standard approximations in machine learning practice leveraging Euclidean distributions. We show that despite the limitations inherent to a simple mean field approximation, regression and latent variable modelling using the $m\mathcal{G}v\mathcal{M}$ for circular data outperforms the approaches rooted in Euclidean distributions. Moreover, the relative success of the mean field approximation allows us to use it as a benchmark for evaluating different approaches in subsequent chapters.

4.1 Inference

This section is organised as follows. Section 4.1.1 provides a brief exposition of the Variational Free Energy framework to the novice following the treatment of Barber (2012) and MacKay (2003). The expert reader may start directly in Section 4.1.2, where we derive the mean field distribution associated with the $m\mathcal{G}v\mathcal{M}$. The chapter concludes by providing experiments in Section 4.2 of the approximation variational inference updates devised for the $m\mathcal{G}v\mathcal{M}$ against standard Euclidean approaches to

handling circularity. The experimental datasets span both synthetic and real-world cases.

4.1.1 Variational Free Energy

A simple motivation for the Variational Free Energy framework stems from approximating the log marginal likelihood from below. This result can be obtained by first interpreting the marginal likelihood $p(\mathbf{y}|\theta_p)$ as the resulting distribution obtained from marginalising a latent variable $\mathbf{x}$,

$$p(\mathbf{y}|\theta_p) = \int p(\mathbf{x}, \mathbf{y}|\theta_p)\mathrm{d}\mathbf{x} \tag{4.1}$$

Drawing on Jensen's inequality and the fact that log is a concave function, it can be established that the expectation relationship

$$\log\left[\int f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}\right] \geq \int \log[f(\mathbf{x})]p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{4.2}$$

holds for any distribution and $f : \mathbb{R}^{\mathrm{N}} \mapsto \mathbb{R}_{++}$. Therefore,

$$\log p(\mathbf{y}|\theta_p) = \log \int p(\mathbf{y}, \mathbf{x}|\theta_p)\mathrm{d}\mathbf{x} \tag{4.3}$$

$$= \log \int \frac{p(\mathbf{y}, \mathbf{x}|\theta_p)q(\mathbf{x}|\theta_q)}{q(\mathbf{x}|\theta_q)}\mathrm{d}\mathbf{x} \tag{4.4}$$

$$\geq \int q(\mathbf{x}|\theta_q) \log \frac{p(\mathbf{y}, \mathbf{x}|\theta_p)}{q(\mathbf{x}|\theta_q)}\mathrm{d}\mathbf{x} = \mathcal{F}(q, \theta_p) \tag{4.5}$$

where we have used $f(\mathbf{x}) = p(\mathbf{y}, \mathbf{x}|\theta_p)/q(\mathbf{x}|\theta_q)$. Here $\mathcal{F}$ is defined as the Free Energy functional whose arguments are the approximating distribution q and the parameters θ_p of the posterior.

Another way to derive the bound of Equation (4.5) is to minimise the error between a posterior distribution $p(\mathbf{x}|\mathbf{y}, \theta_p)$ and an approximating distribution q. The canonical information theoretic metric for performing measuring the approximation error is the Kullback-Leiber (KL) divergence,

$$\mathrm{KL}(q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y}, \theta_p)}\mathrm{d}\mathbf{x}. \tag{4.6}$$

The Kullback-Leibler divergence is asymmetric and, while it may be argued that the best representation for the approximation error when approximating p with q is $\mathrm{KL}(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x}))$, typically the posterior p is intractable and renders the computation

of $\mathrm{KL}(p(\mathbf{x}|\mathbf{y})||q(\mathbf{x}))$ intractable. Hence, in the Variational Free Energy framework it is customary to use the divergence $\mathrm{KL}(q(\mathbf{x})||p(\mathbf{x}|\mathbf{y}))$.

We can perform algebraic manipulations on the KL divergence to recover the Free Energy function, that is

$$\begin{aligned}
\mathrm{KL}(q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})) &= -\int q(\mathbf{x}|\theta_q)\log p(\mathbf{x}|\mathbf{y},\theta_p)\mathrm{d}\mathbf{x} + \int q(\mathbf{x}|\theta_q)\log q(\mathbf{x}|\theta_q)\mathrm{d}\mathbf{x} && (4.7)\\
&= -\int q(\mathbf{x}|\theta_q)\log p(\mathbf{x},\mathbf{y}|\theta_p)\mathrm{d}\mathbf{x} + \int q(\mathbf{x}|\theta_q)\log p(\mathbf{y}|\theta_p)\mathrm{d}\mathbf{x} && (4.8)\\
&\quad + \int q(\mathbf{x}|\theta_q)\log q(\mathbf{x}|\theta_q)\mathrm{d}\mathbf{x}\\
&= \log p(\mathbf{y}|\theta_p) - \underbrace{\Big(\langle \log p(\mathbf{x},\mathbf{y}|\theta_p)\rangle_{q(\mathbf{x}|\theta_q)} + \mathcal{H}(q)\Big)}_{\mathcal{F}(q,\theta_p)} && (4.9)
\end{aligned}$$

where we adopted the angled bracket shorthand notation for denoting expectations as is customary in physics. The lower bound is then found using the fact that the Kullback-Leiber divergence is non-negative and equal to zero only the approximating distribution is equal to the approximated distribution,

$$\log p(\mathbf{y}) = \mathcal{F}(q,\theta_p) + \mathrm{KL}(q(\mathbf{x})||p(\mathbf{x}|\mathbf{y})) \geq \mathcal{F}(q,\theta_p). \tag{4.10}$$

To find the optimal approximating distribution q we turn to calculus of variations and form the Lagrangian for maximising the Free Energy. Maximising the Free Energy is equivalent to minimise the divergence between p and q. The resulting Lagrangian has the form

$$\mathcal{L}(q,\lambda) = \mathcal{F}(q,\theta_p) - \lambda\left(\int q(\mathbf{x})\mathrm{d}\mathbf{x} - 1\right) \tag{4.11}$$

where we highlight the constraint that $q(\mathbf{x})$ is a valid, normalised distribution. Imposing the first order optimality condition and finding the stationary points of the Lagrangian,

$$\frac{\delta}{\delta q}\mathcal{L}(q,\lambda) = 0 \tag{4.12}$$

requires specifing the class of valid distributions q can be chosen from *a priori*. Typically, this distribution will either be a parametric distribution, for example a multivariate Gaussian, a distribution expressed as the fully factorised product of parametric distributions, or a distribution with a conditonal independence assumption build into it. The fully factorised case, i.e.

$$q(\mathbf{x}) = \prod_{\mathrm{d}=1}^{\mathrm{D}} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}}) \tag{4.13}$$

leads to the solution known as the mean-field approximation. This moniker arises from the fact that the solution that the fixed points of the Lagrangian under this assumption imply that each distribution $q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})$ only depends on expectations of the remaining variables $\mathsf{x}_{\ell \neq \mathrm{d}}$.

To verify this result, we apply Equation (4.13) to Equation (4.12) and expanding yields

$$\begin{aligned} \frac{\delta}{\delta q_\ell}\Big[\langle \log p(\mathbf{x}, \mathbf{y}) \rangle_{\prod_{\mathrm{d}} q_{\mathrm{d}}(\mathbf{x}_{\mathrm{d}})} - \int \prod_{\mathrm{d}=1}^{\mathrm{D}} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}}) \left(\sum_{\mathrm{d}=1}^{\mathrm{D}} \log q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}}) \right) \mathrm{d}\mathbf{x} \\ - \sum_{\mathrm{d}=1}^{\mathrm{D}} \lambda_{\mathrm{d}} \int q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}}) \mathrm{d}\mathsf{x}_{\mathrm{d}} \Big] = 0 \end{aligned} \tag{4.14}$$

simplifying the differentials gives

$$\langle \log p(\mathbf{x}, \mathbf{y}) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})} - \log q_\ell(\mathsf{x}_\ell) - 1 - \lambda_\ell = 0 \tag{4.15}$$

which in turn can be expressed as

$$q_\ell(\mathsf{x}_\ell) = \frac{1}{\exp(\lambda + 1)} \exp \left\{ \langle \log p(\mathbf{x}, \mathbf{y}) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})} \right\} \tag{4.16}$$

resulting in the set of distributions known as mean field approximation.

The usefulness of the mean field approximation is strongly related to the tractability of the conditional distributions of the model. To see this, we expand Equation (4.16) into

$$q_\ell(\mathsf{x}_\ell) \propto \exp \left\{ \langle \log p(\mathbf{x}, \mathbf{y}) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})} \right\} \tag{4.17}$$

$$\propto \exp \left\{ \langle \log \left(p(\mathsf{x}_\ell | \mathbf{x}_{\mathrm{d} \neq \ell}, \mathbf{y}) p(\mathbf{x}_{\mathrm{d} \neq \ell}, \mathbf{y}) \right) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})} \right\} \tag{4.18}$$

$$\propto \exp \left\{ \langle \log p(\mathsf{x}_\ell | \mathbf{x}_{\mathrm{d} \neq \ell}, \mathbf{y}) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})} + \langle \log p(\mathbf{x}_{\mathrm{d} \neq \ell}, \mathbf{y}) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathsf{x}_{\mathrm{d}})} \right\} \tag{4.19}$$

$$\propto \exp \left\{ \langle \log p(\mathsf{x}_\ell | \mathbf{x}_{\mathrm{d} \neq \ell}, \mathbf{y}) \rangle_{\prod_{\mathrm{d} \neq \ell} q_{\mathrm{d}}(\mathbf{x}_{\mathrm{d}})} \right\} \tag{4.20}$$

Equation (4.20) implies that if the conditionals are tractable, then the resulting variational approximation should be tractable as well.

The fully-factored representation of the mean field approximation cannot capture well the correlations between variables. Hence, mean field approximations possess a bias towards under-representing the uncertainty present in the true posterior. A thorough analysis of such problems is given by Turner et al. (2008). For brevity of exposition, we illustrate this issue through a simple concrete example. Consider the

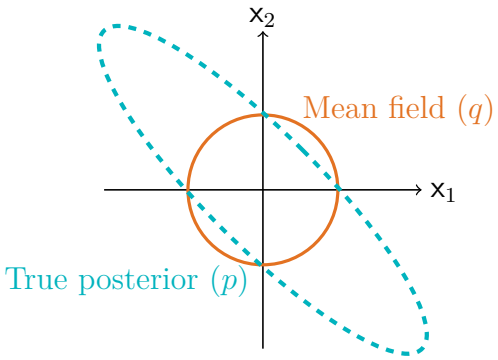


Fig. 4.1 Diagrammatic representation of the bias present in a mean-field approximation, showing equivalent level sets for the posterior p and approximation q. The correlation structure of the posterior is not captured by the mean field approximation.

posterior

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left\{ \begin{bmatrix} \mathsf{x}_1 \\ \mathsf{x}_2 \end{bmatrix}^\top \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} \begin{bmatrix} \mathsf{x}_1 \\ \mathsf{x}_2 \end{bmatrix} \right\} \tag{4.21}$$

for some parameters α, β and γ potentially dependent on $\mathbf{y}$. The model of Equation (4.21) has the mean field distributions

$$q(\mathbf{x}) \propto \exp\left\{ \begin{bmatrix} \mathsf{x}_1 - m_1(\langle \mathsf{x}_2 \rangle_{q_2(\mathsf{x}_2)}) \\ \mathsf{x}_2 - m_2(\langle \mathsf{x}_1 \rangle_{q_1(\mathsf{x}_1)}) \end{bmatrix}^\top \begin{bmatrix} \alpha & 0 \\ 0 & \gamma \end{bmatrix} \begin{bmatrix} \mathsf{x}_1 - m_1(\langle \mathsf{x}_2 \rangle_{q_2(\mathsf{x}_2)}) \\ \mathsf{x}_2 - m_2(\langle \mathsf{x}_1 \rangle_{q_1(\mathsf{x}_1)}) \end{bmatrix} \right\} \tag{4.22}$$

which removes the cross dependency imposed by the beta parameter leading to over-confident situations as illustrated in Figure 4.1.

Prior work has applied variational procedures for probabilistic models using circular distributions has been conducted focusing the applications of phase inference and clustering. As an example, Taghia et al. (2013) used a simple variational Bayes procedure to approximate the posterior over the phase in a hierarchical model. The use of von Mises priors over phase variables can be also found in signal processing algorithms, including generalisations of previously known algorithms such as turbo synchronisation[1] Herzet et al. (2007).

[1]Turbo synchronisation is a an information-theoretic algorithm for detecting and extracting the phases component of real-valued signals. Areas of interest related to algorithms such as this one include coding theory, compression, information retrieval and, more broadly, signal processing.

Another application of the variational framework in circular statistics is found in clustering. For example, Tang et al. (2009) used a mixture of von Mises-Fisher distributions to perform the clustering of speakers from audio data.

4.1.2 Mean field variational inference for the $m\mathcal{G}v\mathcal{M}$

The results obtained in Section 4.1.1 are distribution-independent, hence can also be applied to a $m\mathcal{G}v\mathcal{M}$ posterior

$$p(\boldsymbol{\phi}) \propto \exp\left\{ \boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi} - \boldsymbol{\mu}) - \frac{1}{2} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}^\top \mathbf{K}^{-1} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \right\}. \tag{4.23}$$

If this $m\mathcal{G}v\mathcal{M}$ posterior is approximated using a fully-factored approximation $q(\boldsymbol{\phi}) = \prod_{\mathrm{d}=1}^{\mathrm{D}} q_{\mathrm{d}}(\phi_{\mathrm{d}})$, and the derivations for the optimal distribution are perfomed applying calculus of variations as described in the previous section, it is possible to demonstrate that the functional form for each of mean field conditional q_{d} is a $\mathcal{G}v\mathcal{M}$ distribution, that is,

$$q(\phi_{\mathrm{d}}|\boldsymbol{\phi}_{\neq \mathrm{d}}) = \mathcal{G}v\mathcal{M}\left(\phi_{\mathrm{d}}; \eta_{1,\mathrm{d}}, \eta_{2,\mathrm{d}}, \nu_{1,\mathrm{d}}, \nu_{2,\mathrm{d}}\right) \tag{4.24}$$

where the parameters $\eta_{1,\mathrm{d}}, \eta_{2,\mathrm{d}}, \nu_{1,\mathrm{d}}$ and $\nu_{2,\mathrm{d}}$ can be obtained directly taken from the original $m\mathcal{G}v\mathcal{M}$ distribution. More specifically, the parameters can be calculated through the relations

$$\eta_{1,\mathrm{d}} \cos\nu_{1,\mathrm{d}} = \kappa_{\mathrm{d}} \cos(\mu_{\mathrm{d}}) - \frac{1}{2} \sum_{\ell \neq \mathrm{d}}^{\mathrm{D}} \left\langle (\mathbf{K}^{-1})_{\mathrm{d},\ell} \cos\boldsymbol{\phi}_\ell + (\mathbf{K}^{-1})_{\mathrm{d},\ell+\mathrm{D}} \sin\boldsymbol{\phi}_\ell \right\rangle_{q_\ell}, \tag{4.25}$$

$$\eta_{1,\mathrm{d}} \sin\nu_{1,\mathrm{d}} = \kappa_{\mathrm{d}} \sin(\mu_{\mathrm{d}}) - \frac{1}{2} \sum_{\ell \neq \mathrm{d}}^{\mathrm{D}} \left\langle (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell} \cos\boldsymbol{\phi}_\ell + (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell+\mathrm{D}} \sin\boldsymbol{\phi}_\ell \right\rangle_{q_\ell}, \tag{4.26}$$

$$\eta_{2,\mathrm{d}} \cos 2\nu_{2,\mathrm{d}} = \frac{1}{4} \left[(\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}} - (\mathbf{K}^{-1})_{\mathrm{D}+\mathrm{d},\mathrm{d}+\mathrm{D}} \right] \text{ and} \tag{4.27}$$

$$\eta_{2,\mathrm{d}} \sin 2\nu_{2,\mathrm{d}} = \frac{1}{2} (\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}+\mathrm{D}}. \tag{4.28}$$

The dependencies Equation (4.25) to Equation (4.25) can be readily obtained from the optimal fully factored approximation in Equation (4.20) using the results from Chapter 2 for the conditionals, that is, Section 2.2.2. To see this, notice that the log-unnormalised distribution corresponding to Equation (4.23) is a quadratic in $\cos\boldsymbol{\phi}$

and $\sin \boldsymbol{\phi}$, that is, it can be expressed as

$$\begin{aligned}
\log p^*(\boldsymbol{\phi}) = &\sum_{\mathrm{d}=1}^{\mathrm{D}} \Bigg[\kappa_{\mathrm{d}} (\cos \phi_{\mathrm{d}} \cos \mu_{\mathrm{d}} + \sin \phi_{\mathrm{d}} \sin \mu_{\mathrm{d}}) \Bigg] \\
&- \frac{1}{2} \sum_{\mathrm{d}=1}^{\mathrm{D}} \Bigg[\cos^2 \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}} + \sin^2 \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\mathrm{d}+\mathrm{D}} \Bigg] \\
&- \frac{1}{2} \sum_{\mathrm{d}=1}^{\mathrm{D}} \Bigg[2 \sin \phi_{\mathrm{d}} \cos \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}+\mathrm{D}} \Bigg] \\
&- \frac{1}{2} \sum_{\mathrm{d}=1}^{\mathrm{D}} \sum_{\substack{\ell=1 \\ \ell \neq \mathrm{d}}}^{\mathrm{D}} \Bigg[\cos \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d},\ell} \cos \phi_{\ell} + \cos \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d},\ell+\mathrm{D}} \sin \phi_{\ell} \Bigg] \\
&- \frac{1}{2} \sum_{\mathrm{d}=1}^{\mathrm{D}} \sum_{\substack{\ell=1 \\ \ell \neq \mathrm{d}}}^{\mathrm{D}} \Bigg[\sin \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell} \cos \phi_{\ell} + \sin \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell+\mathrm{D}} \cos \phi_{\ell} \Bigg].
\end{aligned} \tag{4.29}$$

Hence, the unidimensional log-unnormalised distribution for ϕ_{d} is given as

$$\begin{aligned}
\log p^*(\phi_{\mathrm{d}} | \boldsymbol{\phi}_{\ell \not= \mathrm{d}}) = &\kappa_{\mathrm{d}} (\cos \phi_{\mathrm{d}} \cos \mu_{\mathrm{d}} + \sin \phi_{\mathrm{d}} \sin \mu_{\mathrm{d}}) \\
&- \frac{1}{2} \Bigg[\cos^2 \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}} + \sin^2 \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\mathrm{d}+\mathrm{D}} \Bigg] \\
&- \frac{1}{2} \cos \phi_{\mathrm{d}} \sum_{\substack{\ell=1 \\ \ell \neq \mathrm{d}}}^{\mathrm{D}} \Bigg[(\mathbf{K}^{-1})_{\mathrm{d},\ell} \cos \phi_{\ell} + (\mathbf{K}^{-1})_{\mathrm{d},\ell+\mathrm{D}} \sin \phi_{\ell} \Bigg] \\
&- \frac{1}{2} \Bigg[2 \sin \phi_{\mathrm{d}} \cos \phi_{\mathrm{d}} (\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}+\mathrm{D}} \Bigg] \\
&- \frac{1}{2} \sin \phi_{\mathrm{d}} \sum_{\substack{\ell=1 \\ \ell \neq \mathrm{d}}}^{\mathrm{D}} \Bigg[(\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell} \cos \phi_{\ell} + (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell+\mathrm{D}} \cos \phi_{\ell} \Bigg],
\end{aligned} \tag{4.30}$$

which after simplifications leveraging the double angle formula becomes

$$\begin{aligned}
\log p^*(\phi_{\mathrm{d}} | \boldsymbol{\phi}_{\ell \not= \mathrm{d}}) = &\cos \phi_{\mathrm{d}} \Bigg[\kappa_{\mathrm{d}} \cos \mu_{\mathrm{d}} - \frac{1}{2} \sum_{\substack{\ell=1 \\ \ell \neq \mathrm{d}}}^{\mathrm{D}} (\mathbf{K}^{-1})_{\mathrm{d},\ell} \cos \phi_{\ell} + (\mathbf{K}^{-1})_{\mathrm{d},\ell+\mathrm{D}} \sin \phi_{\ell} \Bigg] \\
&+ \sin \phi_{\mathrm{d}} \Bigg[\kappa_{\mathrm{d}} \sin \mu_{\mathrm{d}} - \frac{1}{2} \sum_{\substack{\ell=1 \\ \ell \neq \mathrm{d}}}^{\mathrm{D}} (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell} \cos \phi_{\ell} + (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell+\mathrm{D}} \cos \phi_{\ell} \Bigg] \\
&+ \cos 2\phi \Bigg[\frac{(\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}} - (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\mathrm{d}+\mathrm{D}}}{4} \Bigg] \\
&+ \sin 2\phi_{\mathrm{d}} \Bigg[\frac{(\mathbf{K}^{-1})_{\mathrm{d},\mathrm{d}+\mathrm{D}}}{2} \Bigg].
\end{aligned} \tag{4.31}$$

Taking the expectation of Equation (4.31) with respect to the distributions of the other factors results in the expressions from Equation (4.25) to Equation (4.28).

The parameter dependencies shown from Equation (4.25) to Equation (4.28) suggest that the mean field conditionals may admit further simplifications when the matrix $\mathbf{K}$ has special structure. In particular, when

$$\mathbf{K} = \begin{bmatrix} \mathbf{Q} & 0 \\ 0 & \mathbf{Q} \end{bmatrix} \tag{4.32}$$

for some positive-definite matrix $\mathbf{Q}$, the $m\mathcal{G}v\mathcal{M}$ becomes the particular case of the Toroidal Normal distribution and the second-harmonic terms of the $\mathcal{G}v\mathcal{M}$ approximations vanish. Hence, the mean field model becomes a product of von Mises conditionals where

$$q(\phi_{\mathrm{d}}|\boldsymbol{\phi}_{\neq \mathrm{d}}) = v\mathcal{M}\left(\phi_{\mathrm{d}}; \eta_{1,\mathrm{d}}, \nu_{1,\mathrm{d}}\right). \tag{4.33}$$

Here, the parameters $\eta_{1,\mathrm{d}}$ and $\nu_{1,\mathrm{d}}$ are as defined by Equation (4.25) and Equation (4.26). After applying the simplifications that arise from the kernel structure, Equation (4.25) and Equation (4.26) become

$$\begin{aligned} \eta_{1,\mathrm{d}} \cos \nu_{1,\mathrm{d}} &= \kappa_{\mathrm{d}} \cos(\mu_{\mathrm{d}}) - \frac{1}{2} \sum_{\ell \neq \mathrm{d}}^{\mathrm{D}} \left\langle (\mathbf{K}^{-1})_{\mathrm{d},\ell} \cos \boldsymbol{\phi}_{\ell} \right\rangle_{q_{\ell}} \\ \eta_{1,\mathrm{d}} \sin \nu_{1,\mathrm{d}} &= \kappa_{\mathrm{d}} \sin(\mu_{\mathrm{d}}) - \frac{1}{2} \sum_{\ell \neq \mathrm{d}}^{\mathrm{D}} \left\langle (\mathbf{K}^{-1})_{\mathrm{d}+\mathrm{D},\ell+\mathrm{D}} \sin \boldsymbol{\phi}_{\ell} \right\rangle_{q_{\ell}}. \end{aligned} \tag{4.34}$$

To compute the expectations required in the mean field approximations, we leverage the expressions available for the trigonometric moments[2] of the $\mathcal{G}v\mathcal{M}$ and $v\mathcal{M}$.

In practice, we found the series expansions provided by Gatto (2008) for the trigonometric moments associated with the $\mathcal{G}v\mathcal{M}$ were numerically unreliable for large concentration parameters $\eta_{1,\mathrm{d}}$ and $\eta_{2,\mathrm{d}}$. The situation of concentrated data is common in practice as discussed by Sra (2012) in the context of estimating von Mises-Fisher parameters. Empirically, we verified that by either capping the values of the concentration parameters or using instead a sub-optimal von Mises mean field provided reasonable results when approximating a general $m\mathcal{G}v\mathcal{M}$. For a more extensive analysis of the computation of the trigonometric moments of $\mathcal{G}v\mathcal{M}$ using the series expansions and how to render them tractable, see Appendix C.

[2]Recall from Chapter 1 that the usual n-th order moments for a circular distribution $p(\phi)$ are defined as expectations over the $\mathbb{E}[\cos n\phi]$ and $\mathbb{E}[\sin n\phi]$.

4.2 Experimental Results

Experimental results in this chapter are segregated into both regression, presented in Section 4.2.1, and latent variable modelling, presented in Section 4.2.2. In subsequent chapters, as the overall focus shifts towards non-naive approaches to approximating the posterior, the experiments are limited only to the regression settings for convenience.

4.2.1 Circular regression

In this section, we investigate the advantages of employing the $m\mathcal{G}v\mathcal{M}$ regression model discussed in Chapter 2 over two common approaches to handling circular data in machine learning contexts.

The first approach is to ignore the circular nature of the data and fit a non-circular model. This approach is not infrequent as it is reasonable in contexts where angles are constrained to a small region of the unit circle and there is no wrapping. A typical example of the motivation for such models is the use of a first-order Taylor approximation to the rate of change of an angle as can be found in classical aircraft control applications (see, e.g. Skogestad and Postlethwaite, 2005). To represent this approach to modelling, we will fit a one-dimensional GP (1D-GP) to the data sets.

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

Five data sets were used in this evaluation as outlined in Table 4.1. A Wrapped hat data set generated by wrapping a Mexican hat function around the unit circle, a dataset consisting Uber ride requests in NYC in April 2014[3], the tide levels predictions from the UK Hydrographic Office in 2016[4] as function of the latitude and longitude of a given port, the first side chain angle of aspartate as a function of backbone angles in proteins (Harder et al., 2010), and yeast cell cycle phase as a function of gene expression (Santos et al., 2015).

To assess how well the fitted models approximate the distribution of the data, a subset of the data points was kept for validation and the models scored in terms of

[3]https://github.com/fivethirtyeight/uber-tlc-foil-response
[4]http://www.ukho.gov.uk/Easytide/easytide/SelectPort.aspx

Table 4.1 Information pertaining the data sets for the regression experiments including number of training points (N_{train}), number prediction locations (N_{pred}), input dimensions (D) and Kernel used in the regression.

Data set	N_{train}	N_{pred}	D	Kernel
Wrapped hat	14	100	1	SE + White noise
Tides	182	175	2	SE + White noise
Protein	1.000	1.000	2	SE $\times$ Periodic
Yeast	10	10	4.490	Linear
Uber	10.000	5.000	2	SE + White Noise

Table 4.2 Log-likelihood score for regression with the mGvM, 1D-GP and 2D-GP on validation data.

Data set	$m\mathcal{G}v\mathcal{M}$	1D-GP	2D-GP
Wrapped hat	$+\mathbf{2.02 \cdot 10^4}$	$-1.62 \cdot 10^3$	$+8.28 \cdot 10^2$
Uber	$+\mathbf{3.29 \cdot 10^4}$	$-1.49 \cdot 10^3$	$-2.83 \cdot 10^2$
Tides	$+\mathbf{1.25 \cdot 10^4}$	$-6.46 \cdot 10^4$	$-8.41 \cdot 10^1$
Protein	$+\mathbf{1.42 \cdot 10^5}$	$-3.34 \cdot 10^5$	$+1.28 \cdot 10^5$
Yeast	$+\mathbf{1.33 \cdot 10^2}$	$-1.46 \cdot 10^2$	$-1.65 \cdot 10^1$

the log likelihood of the validation data set. To guarantee fairness in the comparison, the likelihood of the 2D-GP was projected back to the unit circle by numerically marginalising the radial component of the model for each point. This converts the 2D-GP into a one-dimensional projected Gaussian distribution over angles. The results are summarised in Table 4.2.

The results shown in Table 4.2 indicate that the $m\mathcal{G}v\mathcal{M}$ provides a better overall fit than the 1D-GP and the 2D-GP in in all experiments in terms of the validation log-likelihood. The 1D-GP approach performs poorly in every case studied as it cannot account for the wrapping behaviour of circular data and it assigns all wrapping behaviour to its noise component. The 2D-GP performs better than the 1D-GP, however in the Uber, Tides and Yeast datasets and its performance is substantially closer to the one presented by the 1D-GP case rather than the mGvM. The Wrapped hat fits are examined in Figure 4.2 and the Tides dataset conclusions shown in Figure 4.3. From the Wrapped hat dataset, it is possible to conclude that the 2D-GP may learn a different, yet reasonable, underlying function. It should also be noted that as discussed during the derivations for the mean field updates, the fit is over-confident as seen by the highly concentrated probability regions between $x = 0.25$ and $x = 0.50$. However, the 2D-GP performs better in terms of predictive performance 3 of the 5 datasets

Table 4.3 Root Mean Squared Error for regression with the mGvM, 1D-GP and 2D-GP on validation data. To make the error amenable to the angle space, the error is taken as the norm of the difference of data vector $[\cos\psi^*, \sin\psi^*]$ and the predictions $[\cos\phi^*, \sin\phi^*]$.

Data set	$m\mathcal{G}v\mathcal{M}$	1D-GP	2D-GP
Wrapped hat	0.53	1.36	**0.11**
Uber	1.32	1.26	**1.25**
Tides	1.46	1.43	**1.38**
Protein	**0.81**	1.25	1.40
Yeast	**0.79**	1.38	1.37

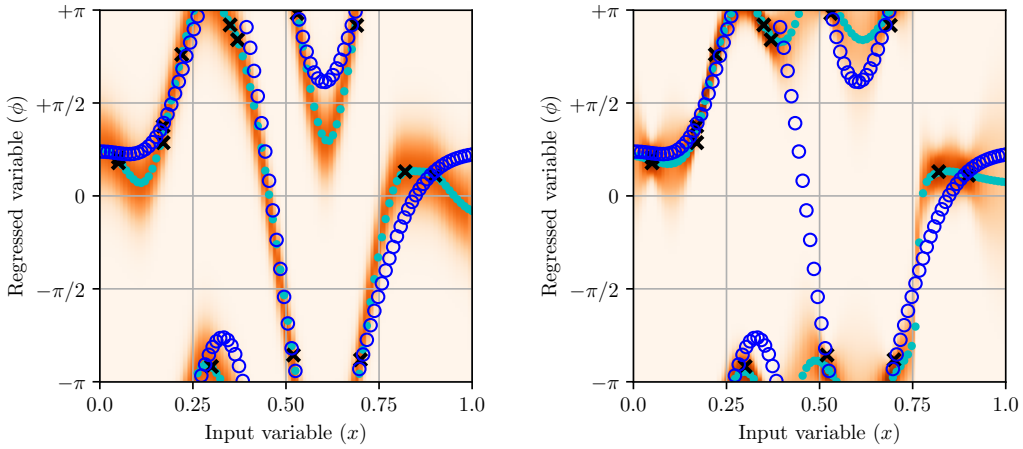


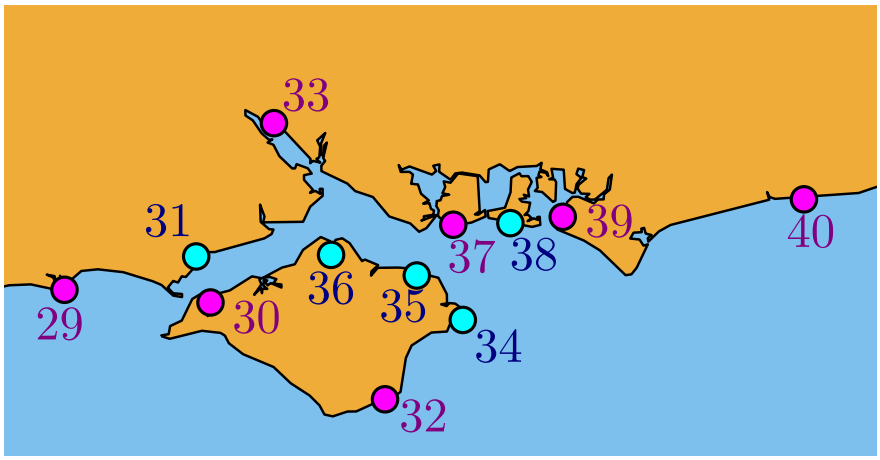
Fig. 4.2 Regression on a Wrapped hat data set using the mGvM (left) and 2D-GP (right): data points are denoted by crosses, the true function by circles and predictions by solid dots.

as shown in Table 4.3. It is important to highlight that even in the cases where the $m\mathcal{G}v\mathcal{M}$ is outperformed in terms of errors, the predictive distribution better represents the underlying data than the GP models.

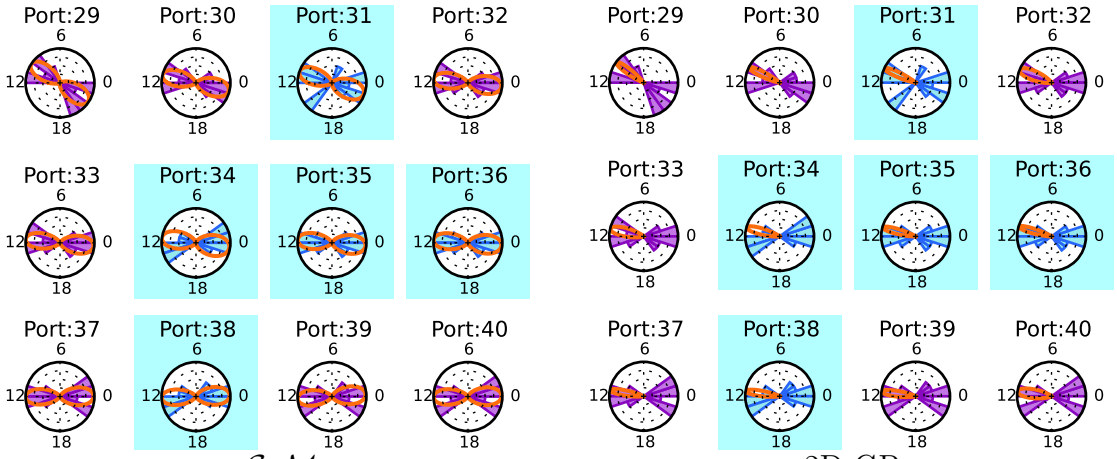
From the fit obtained for the Tides dataset as shown in Figure 4.3, it is possible to establish further that the 2D-GP case can never account for bimodality in the data, whereas by considering a Generalised von Mises likelihood as discussed in Chapter 2 we can model this data characteristic.

4.2.2 Latent variable modelling

To demonstrate the dimensionality reduction application, we analysed two data sets: one motion capture dataset comprising marker positions placed on a subject's arm and



Selected port locations



$m\mathcal{G}v\mathcal{M}$ 2D-GP

Fig. 4.3 Tide time predictions on the UK coast: port location for a subset of the data set (left), $m\mathcal{G}v\mathcal{M}$ fit (left) and 2D-GP (right). The ports whose data was supplied for training are displayed in magenta (darker) rose diagrams whereas the ports held out for prediction are displayed in cyan (lighter). The regression model predictive density is plotted as the orange lines.

Fig. 4.4 Capturing 2D motion: the datasets was generated by recording the motion of a subject with markers on its body then using a colour threshold algorithm and taking the location of the centre of mass of the filtered region.

captured through a low resolution camera as depicted in Figure 4.4, while another set comprising a noisy simulation of a 4-DOF robot arm was generated under the same noise corruption conditions of the motion capture. In the motion capture data sets, we applied a colour filter to the resulting images to isolate each marker. Then, each marker's position was found by calculating the centre of mass of each marker as shown in Figure 4.4. We compared the model using point estimates for the matrices $\mathbf{A}$ and $\mathbf{B}$, a variational Bayes approach by including ARD priors for $\mathbf{A}$ and $\mathbf{B}$, Probabilistic Principal Component Analysis (PPCA) (Tipping and Bishop, 1999) and the Gaussian Process Latent Variable Model (GP-LVM) (Lawrence, 2004) using a squared exponential kernel and a linear kernel. The models using the mGvM require special attention to initialisation. To initialise the test, we used a greedy clustering algorithm to estimate the matrices $\mathbf{A}$ and $\mathbf{B}$. The variational Bayes model was initialised using the learned parameters for the point estimate model.

The performance of each model was assessed by denoising the original dataset corrupted by additional Gaussian noise of 2.5, 5 and 10 pixels and comparing the signal-to-noise ratio (SNR) on a test dataset. The best results after initializing the models at 3 different initial starting points are summarized in Table 4.4 and additional experiments for a wider range of noise levels are available in Figure 4.5.

In Table 4.4, the point estimate cPCA model performs best and is followed by its variational Bayes version for both datasets. The variational Bayes performance is only worse in the motion capture case, whereas it not significantly different in the robot case. This is likely to be an effect of isotropic noise assumption not being valid under the motion capture dataset. By violating this assumption, the structure of the pendulum that gave rise to the data becomes difficult to identify. To further expand on this point, if the noise is assumed fixed, but indeed changes through time, an

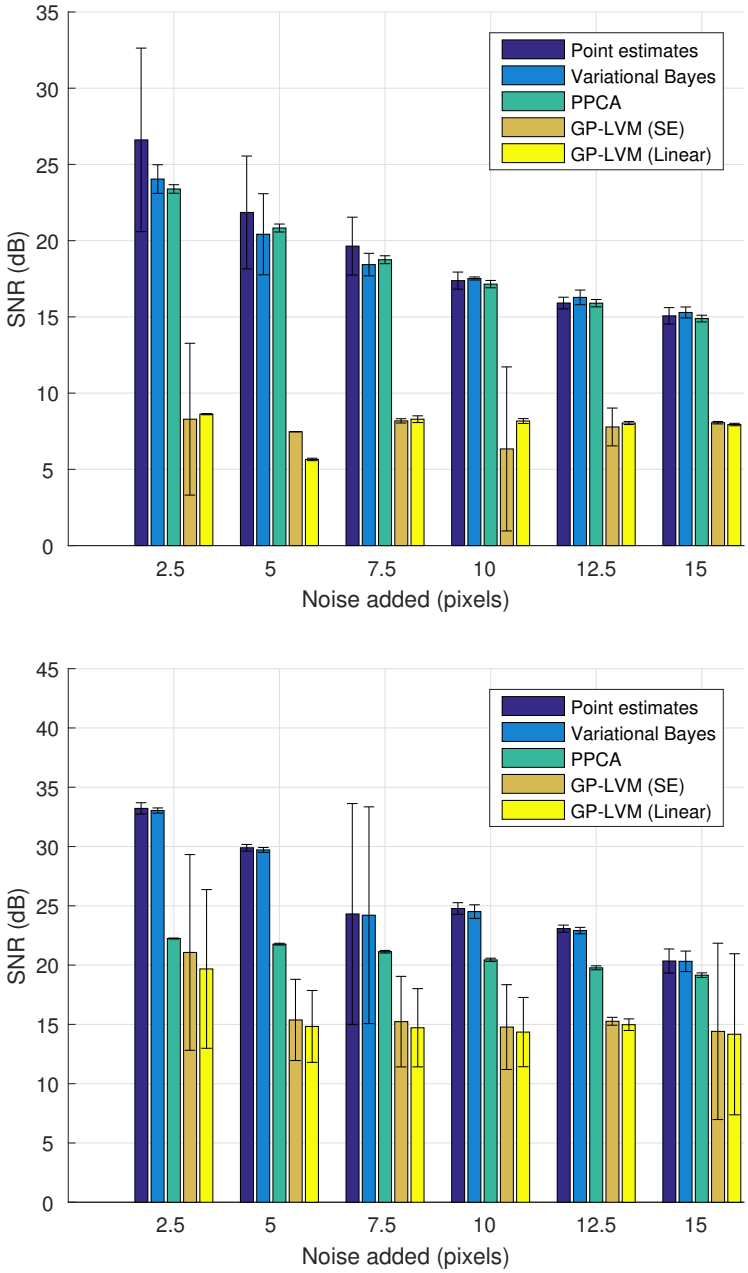


Fig. 4.5 Signal-to-noise ratio with 3 standard deviations for the latent variable modelling datasets: filmed subject (top) and simulated robot arm dataset (bottom).

Table 4.4 Signal-to-noise ratio (dB) of the learned latent structure after denoising corrupted signals with by Gaussian noise.

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

N-joint pendulum can be more simply explained by a $N+1$-joint pendulum, where the extra joint is not observed. The representation of this extra pendulum joint makes matrices $\mathbf{A}$ and $\mathbf{B}$ assume a different form. This phenomenon can account for the poorer performance of variational bayes compared to point estimates. Furthermore, in the motion capture dataset, as the latent angles are highly concentrated. Under these circumstances, the small-angle approximation for sine and cosine provides good results and the cPCA model degenerates into the PPCA model as shown previously. This behaviour is reflected in the proximity of the PPCA and cPCA signal to noise ratios in Table 4.4. In the robot dataset, the latent angles are less concentrated. As a result, the behaviour of the PPCA and cPCA models is different which explains the larger gap between the results obtained for these models.

4.3 Summary

In this chapter we introduced the Variational Free Energy framework for performing inference in both the regression and the latent variable model setting of the $m\mathcal{G}v\mathcal{M}$. We derived the mean-field distribution for the $m\mathcal{G}v\mathcal{M}$ and showed that despite its shortcomings, it is able to outperform standard Euclidean approaches for modelling circular data.

Moreover, the results obtained by the mean field approach developed in this chapter establish an important benchmark for comparing other models both in terms of quantitative predictive performance as well as qualitative evaluating the quality of the posterior approximation.

Chapter 5

Expectation Propagation inference for the $m\mathcal{G}v\mathcal{M}$

In the previous chapter, we introduced the Variational Free Energy framework and approximate inference for the multivariate Generalised von Mises using Mean-Field Variational Inference. Despite the efficiency and convergence guarantees of such approach, a fully-factored von Mises approximation to a $m\mathcal{G}v\mathcal{M}$ posterior is necessarily unimodal. While this property is not a problem for Toroidal Normal models, they will in general yield pauper approximations to a general posterior. It may also be argued that even for Toroidal Normal models, the fact that the approximations are fully factored add to the well-known tendency of Variational Free Energy methods to underestimate the posterior variance discussed in Section 4.1.1.

To mitigate these problems, in this chapter, we turn our attentions to forming approximations of a $m\mathcal{G}v\mathcal{M}$ using Expectation Propagation (EP). While EP methods are only guaranteed to converge distributions than can be posed as a acyclic exponential family graphical model, they are often able to provide a better characterisation of the approximated distribution than mean field variational inference. In particular, we draw on the approximations by Opper and Winther (2005) to produce an algorithm for performing inference on the $m\mathcal{G}v\mathcal{M}$. We also show the connections between the augmented representation of the $m\mathcal{G}v\mathcal{M}$ proposed in Chapter 3 and the inference algorithm presented.

5.1 Inference

In this section, we provide a short introduction to Expectation Propagation and its relationship to variational inference in Section 5.1.1, followed by the derivation and

discussion of more sophisticated approaches in Section 5.1.2 and Section 5.1.2. The content of Section 5.1.1 is directed to the EP novice, and the experienced reader may skip it altogether.

5.1.1 Introduction to EP and its connection to variational inference

The Expectation Propagation algorithm was initially proposed by Minka (2000) as an extension of both the Assumed Density Filter and Belief Propagation algorithms. Algorithmically, EP inference can be succinctly described as forming a product representation of the distribution we wish to approximate

$$p(\mathbf{x}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} f_{\mathrm{n}}(\mathbf{x}), \tag{5.1}$$

constructing an approximation

$$q(\mathbf{x}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} \tilde{f}_{\mathrm{n}}(\mathbf{x}|\boldsymbol{\theta}_{\mathrm{n}}), \tag{5.2}$$

and iteratively refining the approximations by constructing a cavity function

$$q_{\neq \mathrm{d}}(\mathbf{x}) = \prod_{\mathrm{n}\neq \mathrm{d}}^{\mathrm{N}} \tilde{f}_{\mathrm{n}}(\mathbf{x}) \tag{5.3}$$

and obtaining the new parameters $\boldsymbol{\theta}_{\mathrm{d}}^{\star}$ for $\tilde{f}_{\mathrm{d}}$ such that

$$\boldsymbol{\theta}_{\mathrm{d}}^{\star} = \arg\min_{\theta_{\mathrm{d}}} \mathrm{KL}\left(f(\mathbf{x})q_{\neq \mathrm{d}}(\mathbf{x}) \middle\| \tilde{f}(\mathbf{x})q_{\neq \mathrm{d}}(\mathbf{x})\right) \tag{5.4}$$

where KL is the Kullback-Leibler divergence.

It is common to require q to be an exponential family distribution as the optimisation in Equation (5.4) reduces to moment matching. Additionally, when both p and q are exponential family distributions, Expectation Propagation admits a variational interpretation as shown by Wainwright and Jordan (2008). For the remainder of this section, we will briefly review this variational interpretation of EP. When the approximated distribution p admits a exponential family representation, it can be written in the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^{\top}\mathsf{T}(\mathbf{x}) - \log \mathcal{Z}(\boldsymbol{\theta})\right\} \tag{5.5}$$

where the $\boldsymbol{\theta}$ are the parameters of the distribution p, $\mathcal{Z}$ is the partition function associated with p and T are the sufficient statistics of the distribution. Recall the free energy functional presented in Chapter 4

$$\mathcal{F}(q) = \langle \log p(\mathbf{x}|\boldsymbol{\theta}) \rangle_{q(\mathbf{x})} + \mathcal{H}(q), \tag{5.6}$$

and, that in variational inference our objective is to maximise it with respect to the variational approximation q. For the exponential family model of Equation (5.5), this objective becomes

$$\max_q \mathcal{F}(q) = \max_q \boldsymbol{\theta}^\top \langle \mathsf{T}(\mathbf{x}) \rangle_{q(\mathbf{x})} - \log \mathcal{Z}(\boldsymbol{\theta}) + \mathcal{H}(q). \tag{5.7}$$

Wainwright and Jordan (2008) showed that under mild regularity conditions, the Expectation Propagation algorithm can be seen as maximising a relaxed form of the Equation (5.7). If we assume that the approximating distribution q can be factored, we and split the factors into two sets $\mathcal{A}$ and $\mathcal{B}$, the EP relaxation effectively substitutes the Shannon entropy functional $\mathcal{H}$ in Equation (5.7) for the Bethe entropy. The Bethe entropy only into account only pair-wise interactions between all factors of sets $\mathcal{A}$ and $\mathcal{B}$ and is defined as

$$\mathcal{H}_{\text{Bethe}}(\boldsymbol{q}_\mathcal{A}, \boldsymbol{q}_\mathcal{B}) = \mathcal{H}\left(\boldsymbol{q}_\mathcal{A}\right) + \sum_{\mathrm{j} \in \mathcal{B}} \left[\mathcal{H}\left(\boldsymbol{q}_\mathcal{A}, \boldsymbol{q}_\mathrm{j}\right) - \mathcal{H}\left(\boldsymbol{q}_\mathcal{A}\right) \right]. \tag{5.8}$$

If the true distribution p and the approximating distribution q are exponential family members, moment matching between these distributions solves the first-order optimality condition of the Lagrangian associated with the EP free energy, that is,

$$\max_q \mathcal{F}_{\text{EP}}(q) = \max_q \boldsymbol{\theta}^\top \langle \mathsf{T}(\mathbf{x}) \rangle_{q(\mathbf{x})} - \log \mathcal{Z}(\boldsymbol{\theta}) + \mathcal{H}_{\text{Bethe}}(q). \tag{5.9}$$

There are multiple technicalities required to understand the theoretical underpinnings of the approximations outlined, such as their relationship to probability polytopes. Such topics are important in a deeper understanding of Expectation Propagation and other message passing algorithms, however such discussions lie beyond the scope of a first introduction to these methods. Therefore, we have opted for providing the reader with an intuition and referring the interested reader to Wainwright and Jordan (2008) and Murphy (2012) for a thorough exposition of these technicalities.

In the context of circular statistics, EP updates for the multivariate von Mises have been derived by Razavian et al. (2011) and Razavian et al. (2012). However, despite

considerable efforts, we could neither obtain converging results using the algorithm outlined by these researchers nor a naïve implementation of the EP algorithm for the $m\mathcal{G}v\mathcal{M}$ even when annealing and damped updates are introduced. This lack of convergence of such algorithms can be traced to the fact that they utilise mean-field type of factorisation for the $m\mathcal{G}v\mathcal{M}$ and such approximations have been shown to fail, for example, the multivariate Gaussian case is discussed by Cseke and Heskes (2008).

Another limitation of these methods is that they consider every factor to be an independent von Mises distribution. This choice of approximation is problematic both in terms of representation power and numerical stability. Representations arising from a product of independent von Mises are limited to unimodal, symmetric distributions which does not capture the multimodality of the $m\mathcal{G}v\mathcal{M}$. Numerical issues also arise since moment matching requires inverting modified Bessel functions of first kind, since their inverse does not bear an analytic form. Such quantities need to be estimated through a root-finding procedure such as the Newton-Raphson method requiring multiple evaluations of Bessel functions of possibly high-valued arguments. The use of such methods adds another source of numerical stability and can be slow.

In the following sections, we will derive two alternative factorised representations of EP and how to do moment matching in the $m\mathcal{G}v\mathcal{M}$ through a lifting transformation.

5.1.2 Lifted Expectation Propagation approximation for the $m\mathcal{G}v\mathcal{M}$

Drawing on previous work in statistical physics, Opper and Winther (2005) proposed structuring Expectation Propagation factors of an intractable density $p(\mathbf{x})$ into two parts, one a collection of analytically-tractable factors $f_s(\mathbf{x})$, while the second part agglomerates constraint-encoding factors $f_c(\mathbf{x})$, which are defined through the use of Dirac Deltas.

Such constructions can prove useful when distributions can be posed in terms of a series of transformations in generative models. For example, if we adopt the generative view of the $m\mathcal{G}v\mathcal{M}$ distribution provided in Section 2.1.1, a random vector $\boldsymbol{\phi}$ following a $m\mathcal{G}v\mathcal{M}$ distribution can be viewed as a 2N-variate Gaussian conditioned to the

N-torus,

$$p(\mathsf{x}) = \mathcal{N}(\mathsf{x}; \boldsymbol{m}, \boldsymbol{\Sigma}) \tag{5.10}$$

$$p(\mathsf{z}|\mathsf{x}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} \delta\left(\mathsf{z}_{\mathrm{n}}^2 + \mathsf{z}_{\mathrm{N+n}}^2 - 1\right) \delta\left(\mathsf{z}_{\mathrm{n}} - \mathsf{x}_{\mathrm{n}}\right) \delta\left(\mathsf{z}_{\mathrm{n+N}} - \mathsf{x}_{\mathrm{N+n}}\right) \tag{5.11}$$

$$p(\boldsymbol{\phi}|\mathsf{z}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} \delta(\cos\phi_{\mathrm{n}} - \mathsf{z}_{\mathrm{n}})\delta(\sin\phi_{\mathrm{n}} - \mathsf{z}_{\mathrm{N+n}}), \tag{5.12}$$

such that marginal distribution for $\boldsymbol{\phi}$ can be recovered by marginalising both z and x,

$$p(\boldsymbol{\phi}) = \int p(\boldsymbol{\phi}|\mathsf{z})p(\mathsf{z}|\mathsf{x})p(\mathsf{x})\mathrm{d}\mathsf{x}\mathrm{d}\mathsf{z}. \tag{5.13}$$

$$= \int p(\boldsymbol{\phi}|\mathsf{x})p(\mathsf{x})\mathrm{d}\mathsf{x}. \tag{5.14}$$

A key insight of this factorisation is that Expectation Propagation can be performed in the *lifted* space comprising both the simple variates and the constraints, i.e., the tuple $(\boldsymbol{\phi}, \mathsf{z}, \mathsf{x})$ in the $m\mathcal{G}\upsilon\mathcal{M}$ case. Each of the Dirac Delta constraints are then approximated by Gaussians centred at the Delta's argument. These approximations lead to an interpretation that Opper and Winther (2005)'s factorisation is, essentially, an approximate inference procedure over a relaxed version of the original constrained problem. Moreover, because the Gaussian relaxations of the Dirac Delta's have their means at the exact constraint locations, it can be proved that the expectations of the factored approximation will be consistent with the original model (hence the name "expectation-consistent" approximations used in the original paper).

Specifically for the $m\mathcal{G}\upsilon\mathcal{M}$ case, we can derive the updates over a simpler lifted space by analytically integrating over the $\boldsymbol{z}$ as in Equation (5.14), that is, we can consider a lifted space comprised of

$$\begin{aligned} p(\mathsf{x}) &= \mathcal{N}(\mathsf{x}; \boldsymbol{m}_p, \boldsymbol{\Sigma}_p) \\ p(\boldsymbol{\phi}|\mathsf{x}) &= \prod_{\mathrm{n}=1}^{\mathrm{N}} \delta(\cos\phi_{\mathrm{n}} - \mathsf{x}_{\mathrm{n}})\delta(\sin\phi_{\mathrm{n}} - \mathsf{x}_{\mathrm{N+n}})\delta\left(\mathsf{x}_{\mathrm{n}}^2 + \mathsf{x}_{\mathrm{N+n}}^2 - 1\right). \end{aligned} \tag{5.15}$$

Under the lifted view of the augmented space, we can write a von Mises or a Generalised von Mises likelihood term for each data point ψ as

$$\begin{aligned} p(\mathsf{x}) &= \mathcal{N}(\mathsf{x}; \boldsymbol{m}_\ell, \boldsymbol{\Sigma}_\ell) \\ p(\boldsymbol{\psi}|\mathsf{x}) &= \prod_{\mathrm{n}=1}^{\mathrm{N}} \delta(\cos\psi_{\mathrm{n}} - \mathsf{x}_{\mathrm{n}})\delta(\sin\psi_{\mathrm{n}} - \mathsf{x}_{\mathrm{N+n}})\delta\left(\mathsf{x}_{\mathrm{n}}^2 + \mathsf{x}_{\mathrm{N+n}}^2 - 1\right) \end{aligned} \tag{5.16}$$

where $\mathbf{\Sigma}_\ell$ is a $\mathbb{R}^{2\times 2}$ covariance matrix that results in the κ_1 and κ_2 parameters of the $m\mathcal{G}v\mathcal{M}$. Combining Equation (5.15) and Equation (5.16) and drawing on conjugacy of Gaussian distributions, we obtain the posterior model

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}; \tilde{\boldsymbol{m}}, \tilde{\mathbf{K}}) \\ p(\boldsymbol{\phi}|\mathbf{x}) &= \prod_{\mathrm{n}=1}^{\mathrm{N}} \Big[\delta(\cos\phi_{\mathrm{n}} - \mathsf{x}_{\mathrm{n}})\delta(\sin\phi_{\mathrm{n}} - \mathsf{x}_{\mathrm{N+n}}) \\ &\qquad \times \delta(\cos\psi_{\mathrm{n}} - \mathsf{x}_{\mathrm{n}})\delta(\sin\psi_{\mathrm{n}} - \mathsf{x}_{\mathrm{N+n}})\delta\left(\mathsf{x}_{\mathrm{n}}^2 + \mathsf{x}_{\mathrm{N+n}}^2 - 1\right) \Big] \end{aligned} \tag{5.17}$$

for appropriate posterior parameters $\tilde{\mathbf{K}}$ and $\tilde{\boldsymbol{m}}$. This resulting model can be succintly represented as a factor graph in Figure 5.1.

The next step in this algorithmic setup is to construct individual approximations to the constraints involving of Equation (5.17). A natural choice is to use Gaussian distributions over the terms that appear in the argument of the delta functions, for example,

$$\delta(\mathsf{x}_{\mathrm{n}}^2 + \mathsf{x}_{\mathrm{N+n}}^2 - 1) \approx \mathcal{N}\left(\begin{bmatrix} \mathsf{x}_{\mathrm{n}} \\ \mathsf{x}_{\mathrm{N+n}} \end{bmatrix}; \boldsymbol{m}_{\mathrm{n}}, \sigma_{\mathrm{n}}^2 \mathbf{I} \right) \tag{5.18}$$

since $\mathcal{N}(x, 0, \sigma^2) \to \delta(x)$ as $\sigma^2 \to 0$. In the case of the $m\mathcal{G}v\mathcal{M}$'s delta functions, the

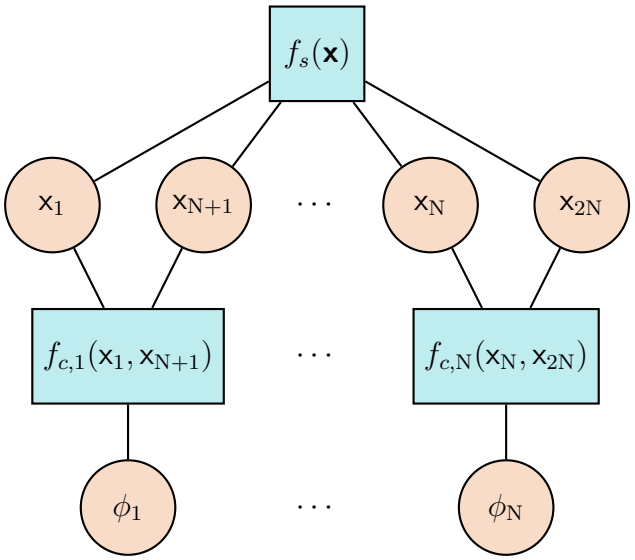


Fig. 5.1 The factor graph for the $m\mathcal{G}v\mathcal{M}$ model under the Lifted approximation indicating the factors f_s comprising the joint Gaussian distribution, and the constraint factors f_c for the delta functions.

$\boldsymbol{m}$ and σ_{n}^2 parameter bears additional significance, as the mean indicate the location

on the unit circle a tangent approximation to the delta function is to be constructed and the variance represents the degree to which the delta function has been relaxed. This interpretation in shown graphically in Figure 5.2

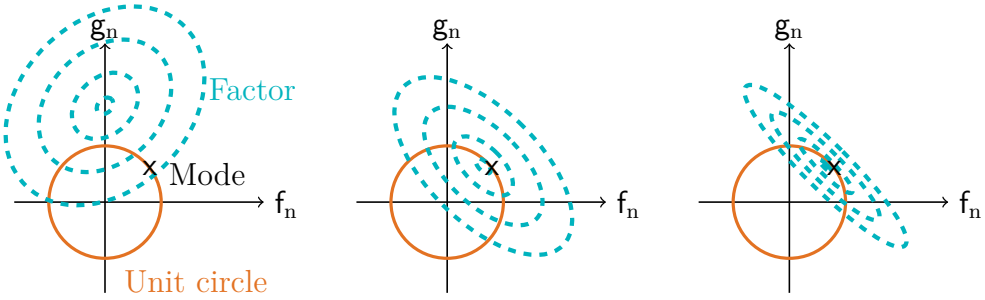


Fig. 5.2 Illustration of different delta function approximations in the Expectation-Consistent framework for the $m\mathcal{G}v\mathcal{M}$: before the algorithm iterates, factors have large variances and are not centred on the unit circle (left), as the algorithm progresses, the approximations tend to assign the factor mean to locations on the unit circle (middle), and final steps focus on reducing the factor variance to avoid assigning high-probability to the regions far from the unit circle (right).

With Equation (5.18), the EP approximation can be simply written as a product of Gaussians

$$\begin{aligned} q(\mathbf{x}) &= p(\mathbf{x}) \prod_{n=1}^{N} \tilde{f}_n(\mathsf{x}_n, \mathsf{x}_{N+n}) \\ &= \mathcal{N}(\mathbf{x}, \boldsymbol{m}, \mathbf{K}) \prod_{n=1}^{N} \mathcal{N}\left(\begin{bmatrix} \mathsf{x}_n \\ \mathsf{x}_{N+n} \end{bmatrix}; \begin{bmatrix} m_{1,n} \\ m_{2,n} \end{bmatrix}, \begin{bmatrix} \sigma_{1,n}^2 & \rho_n \sigma_{1,n} \sigma_{2,n} \\ \rho_n \sigma_{1,n} \sigma_{2,n} & \sigma_{1,n}^2 \end{bmatrix} \right) \end{aligned} \tag{5.19}$$

such that the EP updates adjust the relaxations on the model constraints as shown schematically in Figure 5.2. Notice that the moments of the true underlying distribution which involve delta functions are obtained directly from a $\mathcal{G}v\mathcal{M}$ distribution, since

$$\int \delta\left(\cos\phi_n^2 - \mathsf{x}_n\right) \delta\left(\sin\phi_{N+n}^2 - \mathsf{x}_{N+n}\right) \delta\left(\mathsf{x}_n^2 + \mathsf{x}_n^2 - 1\right) p(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathcal{G}v\mathcal{M}\left(\phi_n; \mu_1, \mu_2, \kappa_1, \kappa_2\right) \tag{5.20}$$

with the parameters are given by Equation (5.21) to Equation (5.24)

$$\kappa_{1,\mathrm{d}} \cos \mu_{1,\mathrm{d}} = 2 \left(\frac{\rho_d}{\sigma_{1,\mathrm{d}} \sigma_{2,\mathrm{d}}} m_{2,\mathrm{d}} - \frac{1}{\sigma_{1,\mathrm{d}}^2} m_{1,\mathrm{d}} \right) \tag{5.21}$$

$$\kappa_{1,\mathrm{d}} \sin \mu_{1,\mathrm{d}} = 2 \left(\frac{\rho_d}{\sigma_{1,\mathrm{d}} \sigma_{2,\mathrm{d}}} m_{1,\mathrm{d}} - \frac{1}{\sigma_{2,\mathrm{d}}^2} m_{2,\mathrm{d}} \right) \tag{5.22}$$

$$\kappa_{2,\mathrm{d}} \cos 2\mu_{2,\mathrm{d}} = \frac{1}{2} \left(\frac{1}{\sigma_{1,\mathrm{d}}^2} - \frac{1}{\sigma_{2,\mathrm{d}}^2} \right) \tag{5.23}$$

$$\kappa_{2,\mathrm{d}} \sin 2\mu_{2,\mathrm{d}} = -\frac{\rho_d}{\sigma_{1,\mathrm{d}} \sigma_{2,\mathrm{d}}}. \tag{5.24}$$

and hence, the moments of the approximating distribution must be matched to the moments of a $\mathcal{G}\upsilon\mathcal{M}$ distribution. We term the approximation derived above as well as its associated Expectation Propagation updates as Lifted Expectation Propagation for the $m\mathcal{G}\upsilon\mathcal{M}$ (LEP-$m\mathcal{G}\upsilon\mathcal{M}$).

The approach described above can be seen as an extension of the algorithm proposed by Turner and Sahani (2011). In their original algorithm, they proposed a constraint approximation to demodulate phase and obtain the signal amplitude using a constrained Markov process. The algorithm presented in this chapter differs from this previous work both in focus and generality. Here, we focus only on modelling the circular component and further constrain the amplitude to unity. At the same time, we extend the Markov framework to allow more general correlation structures present in Gaussian processes.

Relationship to the augmented representation of the $m\mathcal{G}\upsilon\mathcal{M}$

Recall that the augmented representation of the $m\mathcal{G}\upsilon\mathcal{M}$ distribution presented in the previous chapter eliminates the difficulties associated with the $m\mathcal{G}\upsilon\mathcal{M}$'s quadratic term by introducing additional Gaussian states. Applying this augmentation to a posterior distribution over $\boldsymbol{\phi}$, $\mathbf{f}$ and $\mathbf{g}$ results in

$$p(\boldsymbol{\phi}, \mathbf{f}, \mathbf{g} | \boldsymbol{\psi}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} ; \boldsymbol{m}, \boldsymbol{\Sigma} \right) \times \prod_{\mathrm{n}=1}^{\mathrm{N}} \upsilon\mathcal{M}(\phi_{\mathrm{n}}; \alpha_{\mathrm{n}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\psi}), \beta_{\mathrm{n}}(\mathbf{f}, \mathbf{g}, \boldsymbol{\psi}) \tag{5.25}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\sigma^2, \mathbf{K}^{-1}, \mathbf{A})$ for $\mathbf{A}$ an arbitrary invertible real-valued matrix, σ^2 is a positive real constant such that $\sigma^2 - \mathbf{K}^{-1}$ is positive-definite, $\boldsymbol{m}$ is a mean vector, and the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are such that

$$\begin{bmatrix} \boldsymbol{\alpha} \odot \cos \boldsymbol{\beta} \\ \boldsymbol{\alpha} \odot \sin \boldsymbol{\beta} \end{bmatrix} = \kappa \begin{bmatrix} \cos \boldsymbol{\psi} \\ \sin \boldsymbol{\psi} \end{bmatrix} + \mathbf{A}^{-1} \left(\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} - \boldsymbol{m} \right) \tag{5.26}$$

with $\odot$ denoting point-wise multiplication.

Based on the transformation employed in the previous section, the circular variables model from Equation (5.25) can be approximated by treating each ϕ_{n} as the angle arising from two Gaussian components u_{n} and v_{n} to yield the model

$$p\Big(\boldsymbol{\phi}, \mathbf{f}, \mathbf{g}\Big) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix}; \boldsymbol{m}, \boldsymbol{\Sigma}\right) \times \prod_{\mathrm{n}=1}^{\mathrm{N}} [\delta(\cos\phi_{\mathrm{n}} - \mathsf{u}_{\mathrm{n}})\delta(\sin\phi_{\mathrm{n}} - \mathsf{v}_{\mathrm{n}})\delta(\mathsf{u}_{\mathrm{n}}^2 + \mathsf{v}_{\mathrm{n}}^2 - 1)] \times \mathcal{N}\left(\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}; \begin{bmatrix}\cos\boldsymbol{\beta}\\ \sin\boldsymbol{\beta}\end{bmatrix}, \begin{bmatrix}\mathrm{diag}(\boldsymbol{\alpha})^{-1} & 0\\ 0 & \mathrm{diag}(\boldsymbol{\alpha})^{-1}\end{bmatrix}\right). \tag{5.27}$$

The Gaussian distributions of Equation (5.27) can be viewed through the eyes of the Expectation Propagation approximation as

$$\begin{aligned} p\Big(\boldsymbol{\phi}, \mathbf{f}, \mathbf{g}\Big) = & \left[\prod_{\mathrm{n}=1}^{\mathrm{N}} \delta(\cos\phi_{\mathrm{n}} - \mathsf{u}_{\mathrm{n}})\delta(\sin\phi_{\mathrm{n}} - \mathsf{v}_{\mathrm{n}})\delta(\mathsf{u}_{\mathrm{n}}^2 + \mathsf{v}_{\mathrm{n}}^2 - 1)\right] \\ & \times \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{g}\end{bmatrix}; \tilde{\boldsymbol{m}}(\mathbf{u}, \mathbf{v}, \boldsymbol{\psi}), \tilde{\boldsymbol{\Sigma}}\right) \end{aligned} \tag{5.28}$$

for appropriate $\tilde{\boldsymbol{m}}$ and $\tilde{\boldsymbol{\Sigma}}$. Equation (5.28) defines a model very similar to that approximated by Equation (5.15) and Equation (5.16). This shows that both models are inherently connected, despite adopting different routes for approximating a $m\mathcal{G}v\mathcal{M}$ model.

Different representations can induce different types of related EP algorithms, as portrayed in Figure 5.3. However not all representations are tractable or yield efficient algorithms. From the algorithms outlined in Figure 5.3, only the LEP-$m\mathcal{G}v\mathcal{M}$ algorithm has an efficient scheme for matching moments. For example, while technically feasible, direct EP for the augmented representation of the $m\mathcal{G}v\mathcal{M}$ requires estimating the moments of the underlying distribution with Monte Carlo methods and as a consequence its performance can be slowed significantly when compared to LEP-$m\mathcal{G}v\mathcal{M}$.

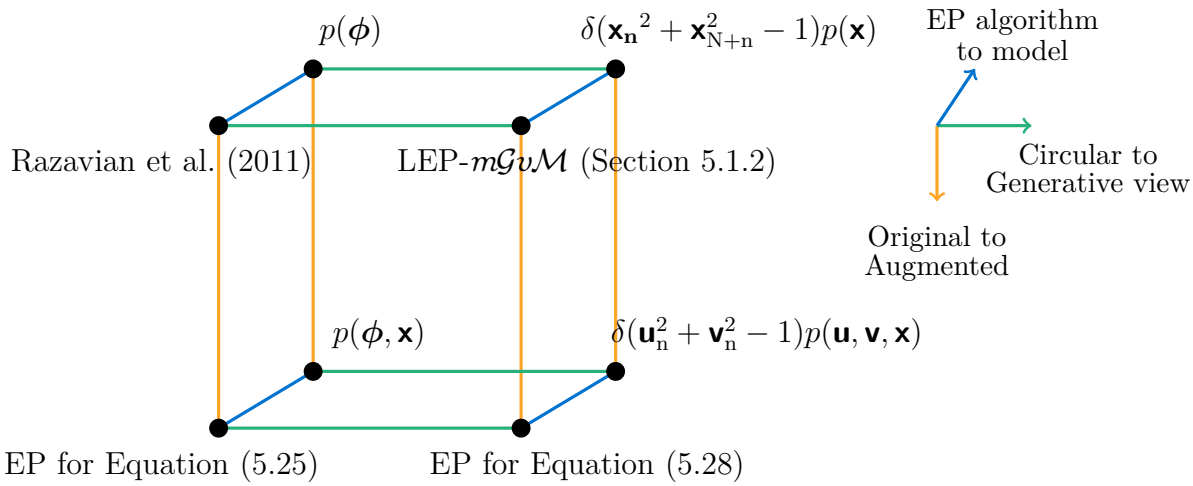


Fig. 5.3 A diagram outlining the relationship between the different Expectation Propagation algorithms, the $m\mathcal{G}\upsilon\mathcal{M}$ models in original circular form and constrained $\mathcal{GP}$ representations, and augmented representations.

5.2 Experimental results

In this section, we present experiments on synthetic and real data sets for the Expectation Propagation algorithm for the $m\mathcal{G}\upsilon\mathcal{M}$ and compare this method's performance to mean-field variational inference. In particular, we evaluate the predictive performance in both quantitative and qualitative grounds using the Mean-Field Variational Inference outlined in Chapter 4 as a comparison reference. To establish a fair comparison, we employ the same data sets and model parameters and also evaluate the total running time pertaining both inference algorithms. We note that while a comparison with naïve applications of Expectation Propagation was attempted based on the EP updates derived by Razavian et al. (2011) and Razavian et al. (2012), we could not obtain converging results using the algorithm outlined by these researchers[1].

Quantitative results for both the inference algorithms are presented in Table 5.1 in terms of the predictive likelihood of a test data set and a running time comparison between both algorithms. Table 5.1 shows that the Expectation Propagation algorithm is slower than the Mean-Field Variational Inference. Profiling the source code of both implementations reveals that the major source for this time discrepancy is the moment calculation step. This discrepancy arises because the Lifted EP algorithm requires

[1]For reproducibility purposes, it is important to remark that we could not rely on the original implementation of these methods. The original implementation was not openly available, nor was it supplied upon request at the time of the writing of this thesis. The results shown correspond to our implementation of the EP updates described in their papers.

Table 5.1 Log-likelihood of the predictions with the mGvM using different algorithms and ratio between the running time of the Lifted Expectation Propagation for the $m\mathcal{G}v\mathcal{M}$ algorithm (LEP-$m\mathcal{G}v\mathcal{M}$) to Mean-Field Variational Inference (MF-VI). Instances of EP that did not converge even in the presence of both annealing and damping are indicated in the table

Data set	MF-VI	Razavian et al. (2011)	LEP-$m\mathcal{G}v\mathcal{M}$	Time Ratio
Wrapped hat	$2.02 \cdot 10^4$	Not Converged	$\mathbf{3.60 \cdot 10^4}$	4.23
Yeast	$1.33 \cdot 10^2$	Not Converged	$\mathbf{1.34 \cdot 10^2}$	3.77
Tides	$1.25 \cdot 10^4$	Not Converged	$\mathbf{3.45 \cdot 10^4}$	1.37
Protein	$1.42 \cdot 10^5$	Not Converged	$\mathbf{2.03 \cdot 10^5}$	$1.47 \cdot 10^2$
Uber	$\mathbf{3.29 \cdot 10^4}$	Not Converged	$1.68 \cdot 10^4$	$2.06 \cdot 10^1$

matching moments of a Generalised von Mises computed through series of modified Bessel functions, whereas the moment calculations in the Mean-Field Variational Inference algorithm rely solely on a simple ratio of modified Bessel functions. The likelihood scores in Table 5.1 show that in general the Expectation Propagation algorithm has superior predictive performance when compared to the MF-VI. This characteristic is a consequence of a better representation of uncertainty surrounding the true underlying function as evidenced by Figure 5.4 We conjecture that the improved performance can be traced to the lifting procedure. This procedure is similar to the one used when fitting the 2D-GP model discussed in the previous chapter, which allows the Lifted EP model to incorporate characteristics from both the uncertainty representation of the 2D-GP with the mean estimates from the MF-VI. An interesting point for further analysis is to investigate whether analytic relationships between the MF-VI, the 2D-GP and the Lifted EP can be established.

It is also possible to view in Figure 5.4 that the underlying mean function learned by LEP-$m\mathcal{G}v\mathcal{M}$ is different from the learned mean function of MF-VI, which the MF-VI over-confidently commits to a single regressed function, without properly weighing the possibility for different functions to fit the structure as discussed in Chapter 4. The Lifted EP algorithm is, instead, forced to assign probability mass asymmetrically to contemplate the hypothesis that the underlying function has wrapped or not at a given point, and weight such hypothesis adequately. This behaviour leads to the differences in the error bars illustrated in Figure 5.4. A similar behaviour is also noted in the Tides data set as demonstrated in Figure 5.5.

It was necessary to introduce damping in the EP updates and anneal the strictness of the constraints to the unit circle to obtain the results above. This annealing procedure consisted of setting the variance terms of the delta factors to fixed values

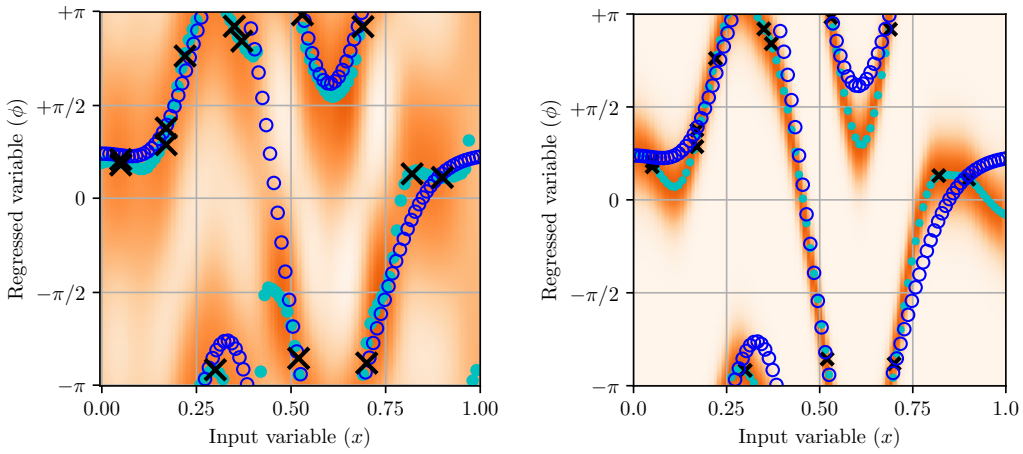


Fig. 5.4 Regression on the Wrapped Mexican Hat function using LEP-$m\mathcal{G}v\mathcal{M}$ (left) and MF-VI (right): data points are denoted by crosses, the true function by circles and mean of each factor for the prediction locations by solid dots (darker regions have higher probability than lighter regions).

and progressively decreasing them until it was completely absent. We denote a set of values used in this procedure as an annealing schedule. Multiple annealing schedules were attempted by considering 10 different variance settings taking values between 10 and 0.01, initially starting with a linear profile and then manually adjusting the values on a log-scale grid. From all the schedules tested, the one which produced the best overall predictive value on a test data set was retained. This procedure is illustrated in Algorithm 2.

Algorithm 2: Annealing and damping scheme utilised in the $m\mathcal{G}v\mathcal{M}$ Expectation Propagation schemes

```
getInput (epFactors, dampingValues, annealValues); // Supply initial values
  for the EP factor parameters, a list of damping values to be used
  and a list of annealing values to be used
for α ∈ dampingValues do
    for β ∈ annealValues do
        epFactors = runEP (noise=β, damping=α, initialFactors=epFactors) ;
         // Run the EP algorithm
    end
end
Output epFactors
```

Annealing has a severe impact on the function that is learned and requires careful attention, as evidenced by Figure 5.6. If no annealing is used, the algorithm will not

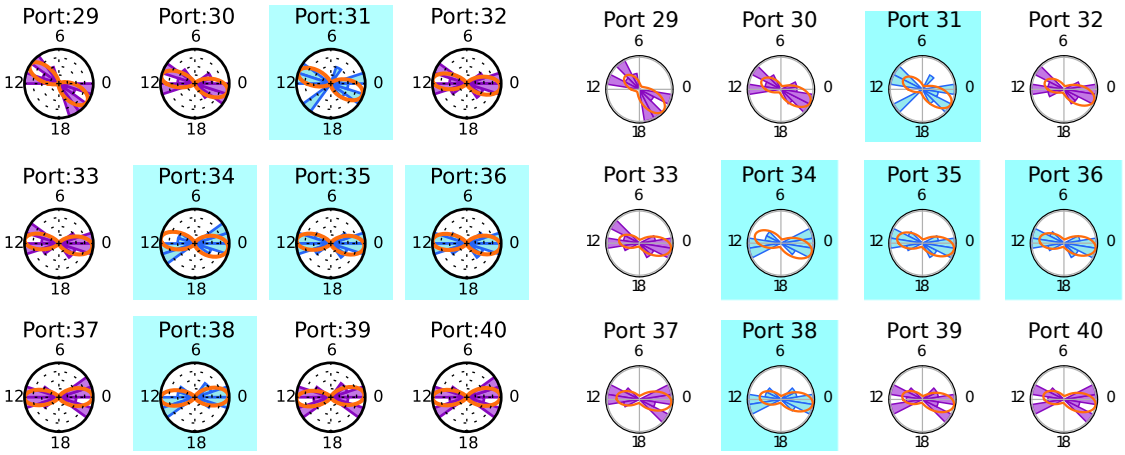


Fig. 5.5 Selected ports for regression on the Tides dataset using LEP-$m\mathcal{G}v\mathcal{M}$ (left) and MF-VI (right): predicted ports are highlighted in blue while training ports are shown in magenta locations denote training ports data points are denoted by crosses, the inferred density for each port is displayed as the orange line. The LEP-$m\mathcal{G}v\mathcal{M}$ inferred densities are typically less certain than the ones obtained by MF-VI, while the modes of both approximations are close.

converge in general, as displayed in the leftmost plot of Figure 5.6. An annealing schedule starting with a stricter variance value for the approximation of the constraints will induce a different learned regressed function, as can be seen comparing the middle plot of Figure 5.6 with the rightmost plot in Figure 5.6. This difference is a direct consequence of the constraints localising the fit on a region of space as portrayed in Figure 5.2 and the non-convex nature of the space of inferred functions. The computation of the moments of the $\mathcal{G}v\mathcal{M}$ distributions also required further attention to render them more numerically stable, including imposing a limitation on the concentration parameter values and developing a variant of the moment equations from Gatto (2008) shown in Appendix C.

5.3 Summary

In this chapter, we have introduced an Expectation Propagation algorithm for $m\mathcal{G}v\mathcal{M}$ based on the Expectation Propagation framework proposed by Opper and Winther (2005) and compared the proposed method with Mean-Field Variational Inference. Through experiments on real and synthetic data sets, we showed that the Lifted Expectation Propagation for the $m\mathcal{G}v\mathcal{M}$ algorithm provides a better approximation to the posterior than the Mean Field distribution at the expense of additional computation time.

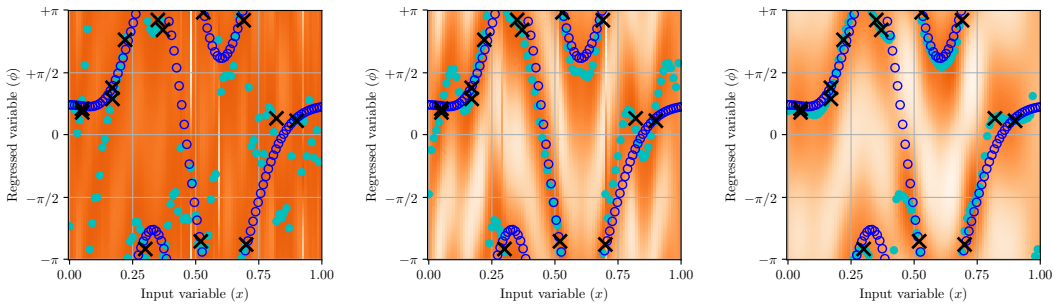


Fig. 5.6 Regression on the Wrapped Mexican Hat function using different annealing schemes representing data points as crosses, the true function by circles and mean of each approximating factor for the predictions by a solid dot: no annealing (left), a schedule with variances of $0.5, 0.1$ and 0.01 (middle), and a slower schedule from 10^3 to 10^{-4} decreasing a power at each iteration (right).

Chapter 6

Approaches based on Markov chain Monte Carlo

In this chapter, we introduce Markov chain Monte Carlo techniques which are suitable for data sets with up to N = 10^2 variables. In particular, we investigate sampling schemes using Gibbs sampling, Hamiltonian Monte Carlo (HMC) and a hybrid method that uses both algorithms. We show that in general sampling directly from the $m\mathcal{G}u\mathcal{M}$ model performs worse than using sampling schemes which employ the augmented representation of the $m\mathcal{G}u\mathcal{M}$. Each sampler is tested in both synthetic and real world data sets and the HMC samplers were tuned using Bayesian Optimisation to maximise sample indepech6/figndence.

6.1 Simulation and inference

Markov chain Monte Carlo (MCMC) are a class of methods based on the seminal work of Metropolis and Ulam (1949), which proposed Monte Carlo methods approximate integrals based on samples, with an insight from Hastings (1970). Hasting's key insight was to construct a Markov chain that converged to a target distribution we wish to draw samples from. Then, samples obtained from the Markov Chain transitions are used to form a Monte Carlo estimate of the target distribution.

There is often an expensive computational load associated with these methods, which also explains why these techniques were only popularised decades later by the influential works of Gelfand and Smith (1990) in statistics and Neal (1993) in machine learning. Furthermore, while this class of techniques produces exact samples from any distribution of interest including intractable posteriors, there is no diagnostic capable of guaranteeing convergence of the underlying Markov chain. Users must rely

on metrics based on analysing the auto-correlation generated by samplers such as the Rubin-Gelman diagnostic simulating multiple chains in parallel (Gelman and Rubin, 1992) or the Geweke test (Geweke, 1992).

The methods utilised in this chapter are already well-established in literature, therefore we provide only a brief overview of each of these methods along with the particulars for their use with the $m\mathcal{G}v\mathcal{M}$. Readers interested in the minutiae of these methods are directed to the excellent texts by Neal (1993), Gilks et al. (1995), Brooks et al. (2011), Gelman et al. (2013), MacKay (2003), Bishop (2006) and Murphy (2012).

6.1.1 Gibbs sampling

Gibbs sampling (Geman and Geman, 1984) is a widely used method that constructs a Markov Chain through the univariate conditionals of the distribution of interest. Namely, if one is interested in the distribution $p(\mathsf{x}_1, \ldots, x_\mathrm{N})$, a Gibbs sampler will draw samples from unidimensional conditionals of p according to Algorithm 3. In

Algorithm 3: Gibbs sampling

1 Define the number of samples to be S.
2 **for** $\mathrm{s} = 1 : \mathrm{S}$**,**
3 **for** $\mathrm{n} = 1 : \mathrm{N}$**,**
4 **Sample** $\mathsf{x}_{\mathrm{s,n}}$ **from** $p(\mathsf{x}_i | \mathsf{x}_{\mathrm{s}-1,1}, \ldots, \mathsf{x}_{\mathrm{s}-1,\mathrm{n}=1}, \mathsf{x}_{\mathrm{s}-1,\mathrm{n}+1}, \ldots, \mathsf{x}_{\mathrm{s}-1}, \mathrm{N})$
5 **Set** $\mathsf{x}_{\mathrm{s}-1,\mathrm{n}}$ **to** $\mathsf{x}_{\mathrm{s,n},}$

the limit as the number of samples approaches infinity, the sampler will converge to the distribution of interest. Hence, samples generated from the conditionals $\mathsf{x}_{\mathrm{n,s}}$ will coincide with samples of $p(\mathsf{x}_1, \ldots, x_\mathrm{N})$.

There are two major advantages of the Gibbs sampler: there is no time wasted generating samples that are not going to be accepted, and there are no free parameters that require tuning. However, these advantages come at a cost of very local transitions, a problem that is further aggravated when the joint distribution exhibits strong correlations (see, e.g. Bishop, 2006; MacKay, 2003).

Gibbs sampler's reliance on tractable conditionals of the distribution of interest and its lack of tuning parameters renders it an attractive baseline for comparing sampling methods for the $m\mathcal{G}v\mathcal{M}$. As discussed in Chapter 2, the one-dimensional conditionals of the $m\mathcal{G}v\mathcal{M}$

$$m\mathcal{G}v\mathcal{M}(\boldsymbol{\phi}; \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{K}) \propto \exp\left\{ \boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi} - \boldsymbol{\mu}) - \frac{1}{2} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix}^\top \mathbf{K}^{-1} \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} \right\} \tag{6.1}$$

are $\mathcal{GvM}$-distributed

$$\mathcal{GvM}(\phi_{\mathrm{n}}|\phi_{\neq \mathrm{n}}) \propto \exp\left\{\tilde{\kappa}_{1,\mathrm{n}}\cos(\phi_{\mathrm{n}}-\tilde{\mu}_{1,\mathrm{n}})+\tilde{\kappa}_{2,\mathrm{n}}\cos(2\phi_{\mathrm{n}}-2\tilde{\mu}_{2,\mathrm{n}})\right\} \tag{6.2}$$

with parameters defined through the relations

$$\begin{aligned}
\tilde{\kappa}_{\mathrm{n},1}\cos\tilde{\mu}_{\mathrm{n},1} &= \kappa_{1,\mathrm{n}}\cos(\mu_{1,\mathrm{n}}) \\
&\quad -\frac{1}{2}\sum_{\mathrm{j}\neq\mathrm{n}}\left[(\mathbf{K}^{-1})_{\mathrm{n,j}}\cos(\phi_{\mathrm{j}})+(\mathbf{K}^{-1})_{\mathrm{n,j+N}}\sin(\phi_{\mathrm{j}})\right] \\
\tilde{\kappa}_{\mathrm{n},1}\sin\tilde{\mu}_{\mathrm{n},1} &= \kappa_{1,\mathrm{n}}\sin(\mu_{1,\mathrm{n}}) \\
&\quad -\frac{1}{2}\sum_{\mathrm{j}\neq\mathrm{n}}\left[(\mathbf{K}^{-1})_{\mathrm{n+N,j}}\cos(\phi_{\mathrm{j}})+(\mathbf{K}^{-1})_{\mathrm{n+N,j+N}}\sin(\phi_{\mathrm{j}})\right] \\
\tilde{\kappa}_{\mathrm{n},2}\cos 2\tilde{\mu}_{\mathrm{n},2} &= -\frac{1}{4}\left[(\mathbf{K}^{-1})_{\mathrm{n,n}}+(\mathbf{K}^{-1})_{\mathrm{n+N,n+N}}\right] \\
\tilde{\kappa}_{\mathrm{n},2}\sin 2\tilde{\mu}_{\mathrm{n},2} &= -\frac{1}{2}(\mathbf{K}^{-1})_{\mathrm{n,n+N}}.
\end{aligned} \tag{6.3}$$

6.1.2 Hamiltonian Monte Carlo

The Hamiltonian Monte Carlo (HMC) method (Duane et al., 1987) relies on two central ideas: considering an intelligent augmentation of the distribution we wish to sample from as a Hamiltonian, and leveraging on the dynamical system associated with a Hamiltonian to drive the exploration of the sample space. More precisely, given a distribution

$$p(\mathbf{x}) \propto \exp\Big\{-E(\mathbf{x})\Big\}, \tag{6.4}$$

we wish to sample from, we define an augmented system

$$p(\mathbf{x},\boldsymbol{v}) \propto \exp\Big\{-H(\mathbf{x},\boldsymbol{v})\Big\} \tag{6.5}$$

where $H(\mathbf{x},\boldsymbol{v}) = E(\mathbf{x}) + K(\boldsymbol{v})$. Within the statistical physics formalism, the function E can be interpreted as potential energy for the states $\mathbf{x}$ while $K(\boldsymbol{v})$ can be regarded as a kinetic energy term with associated momenta $\boldsymbol{v}$.

The function $H(\mathbf{x},\boldsymbol{v})$ induces dynamical system that can be described in terms of the flow equations

$$\frac{\partial H(\mathbf{x},\boldsymbol{v})}{\partial \mathbf{x}} = -\frac{\partial \boldsymbol{v}}{\partial t} \quad \text{and} \quad \frac{\partial H(\mathbf{x},\boldsymbol{v})}{\partial \boldsymbol{v}} = \frac{\partial \mathbf{x}}{\partial t}, \tag{6.6}$$

describing the state and momentum transitions along a simulation time t. The dynamical system of Equation (6.6) represents a principled way to steer $\mathbf{x}$ away from the initial simulation point. Each iteration involves sampling a new initial state by drawing a sample for the momentum variable, integrating the Hamiltonian system and deciding whether to accept the new sample state based on the preservation of the volume of Hamiltonian system as presented in Algorithm 4

Algorithm 4: Hamiltonian Monte Carlo sampling

Define the number of samples to be S.
Define the number of integration steps L.
Define the size of integration steps ξ.
Define a set of accepted samples $\mathcal{A}$.
Set initial state $\mathbf{x}_0$.
while $|\mathcal{A}| \leq$ S,
    Sample $\boldsymbol{v}_0$ from $p(\boldsymbol{v}) \propto \exp -K(\boldsymbol{v})$
    Calculate $h_0 = H(\mathbf{x}_0, \boldsymbol{v}_0)$
    Get final states $\mathbf{x}_\mathrm{L}, \boldsymbol{v}_\mathrm{L}$ by numerically integrating Equation (6.6) with a symplectic integrator using L steps of size ξ.
    Calculate $h_\mathrm{L} = H(\mathbf{x}_\mathrm{L}, \boldsymbol{v}_\mathrm{L})$
    Sample $u = \mathrm{Uniform}(0, 1)$
    if $u \leq \exp(h_0 - h_\mathrm{L})$,
        Add $\mathbf{x}_\mathrm{L}$ to $\mathcal{A}$ and set $\mathbf{x}_0$ to $\mathbf{x}_\mathrm{L}$

Since the method's inception, the kinetic energy term $K(\mathbf{x}, \boldsymbol{v})$ is assumed to be a quadratic function of the momentum variables, i.e. $\frac{1}{2}\boldsymbol{v}^\top \mathbf{M}^{-1}\boldsymbol{v}$ for some matrix $\mathbf{M}^{-1}$. The reasons for maintaining this form are two-fold: it relates directly to the classical formulation of kinetic energy in physical systems, and it results in a multivariate Gaussian distribution which can be easily sampled from.

For the $m\mathcal{G}v\mathcal{M}$, the energy function from Equation (6.4) becomes

$$E(\boldsymbol{\phi}) = -\boldsymbol{\kappa}^\top \cos(\boldsymbol{\phi} - \boldsymbol{\mu}) + \frac{1}{2} \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix}^\top \mathbf{K}^{-1} \begin{bmatrix} \cos \boldsymbol{\phi} \\ \sin \boldsymbol{\phi} \end{bmatrix} \tag{6.7}$$

Different kinetic energy terms for circular distributions can be used, since the algorithm proposed by Duane et al. (1987) does not depend on the kinetic energy to be a Gaussian term. However, empirically we did not find any significant difference when other different, static[1] kinetic energy terms are used. For example, using

[1] In the investigation presented in this chapter we did not use any adaptive weight scheme such as a Riemannian Manifold Hamiltonian Monte Carlo (Girolami and Calderhead, 2011).

$K(\boldsymbol{v}) = \alpha \odot \cos(\boldsymbol{v})$ to sample from an associated von Mises associated distribution yielded similar results to the usual quadratic energy term.

Different from the Gibbs sampler, an HMC sampler possesses a large list of tuning parameters: the number of steps of symplectic integrator for the dynamical system of the Hamiltonian, the size of the steps taken by the symplectic integrator and the weights of the Kinetic energy distribution. Tuning this large number of parameters is often a dauting task. Moreover, there is no consensus on which metric the sampler has to be tuned for. In this study, we chose to prioritise exploration of the variable's space and adopted the Normalised Expected Squared Jump Distance criterion of by Wang et al. (2013), i.e.

$$\text{NESJD} = \frac{1}{\sqrt{\text{L}}}\mathbb{E}_{p(\boldsymbol{x})}\left[\mathsf{x}_{\text{s}-1} - \mathsf{x}_{\text{s}}\right]^2, \tag{6.8}$$

where p is the true distribution of the random of x and L is the size of the integration step in the HMC sampler.

Wang et al. (2013)'s criterion is a modification of the Expected Squared Jump Distance (ESJD) proposed by Pasarica and Gelman (2010),

$$\text{ESJD} = \mathbb{E}_{p(\boldsymbol{x})}\left[\mathsf{x}_{\text{s}-1} - \mathsf{x}_{\text{s}}\right]^2 \tag{6.9}$$

to account for the fact that increasing the number of integration steps in the HMC algorithm necessarily leads to an increase in the ESJD without necessarily improving the exploration of the space.

The sampler parameters are then tuned using a Bayesian Optimisation (BO) procedure as described by Wang et al. (2013). A discussion of Bayesian Optimisation methods are outside the scope of this chapter, however, the interested reader is referred to Brochu et al. (2010), Hoffman and Shahriari (2014) and Hernández-Lobato et al. (2014, 2015) for further details.

6.1.3 MCMC methods for the augmented representation

Recall from Chapter 3 that the augmented representation for the $m\mathcal{G}v\mathcal{M}$ distribution defined in Equation (6.1), has two possible representations. The first representation defines the augmentation states conditioned on the $m\mathcal{G}v\mathcal{M}$-distributed circular variables,

i.e.

$$\begin{aligned} p(\boldsymbol{\phi}) &= m\mathcal{G}v\mathcal{M}(\boldsymbol{\phi}; \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{K}) \\ p(\mathbf{f}|\boldsymbol{\phi}) &= \mathcal{N}\left(\mathbf{f}; \mathbf{A}(\sigma^2\mathbf{I} - \mathbf{K}^{-1}) \begin{bmatrix} \cos\boldsymbol{\phi} \\ \sin\boldsymbol{\phi} \end{bmatrix} + \boldsymbol{m}, \mathbf{A}(\sigma^2\mathbf{I} - \mathbf{K}^{-1})\mathbf{A}^\top\right). \end{aligned} \tag{6.10}$$

The second representation has fully factorised circular variables conditioned on the Gaussian augmentation states,

$$\begin{aligned} p(\mathbf{f}) &= \mathcal{N}\Big(\mathbf{f}; \boldsymbol{m}, \mathbf{A}(\sigma^2\mathbf{I} - \mathbf{K}^{-1})\mathbf{A}^\top\Big) \\ p(\boldsymbol{\phi}|\mathbf{f}) &= \prod_{\mathrm{n}=1}^{\mathrm{N}} \Big[v\mathcal{M}(\phi_\mathrm{n}; \alpha_\mathrm{n}(\mathbf{f}), \beta_\mathrm{n}(\mathbf{f}))\Big] \end{aligned} \tag{6.11}$$

where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given as

$$\boldsymbol{\alpha} \odot v(\boldsymbol{\beta}) = \boldsymbol{\kappa} \odot v(\boldsymbol{\mu}) + \mathbf{A}^{-1}(\mathbf{f} - \boldsymbol{m}) \tag{6.12}$$

and $\odot$ denotes the Hadamard product, i.e. point-wise vector multiplication.

Leveraging representations of Equation (6.10) and Equation (6.11), an efficient Gibbs sampler can be proposed. When sampling the conditionals of the circular states, the sampler resorts to the tractable representation of Equation (6.11). When the sampler needs to obtain samples from circular variables $\boldsymbol{\phi}$, the sampler utilises the representation of Equation (6.10).

One can also leverage the augmented representation to derive a hybrid sampler where HMC is performed on the augmentation states, and the energy and then samples for the circular states are obtained by drawing from von mises representations of Equation (6.10).

It can be hypothesised that the additional structure imparted by the additional variables from the augmentation transformation allows for more efficient sampling. This conjecture resides on the fact that changes in the added Euclidean states lead to longer jumps in the circular space than those that would result from the direct application of these methods. Moreover, similar results have been shown for discrete space models (see, e.g. Pakman and Paninski, 2013) to further support this conjecture.

6.1.4 Contrastive divergence learning

Recall that when using the $m\mathcal{G}v\mathcal{M}$ in a regression setting, the model prior is given by a special case of the $m\mathcal{G}v\mathcal{M}$ distribution, the Toroidal Normal, for which the normalising

constant is not known. As in this chapter we developed samplers for this distribution, we can use these MCMC procedures to implement a learning procedure similar to contrastive divergence.

Contrastive divergence was initially proposed by Ackley et al. (1985); Hinton (1989) for Ising models and more recently enjoyed a revival as a consequence of the popularity of Deep Restricted Boltzmann Machines (Hinton, 2002, 2010; Salakhutdinov and Hinton, 2009). We note that while in the circular context, the Directional-Unit Boltzmann machines proposed by Zemel et al. (1993) already employed a contrastive divergence learning procedure, their method used a mean-field approximation with von Mises distributions. The methods outlined next do not require this additional approximation step since we can compute all expectations required through the sampling procedures outlined in Section 6.1.

The contrastive divergence method hinges on two main concepts: reformulating the derivative of an intractable partition function in terms of the expectation of a gradient and the ability to optimise functions under noisy gradients to perform maximum-likelihood or maximum *a posteriori* learning. The former is addressed by writing either the log likelihood or log joint distribution as the objective function for a gradient-based stochastic optimisation method, most prominently, stochastic gradient ascent. The latter can be derived using log derivatives and the quotient rule for differentiation

$$\frac{\partial}{\partial\theta}\log\mathcal{Z}_{p(\phi|\theta)} = \frac{1}{\mathcal{Z}_{p(\phi|\theta)}}\frac{\partial}{\partial\theta}\int p^*(\phi|\theta)\mathrm{d}\phi \tag{6.13}$$

$$= \frac{1}{\mathcal{Z}_{p(\phi|\theta)}}\int \frac{p^*(\phi|\theta)}{p^*(\phi|\theta)}\frac{\partial}{\partial\theta}p^*(\phi|\theta)\mathrm{d}\phi \tag{6.14}$$

$$= \int p(\phi|\theta)\frac{\partial}{\partial\theta}\log p^*(\phi|\theta)\mathrm{d}\phi \tag{6.15}$$

which can be re-expressed as the expectation

$$\frac{\partial}{\partial\theta}\log\mathcal{Z}_{p(\phi|\theta)} = \left\langle\frac{\partial}{\partial\theta}\log p^*(\phi|\theta)\right\rangle_{p(\phi|\theta)} \tag{6.16}$$

6.2 Experiments

In this section, we conduct experiments with synthetic and real-world data sets to investigate the behaviour of different samplers and the use of contrastive divergence learning for the $m\mathcal{G}u\mathcal{M}$.

6.2.1 Analysis of sampler behaviour

The experiments conducted aimed to investigate the characteristics of both Gibbs and Hamiltonian Monte Carlo samplers applied directly to the $m\mathcal{G}u\mathcal{M}$ as well as their performance on the augmented model.

In order to assess the behaviour of the samplers in detail, four toy data sets were utilised to explore the qualitative behaviour of the sampler under controlled conditions, while another three real-world data sets were used to analyse the sampler's performance in practical settings. In general, the dimensionsionality of the problems tackled by the samplers presented in this section are smaller than the ones in previous chapters. This limitation is a consequence of the difficulty of effectively assessing the convergence of MCMC samplers in high-dimensional spaces (for a more in depth analysis of such convergence issues see, e.g. Rajaratnam and Sparks, 2015).

The toy data sets were created to reflect cases when the $m\mathcal{G}u\mathcal{M}$ can be unimodal or multimodal, and the effect of correlations. Therefore, for each example we sampled 40 points sampled from the empirical distributions on 2-dimensional grids for an uncorrelated, multimodal $m\mathcal{G}u\mathcal{M}$ (Toy-1), a correlated, multimodal $m\mathcal{G}u\mathcal{M}$ (Toy-2), an uncorrelated and unimodal $\mathcal{G}u\mathcal{M}$ (Toy-3), and an unimodal and correlated $m\mathcal{G}u\mathcal{M}$ (Toy-4). We also employed reduced versions of the datasets used in previous chapters in dimensions amenable to the samplers. These data sets included the training points the Wrapped Mexhat, Tides dataset and a subset of 10 angles in protein dataset (Sub-protein). The kernels used were the same as in the previous studies. Models were run with a fixed time budget of one hour and the first 5000 samples were discarted as a burn-in phase.

The performance of the different samplers was analysed using the log-evidence for a validation set comprising of held out data points. This metric was computed by forming the (simple) Monte Carlo estimate as

$$\log p(\boldsymbol{\psi}) = \log \int p(\boldsymbol{\psi}|\boldsymbol{\phi})p(\boldsymbol{\phi})\mathrm{d}\boldsymbol{\phi} \tag{6.17}$$

$$\approx \log \hat{p}(\boldsymbol{\psi}) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \log \sum_{\mathrm{s}=1}^{\mathrm{S}} p(\psi_{\mathrm{n}}|\boldsymbol{\phi}_{\mathrm{s}}) \quad \boldsymbol{\phi}_{\mathrm{s}} \sim p(\boldsymbol{\phi}_{\mathrm{s}}). \tag{6.18}$$

This performance score was chosen as it clearly indicates whether samples generated by a given method represent well the validation set. The results for this measure are tabulated in Table 6.1.

Table 6.1 indicates that the augmented representation model tend to represent the validation set better than their non-augmented counterparts. This evidence is

Table 6.1 Comparison of log-evidence evaluated at held out data set for different sampling schemes for the same running time. Columns referring to an algorithm using the augmented representation of the $m\mathcal{G}u\mathcal{M}$ are denoted by the prefix AR.

	AR-HMC	HMC	AR-Gibbs	Gibbs
Toy-1	$\mathbf{+1.7 \cdot 10^{-1}}$	$+1.31 \cdot 10^{-1}$	$+1.54 \cdot 10^{-1}$	$+1.29 \cdot 10^{-1}$
Toy-2	$\mathbf{+1.3 \cdot 10^{-1}}$	$+1.04 \cdot 10^{-1}$	$+1.07 \cdot 10^{-1}$	$+1.05 \cdot 10^{-1}$
Toy-3	$\mathbf{+4.81 \cdot 10^{-1}}$	$+4.76 \cdot 10^{-1}$	$+4.27 \cdot 10^{-1}$	$+4.55 \cdot 10^{-1}$
Toy-4	$\mathbf{+6.4 \cdot 10^{-1}}$	$+3.82 \cdot 10^{-1}$	$+4.23 \cdot 10^{-1}$	$+3.75 \cdot 10^{-1}$
Wrapped hat	$+1.27 \cdot 10^{-1}$	$+1.16 \cdot 10^{-1}$	$\mathbf{+1.46 \cdot 10^{-1}}$	$+1.10 \cdot 10^{-1}$
Tides	$+2.89 \cdot 10^{+0}$	$+2.81 \cdot 10^{+0}$	$\mathbf{+3.79 \cdot 10^{+0}}$	$+3.67 \cdot 10^{+0}$
Sub-protein	$\mathbf{+2.88 \cdot 10^{+0}}$	$+9.40 \cdot 10^{-1}$	$-5.50 \cdot 10^{-1}$	$-2.00 \cdot 10^{-1}$

supported by the plots shown from Figure 6.1 to Figure 6.8, which examine the sample histograms against the true distribution as well as each sampler's trajectory and trace for a limited number of samples.

In the multimodal cases Toy-1 and Toy-2, both the AR-HMC sampler and the AR-Gibbs samplers represent well all four modes of the true distribution. The optimally tuned HMC sampler can also represent three of the four modes well, but also places a substantial amount of probability mass in lower-probability regions between modes. The Gibbs sampler trace indicates that generally the sampler does not switch modes, as can be seen in the remaining samplers. This indicates the poor representation of the underlying distribution as it places substantial mass on a low-probability zone despite providing the sampler with a long burn-in phase.

In the unimodal cases of Toy-3 and Toy-4, the Gibbs sampler can correctly explore and sample from the underlying distribution and performs well in the unimodal uncorrelated case as expected. However, the samplers using the model with the augmentation transformation perform better in the unimodal correlated case as they tend to spend more time sampling the high-probability regions.

It is worth noting that in the low-dimensional toy data sets, AR-HMC, HMC and AR-Gibbs do not exhibit significant discrepancies regarding their end results. This similarity is explained by the fact that the computational time allowed to generate the samples are sufficient for mixing even in the AR-Gibbs case. It is also the case for Wrapped hat and Tides data set. For larger-dimensional problems such as the Sub-protein data set, a more sophisticated method is required to sample from the high-probability regions. Short-sighted methods such as Gibbs sampling cannot provide an efficient exploration of the sample space and perform substantially worse.

6.2.2 Contrastive divergence learning for the $m\mathcal{G}u\mathcal{M}$

To assess the effectiveness of contrastive divergence learning, we analysed the behaviour of the samplers discussed the previous section and compared it against the true parameter values synthetic data sets. These synthetic data sets were generated by obtaining 200 samples from a known 5-dimensional $m\mathcal{G}u\mathcal{M}$ using the HMC sampler. The stochastic gradient descent parameters utilised in this experiment were a maximum of 500 solver iterations with learning rate of 1.0×10^{-4} and momentum of 5.0×10^{-5}. Convergence was assessed when the norm between successive iterations decreased below 10^{-4}.

The parameters learned by contrastive divergence for data sets whose ground truth parameters are known are outlined in Table 6.2.

Table 6.2 Comparison of between true parameter values and learned values through contrastive divergence with 10 samples (CD-10) and 100 samples (CD-100).

Kernel	**Parameter**	**True**	**CD-10**		**CD-100**	
		Value	**Time**	**Value**	**Time**	**Value**
SE	κ	5.00	39.8	5.59	384.0	5.84
	σ	200.0		179.1		177.4
	ℓ	1.41		1.81		1.80
Matérn 5/2	κ	3.0	34.9	4.73	360.6	4.75
	ℓ	0.75		17.9		20.8

The results for the squared exponential kernel in Table 6.2 show that the contrastive divergence procedure has learned a distribution similar to the original distribution from which the samples were obtained. For the Matérn kernel, the learned parameters differ from those of the true underlying distribution, with the learned length-scale being considerably inflated. The overall kernel influence is downplayed in favour of the concentration parameter, which is learned to be greater than ground truth. This last effect can be attributed to the $m\mathcal{G}u\mathcal{M}$'s over-parameterisation. This feature implies that there exist situations such as the Matérn kernel experiment where it is not simple to disambiguate the effects of the kernel and the ones from the concentration parameter under a limited number of samples. Furthermore, the energy function optimised is intrinsically non-convex, and special attention to initialisation has to be considered as in the Gaussian Process case (e.g. see Rasmussen and Williams, 2006, Section 5.4 and Figure 5.5)

In both experiments, there was no significant difference regarding the values obtained after 500 iterations for using 10 or 100 samples. However, the variance in the learning procedure is much more pronounced for fewer samples as shown in Figure 6.9. In our experiments there was no major difference between results with Gibbs and AR-Gibbs samplers. The AR-HMC and HMC samplers were not considered in this analysis since they require special tuning.

6.3 Conclusions

In this chapter, we introduced two Markov chain Monte Carlo methods for both the multivariate Generalised von Mises and its augmented model representation. We compared each sampler scheme against each other on synthetic and real-world data sets, showing that the samplers using the augmented representation perform better over standard sampling techniques in correlated and multi-modal cases, with their performance advantage growing with the dimensionality of the data. We also showed that there is little difference in the sampling schemes for unimodal uncorrelated cases when using the chosen HMC tuning metric.

Another contribution of this chapter is to present a contrastive divergence learning procedure for the $m\mathcal{G}v\mathcal{M}$, which can properly learn the parameters of the $m\mathcal{G}v\mathcal{M}$ and alleviates the issue of double intractability when performing regression with the $m\mathcal{G}v\mathcal{M}$ model in problems of reduced dimensionality.

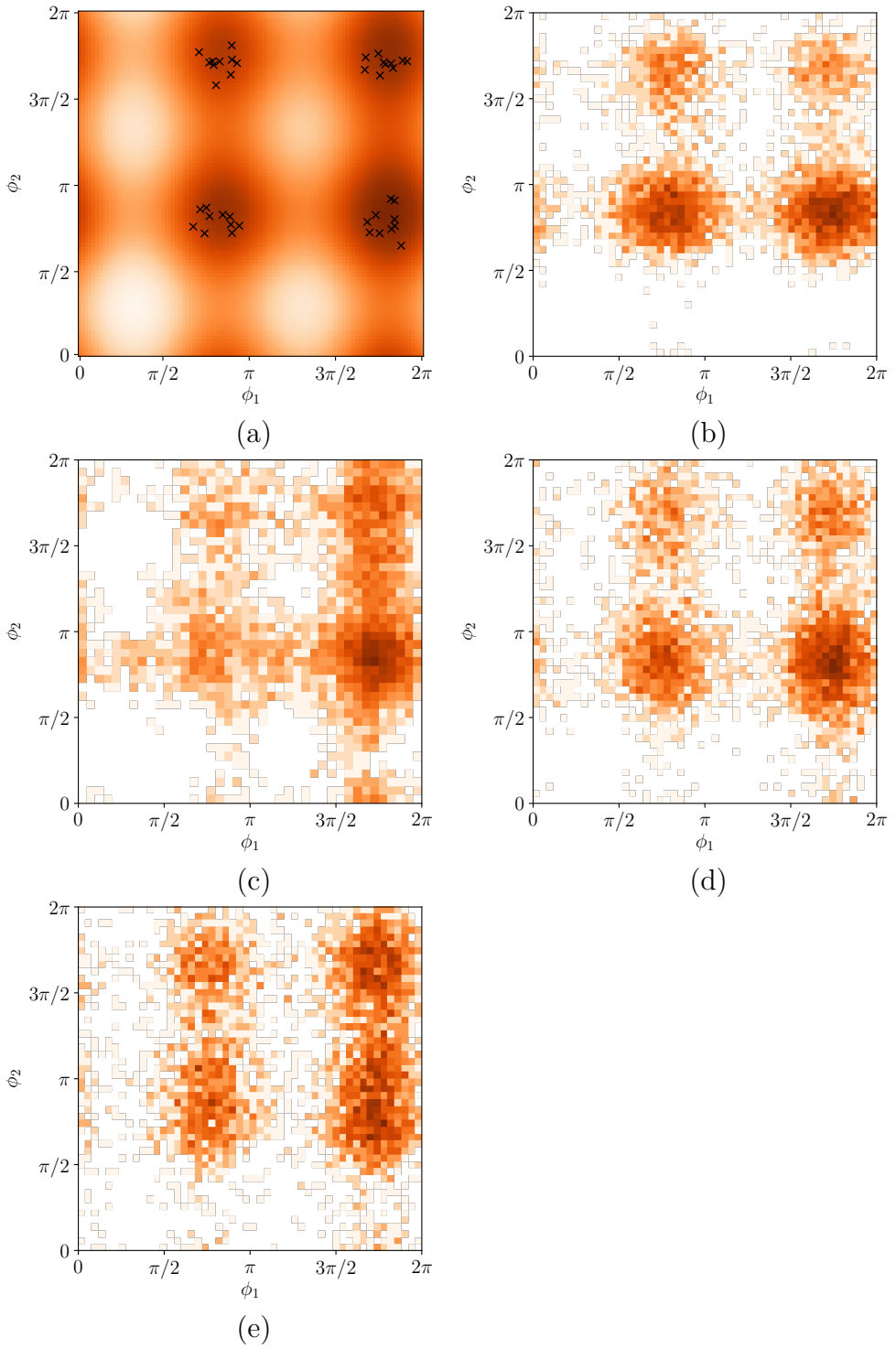


Fig. 6.1 Comparison of bivariate histogram of 5000 samples for an uncorrelated and multimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-1: true distribution with validation samples indicated (a), AR-HMC (b), HMC (c) and AR-Gibbs (d) and Gibbs (e).

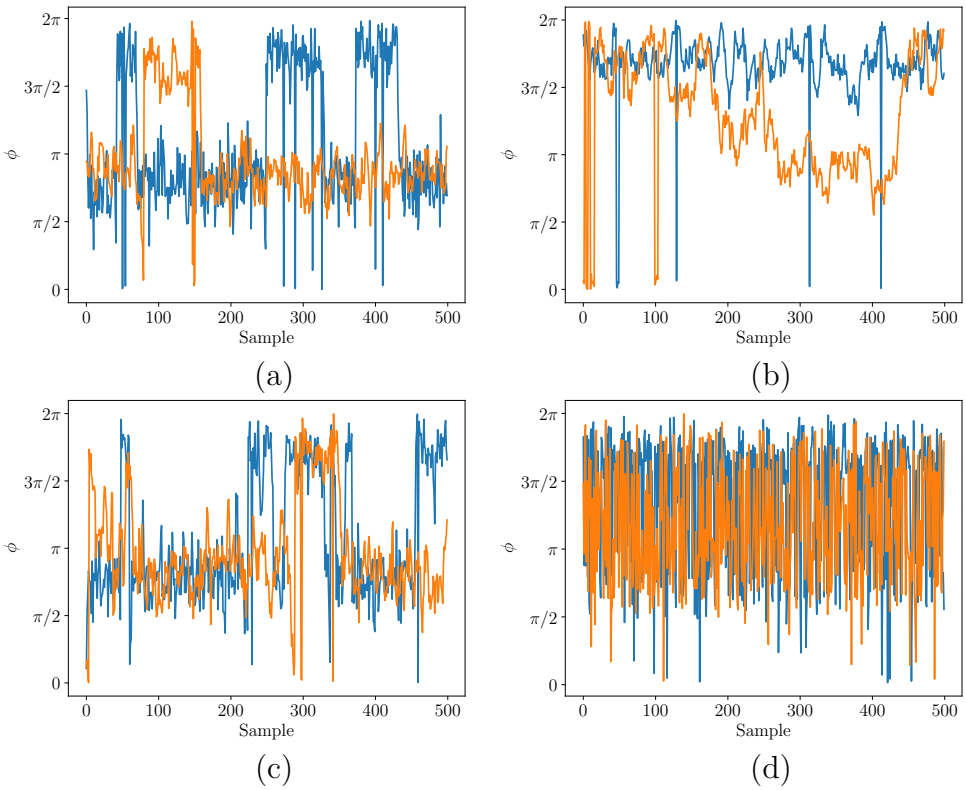


Fig. 6.2 Sampler trace for the 500 first samples for an uncorrelated multimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-1 using AR-HMC (a), HMC (b), AR-Gibbs (c) and Gibbs (d).

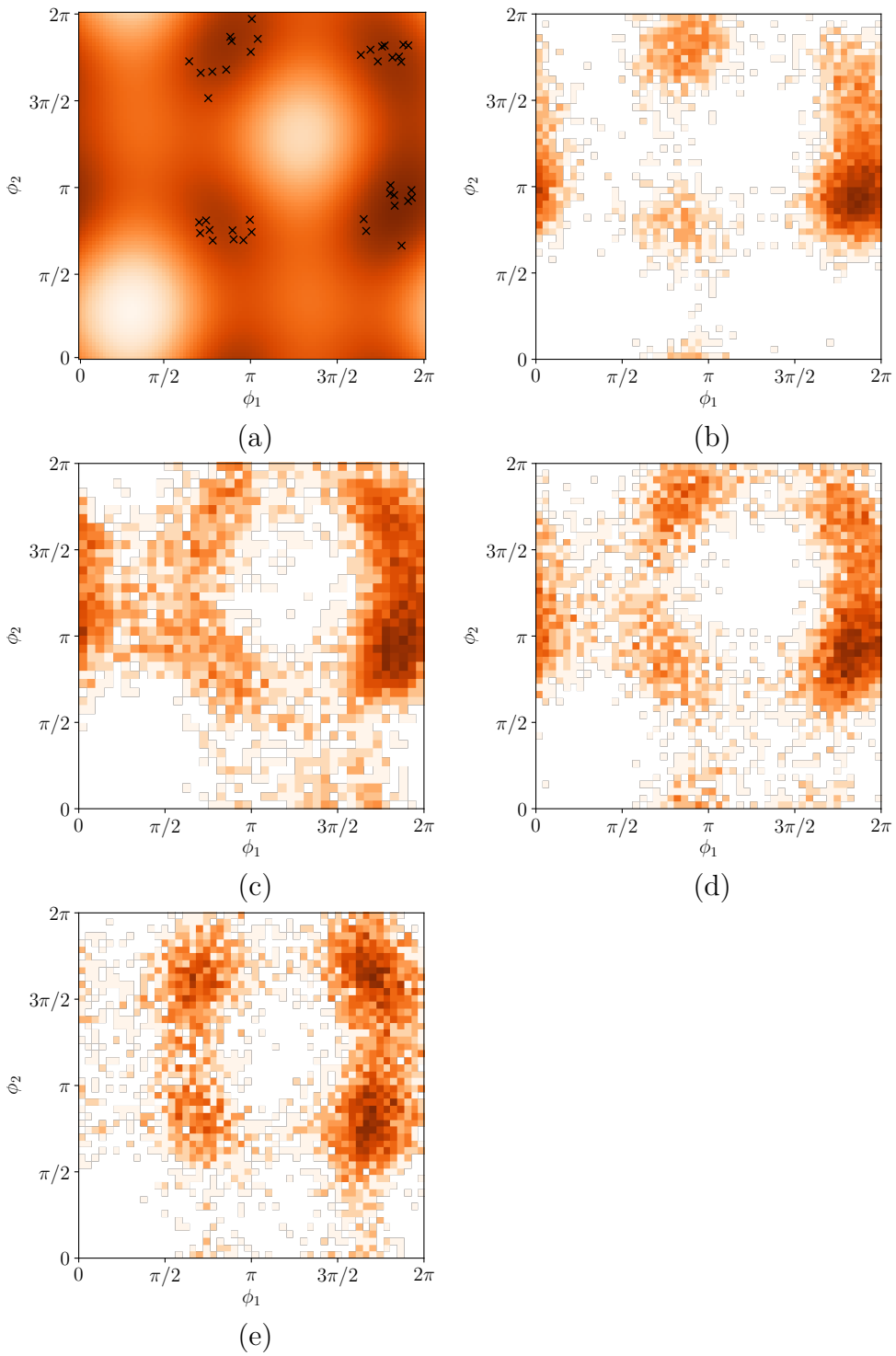


Fig. 6.3 Comparison of bivariate histogram of 5000 samples for a correlated multimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-2: true distribution with validation samples indicated (a), AR-HMC (b), HMC (c) and AR-Gibbs (d) and Gibbs (e).

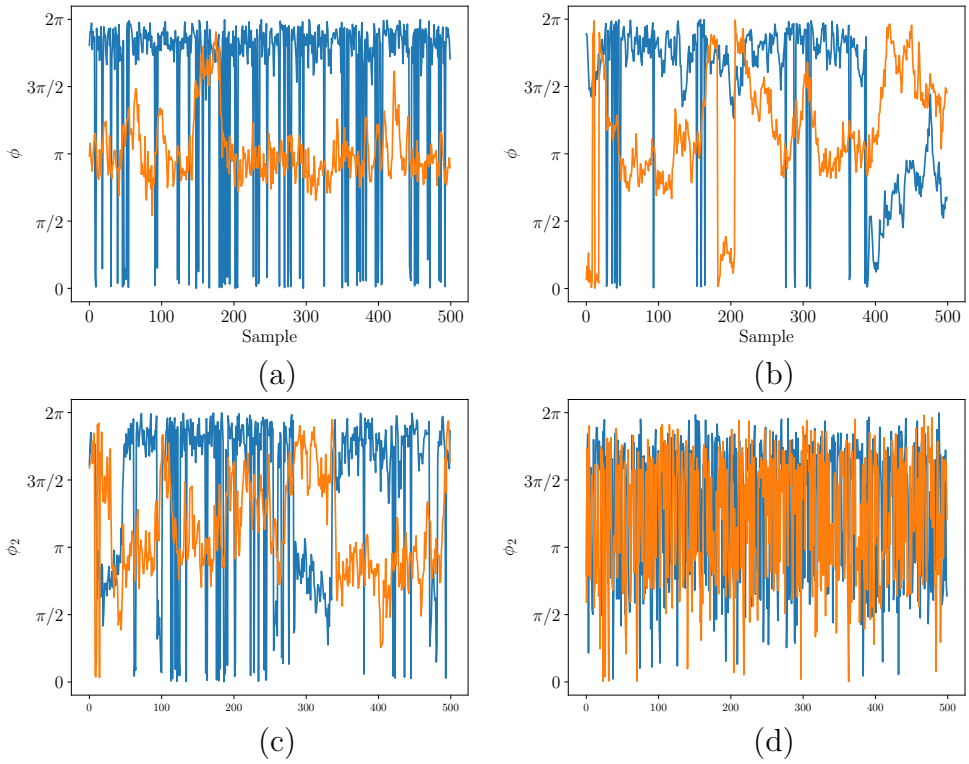


Fig. 6.4 Sampler trace for the 500 first samples for a correlated multimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-2 using AR-HMC (a), HMC (b), AR-Gibbs (c) and Gibbs (d).

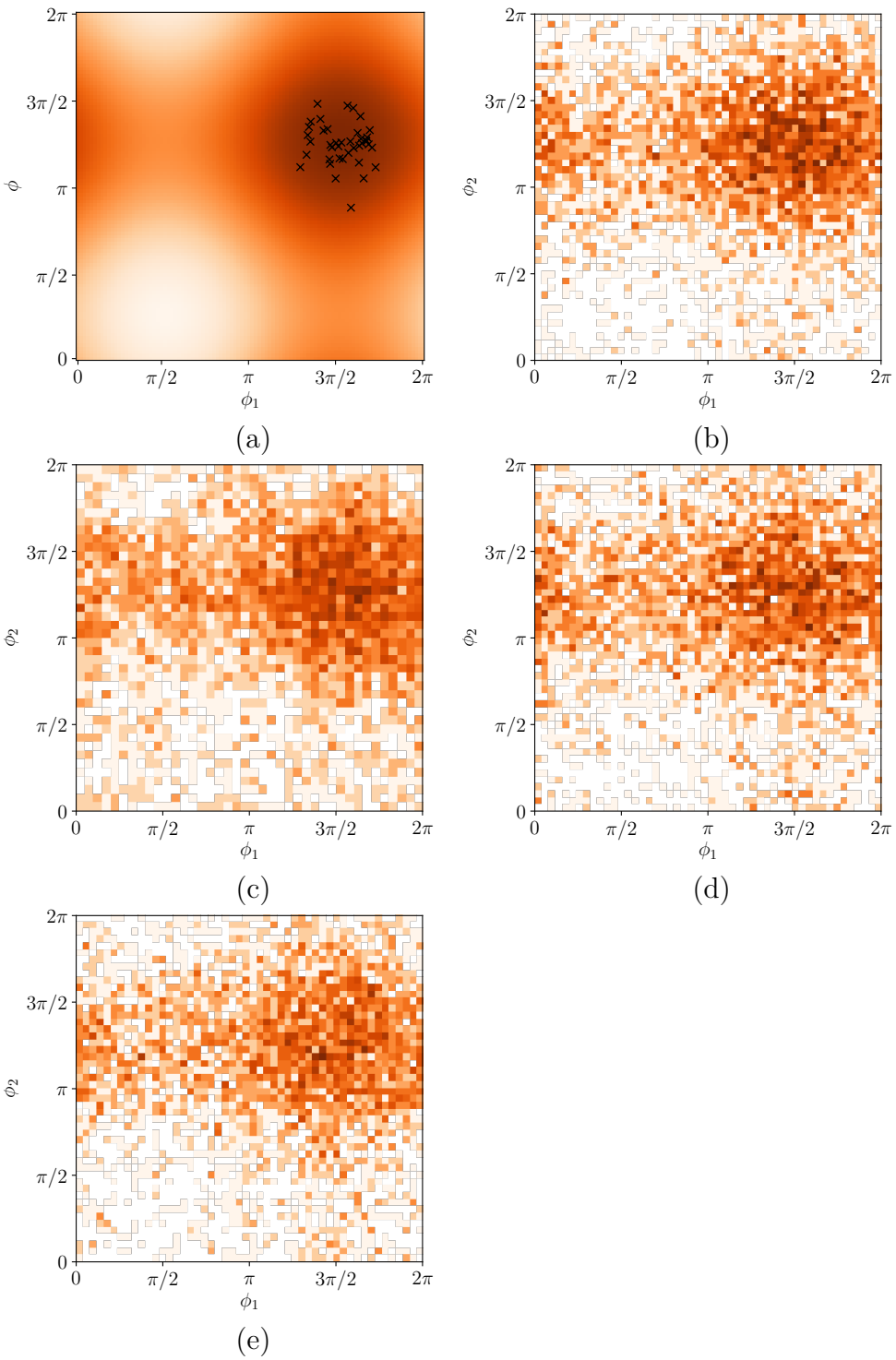


Fig. 6.5 Comparison of bivariate histogram of 5000 samples for an uncorrelated unimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-3: true distribution with validation samples indicated (a), AR-HMC (b), HMC (c) and AR-Gibbs (d) and Gibbs (e).

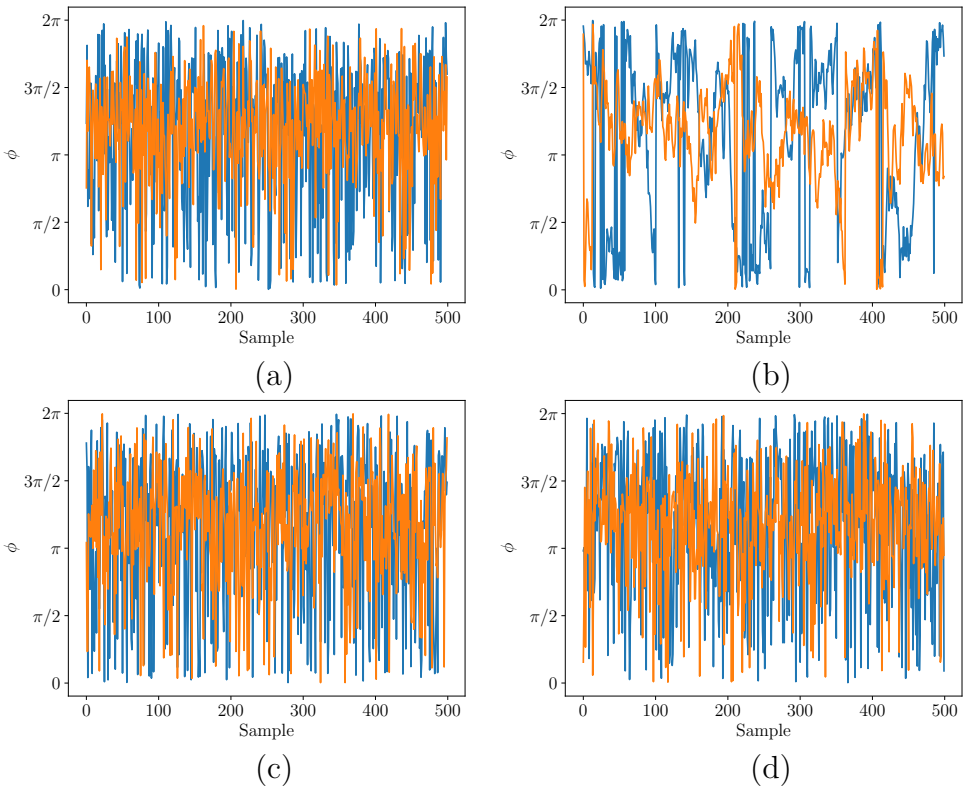


Fig. 6.6 Sampler trace for the 500 first samples for an uncorrelated unimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-3 using AR-HMC (a), HMC (b), AR-Gibbs (c) and Gibbs (d).

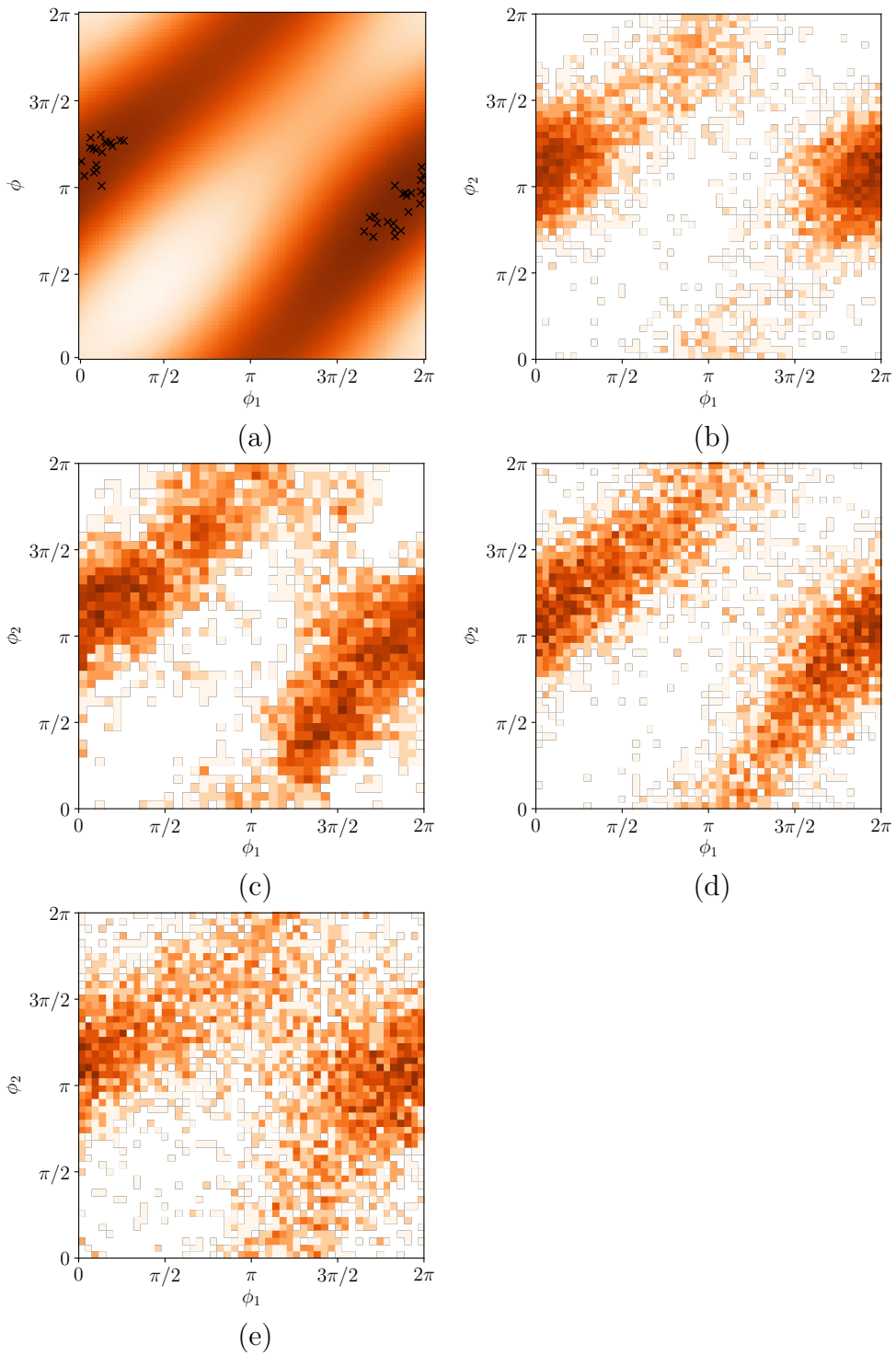


Fig. 6.7 Comparison of bivariate histogram of 5000 samples for a correlated unimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-4: true distribution with validation samples indicated (a), AR-HMC (b), HMC (c) and AR-Gibbs (d) and Gibbs (e).

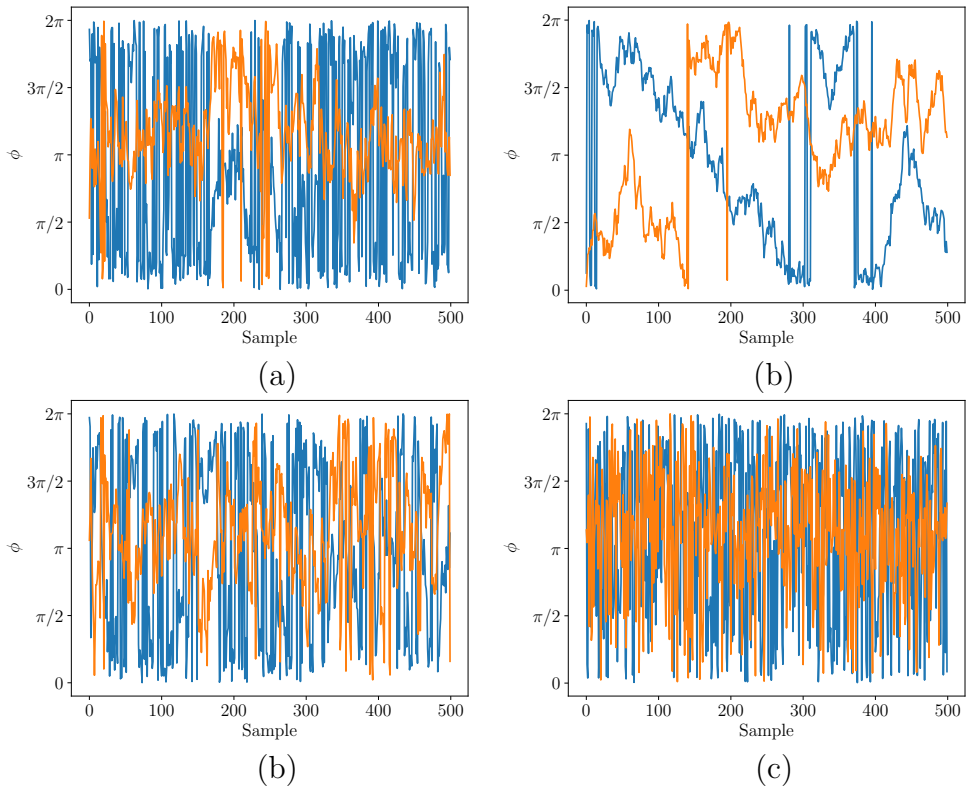


Fig. 6.8 Sampler trace for the 500 first samples for a correlated unimodal bivariate $m\mathcal{G}v\mathcal{M}$ distribution of the data set Toy-4 using AR-HMC (a), HMC (b), AR-Gibbs (c) and Gibbs (d).

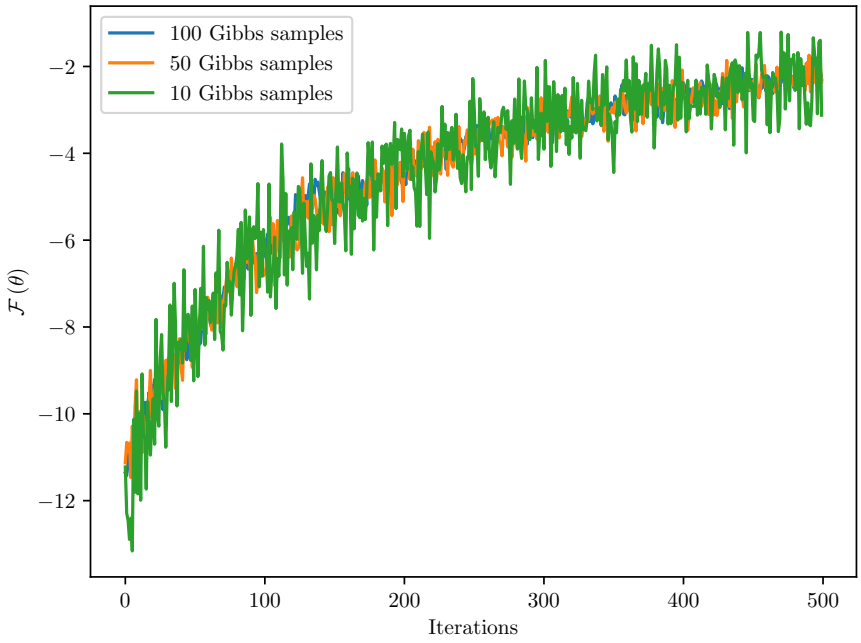


Fig. 6.9 Evolution of free energy when learning the parameters of a $m\mathcal{G}v\mathcal{M}$ with contrastive divergence using different number of Gibbs samples.

Part IV

Conclusions

Chapter 7

Conclusions and further directions

In this chapter, we conclude the thesis by providing a discussing of overarching contributions presented in this work in Section 7.1. Following this discussion, Section 7.2 presents further research avenues stemming from the work presented in this thesis.

7.1 Conclusions arising from the thesis contributions

The thread that unites this thesis is its goal of bridging the gap between the Circular Statistics and the Probabilistic Machine Learning communities. To this intent, several models and inference methods that fuse insights from both areas were developed.

In the modelling front, one of the central developments lies in the introduction of Multivariate Generalised von Mises distribution and the analysis of its properties. This distribution provides a straightforward way to match the need to correctly represent the covariance structure for hyper-toroidal spaces (a key requirement of the Circular Statistics community) with the kernel machinery developed for Gaussian Processes (a workhorse of the Probabilistic Machine Learning community). The $m\mathcal{G}v\mathcal{M}$ allowed for the creation of a probabilistic Principal Component Analysis analogue for circular variables, a previously unsolved problem in circular statistics, and contributing towards unifying the treatment of regression problems in circular statistics through the use of kernels, since kernels admit a myriad of different inputs ranging from Euclidean variables to character sequences which had to be posed under different frameworks before.

While the thesis focuses on $m\mathcal{G}v\mathcal{M}$, it is important to remark that the algorithms and modelling framework developed also extend to the Matrix Fisher-Bingham distribution.

Minor adjustments have to be performed for this more general distribution to be used. These adjustments are mentioned throughout the thesis and can be often seen as taking inner products of D-dimensional unit vectors instead of 2-dimensional unit vectors. Therefore, the thesis can be seen as trying to integrate Circular Statistics in its broadest sense to Machine Learning. That is, to consider both spaces comprised of angles and unit vectors.

The thesis also evidences that it is not trivial to perform inference with circular distributions, in particular the $m\mathcal{G}v\mathcal{M}$. Inference has issues both of numerical and theoretical nature. For example, the numeric evaluation of the moments of Generalised von Mises are non-monotonically decreasing series of Bessel functions, which require substantial effort and to evaluate and implement in robust manner. Such numerical issues have plagued every method attempted. This is often not a problem in typical Machine Learning models.

On the theoretical front, we can point to the transductiveness of the model, that is, the fact that the input locations have to be known before regression is performed. This behaviour is a direct consequence of the way the $m\mathcal{G}v\mathcal{M}$ is constructed, as the the $m\mathcal{G}v\mathcal{M}$ can be seen as imposing a finite number of constraints on a continuous functions. That is, the function represented by the kernels in the $m\mathcal{G}v\mathcal{M}$ need to be on the unit circle only at the input locations. This construction then inherently limits how the $m\mathcal{G}v\mathcal{M}$ generalises, different from more flexible models such as Gaussian Processes.

The power and limitations of the $m\mathcal{G}v\mathcal{M}$ showcase how the use of Circular Statistics models in Probabilistic Machine Learning can yield promising results, as well as the limitations of a purely circular approach. The augmented representations introduced in Chapter 3 presents an alternative approach to modelling circular variables that resonates more with Probabilistic Machine Learning generative models. This model requires additional latent variables to facilitate inference, but turned an intractable model into a tractable one.

While in the modelling front a central goal was to bring Circular Statistics models into evidence for the Probabilistic Machine Learning community, in the algorithms section the aim was to introduce approximate inference methods to Circular Statistics. The Circular Statistics community has greatly relied in exact inference, whereas the Probabilistic Machine Learning has greatly advanced approximate inference methods. With this in mind, we provided simple Variational Inference (Chapter 4), Expectation Propagation (Chapter 5) and Markov chain Monte Carlo (Chapter 6) techniques that are widely diffused in the Probabilistic Machine Learning community.

This explains why we have emphasised in Chapter 4 a standard mean-field approximations for the $m\mathcal{G}v\mathcal{M}$, leaving more complex approximations to Appendix B. The mean-field approximation is a rather simple approximation when compared to the state-of-the-art in contemporary Machine Learning research, but provides an important foundation for a novice to the field. Furthermore, it provides a good foundation to later compare more sophisticated approximations such as the one outlined in Appendix B.

In Chapter 5, we have focused on obtaining an Expectation Propagation algorithm that was fully convergent, as the other existing methods in literature failed to converge even for small problems and using damping. While there is room for exploring other updates and factor approximations, achieving an algorithm that consistently converges is an important initial step. As in the variational free energy case, this sets a standard over which future algorithms can be compared to.

Finally, Chapter 6 concerned itself with algorithms that make use of augmented states in Markov chain Monte Carlo. It is fair to say that despite the use of MCMC techniques in Circular Statistics, in general only the original form of the distribution is used instead of an augmented form—be it in Hamiltonian Monte Carlo or with model augmentation of the kind presented in Chapter 3. Furthermore, we also showed how contrastive divergence can be used to learning Multivariate Generalised von Mises, a technique that also has had little permeation in the Circular Statistics community.

In conclusion, this thesis achieves its goal of establishes a number of benchmarks and points for discussion over the need for integrating Circular Statistics and Probabilistic Machine Learning. The thesis provides a convergent point from which practitioners from both Circular Statistics and Probabilistic Machine Learning can build upon as well as outlining multiple further research points.

7.2 Further work

A recurrent problem throughout when applying the $m\mathcal{G}v\mathcal{M}$ and related distributions to large datasets resides in the numerical stability of algorithms involving modified Bessel functions of first kind. Despite the existence of bounds and asymptotic approximations for such functions in literature, we did not find such methods to be robust in practice. This resonates with the findings of Sra (2012), who showed that in some cases naïve numerical implementations for such functions perform better in large dataset contexts. Hence, we believe that further numerical research should be conducted to obtain stable and accurate methods for calculating modified Bessel functions and their ratios.

In this thesis we focused on fully factored approximations of the $m\mathcal{G}v\mathcal{M}$. A interesting research direction is to use structured approximations that do not require factors to be independent. For example, one could use an augmented representation defined in Appendix B as the approximating distribution to a true $m\mathcal{G}v\mathcal{M}$ in a variational free energy setting. The details surrounding the use of this approximation are sketched Appendix B. While the application of such approximations may seem trivial, preliminary experiments revealed that further reserach is needed before one can use this approximation. These experiments revealed that the gradients of the variational free energy suffer from high variance as a consequence of the Monte Carlo integration of the moments of the Bessel functions. While there are significant improvements in variance reduction for continuous and and categorical variables (see e.g., Greensmith et al., 2005; Jang et al., 2016; Maddison et al., 2016), variance reduction for gradients in circular models remains an open problem.

Another area that could benefit from the results derived in this thesis is probabilistic deconvolution. For example, the distributions and methods derived in this thesis can lead to probabilistic versions of the Discrete-Time Fourier transform and the Z-transform (Ragazzini and Zadeh, 1952). The inverse of such transformations cannot be obtained analytically for general functions, and require the deconvolution of a signal into an angular and an amplitude component. As the $m\mathcal{G}v\mathcal{M}$ allows imposing a functional structure on the phase component, we conjecture that the $m\mathcal{G}v\mathcal{M}$ can be play a significant role in the derivation of probabilistic methods for deconvolving time series and frequency analysis of control systems. Moreover, since modern control methods are strongly reliant on phase and frequency analysis (see, e.g., the celebrated results from McFarlane and Glover, 1992, which found multiple practical applications) we recognise that the use of machine learning in control theory should be able to benefit greatly from the contributions from this thesis.

Finally, a straightforward direction towards extending the work in this thesis is to incorporate further contributions from the Gaussian Process literature to the $m\mathcal{G}v\mathcal{M}$. These include scaling the $m\mathcal{G}v\mathcal{M}$ and its augmented representations to handle very large data sets through the use of sparse representation, the creation of latent variable models for circular spaces akin to the GP-LVM (Lawrence, 2004), the use of $m\mathcal{G}v\mathcal{M}$-distributed time series as the GPSS (Frigola et al., 2014; Turner, 2011), extending Gaussian states in deep Gaussian Processes (Damianou and Lawrence, 2013) and Gaussian Process Auto-Encoders (Eleftheriadis et al., 2017) in a similar vein to the work of Davidson et al. (2018).

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. Wiley, New York, 1972. ISBN 0486612724.

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines*. *Cognitive Science*, 9(1):147–169, 1985. ISSN 1551-6709. doi: 10.1207/s15516709cog0901_7. URL http://dx.doi.org/10.1207/s15516709cog0901_7.

D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581 – 598, 1981. ISSN 0047-259X. URL http://dx.doi.org/10.1016/0047-259X(81)90099-3.

D. J. Aldous. Exchangeability and related topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag, 1985. Lecture Notes in Mathematics 1117.

Z. Bai and G. H. Golub. Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices. *Annals of Numerical Mathematics*, 4:29–38, 1997.

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

E. Batschelet. *Circular Statistics in Biology*. Mathematics in biology. Academic Press, 1981. ISBN 9780120810505.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 2008. ISBN 9780470316870. doi: 10.1002/9780470316870.

C. Bingham. An Antipodally Symmetric Distribution on the Sphere. *The Annals of Statistics*, 2(6):1201–1225, 1974.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 9780387310732.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, Mar. 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT'98, pages 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0. doi: 10.1145/279943.279962.

E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. eprint arXiv:1012.2599, arXiv.org, December 2010.

S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. ISBN 978-1-4200-7941-8. URL https://doi.org/10.1201/b10905.

T. D. Bui and R. E. Turner. Tree-structured gaussian process approximations. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2213–2221. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5459-tree-structured-gaussian-process-approximations.pdf.

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

Y. Cho and L. K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf.

D. R. Cox. Discussion of Professor Mardia's paper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3):380–381, 1975.

B. Cseke and T. Heskes. Bounds on the bethe free energy for gaussian networks. In D. McAllester and P. Myllymaki, editors, *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008)*, pages Helsinki, Finland, July 9 – 12 2008342–350. Helsinki, Finland, July 2008.

A. Damianou and N. Lawrence. Deep gaussian processes. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL http://proceedings.mlr.press/v31/damianou13a.html.

P. J. Daniell. Integrals in an Infinite Number of Dimensions. *Annals of Mathematics*, 20:281–288, 1919.

T. R. Davidson, L. Falorsi, N. D. Cao, T. Kipf, and J. M. Tomczak. Hyperspherical variational auto-encoders, 2018.

B. de Finetti. Funzione Caratteristica Di un Fenomeno Aleatorio. In *Atti della R. Academia Nazionale del Lincei*, 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale, pages 251–299. Academia Nazionale del Linceo, 1931.

V. R. de Sa. Learning classification with unlabeled data. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 112–119. Morgan-Kaufmann, 1994. URL http://papers.nips.cc/paper/831-learning-classification-with-unlabeled-data.pdf.

M. Di Marzio, A. Panzera, and C. C. Taylor. Local polynomial regression for circular predictors. *Statistics & Probability Letters*, 79(19):2066–2075, 2009. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2009.06.014. URL http://www.sciencedirect.com/science/article/pii/S0167715209002417.

M. Di Marzio, A. Panzera, and C. C. Taylor. Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, 141(6):2156–2173, 2011. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2011.01.002. URL http://www.sciencedirect.com/science/article/pii/S037837581100019X.

M. di Marzio, A. Panzera, and C. C. Taylor. Non-parametric Regression for Circular Responses. *Scandinavian Journal of Statistics*, 40(2):238–255, 2012. doi: 10.1111/j.1467-9469.2012.00809.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2012.00809.x.

M. Di Marzio, S. Fensore, A. Panzera, and C. C. Taylor. Nonparametric estimating equations for circular probability density functions and their derivatives. *Electron. J. Statist.*, 11(2):4323–4346, 2017. doi: 10.1214/17-EJS1318. URL https://doi.org/10.1214/17-EJS1318.

T. D. Downs. Orientation statistics. *Biometrika*, 59(3):665–676, 1972. doi: 10.1093/biomet/59.3.665. URL http://dx.doi.org/10.1093/biomet/59.3.665.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 0370-2693. doi: 10.1016/0370-2693(87)91197-X. URL http://dx.doi.org/10.1016/0370-2693(87)91197-X.

D. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, Computational and Biological Learning Laboratory, University of Cambridge, 2014.

D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302,4922v4*, 28, 2013. URL http://arxiv.org/abs/1302.4922.

D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen. Additive Gaussian Processes. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 226–234. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4221-additive-gaussian-processes.pdf.

S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic. Variational gaussian process auto-encoder for ordinal prediction of facial action units. In S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, editors, *Computer Vision – ACCV 2016*, pages 154–170, Cham, 2017. Springer International Publishing. ISBN 978-3-319-54184-6.

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

R. Fisher. Dispersion on a Sphere. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 217(1130):295–305, 1953. ISSN 0080-4630. doi: 10.1098/rspa.1953.0064.

R. Frigola, Y. Chen, and C. E. Rasmussen. Variational gaussian process state-space models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3680–3688. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5375-variational-gaussian-process-state-space-models.pdf.

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

R. Gatto. Some computational aspects of the generalized von Mises distribution. *Statistics and Computing*, 18(3):321–331, sep 2008. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-008-9060-4.

R. Gatto and S. R. Jammalamadaka. The generalized von Mises distribution. *Statistical Methodology*, 4(3):341–353, jul 2007. ISSN 15723127. doi: 10.1016/j.stamet.2006.11.003.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2289776.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, 7(4):457–472, 11 1992. doi: 10.1214/ss/1177011136. URL http://dx.doi.org/10.1214/ss/1177011136.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6(6):721–741, Nov. 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596. URL http://dx.doi.org/10.1109/TPAMI.1984.4767596.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, and J. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–193. Oxford University Press, Oxford, 1992.

Z. Ghahramani and T. L. Griffiths. Infinite latent feature models and the indian buffet process. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, 2006. URL http://papers.nips.cc/paper/2882-infinite-latent-feature-models-and-the-indian-buffet-process.pdf.

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995. ISBN 9780412055515.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. ISSN 1467-9868. URL http://dx.doi.org/10.1111/j.1467-9868.2010.00765.x.

E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *JMLR*, 5:1471–1530, Nov. 2005.

T. Harder, W. Boomsma, M. Paluszewski, J. Frellsen, K. E. Johansson, and T. Hamelryck. Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics*, 11(1):306, jun 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-306.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 44:97–109, Apr. 1970. doi: 10.1093/biomet/57.1.97. URL http://dx.doi.org/10.2307/2280232.

M. Hermans and B. Schrauwen. Recurrent kernel machines: Computing with infinite echo state networks. *Neural Computation*, 24(1):104–133, 2012. doi: 10.1162/NECO_a_00200. URL http://dx.doi.org/10.1162/NECO_a_00200. PMID: 21851278.

J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Neural Information Processing Systems*, 2014.

J. M. Hernández-Lobato, M. A. Gelbart, M. W. Hoffman, R. P. Adams, and Z. Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *the International Conference on Machine Learning*, 2015.

D. Hernandez-Stumpfhauser, F. J. Breidt, and M. J. van der Woerd. The general projected normal distribution of arbitrary dimension: Modeling and bayesian inference. *Bayesian Analysis*, 12(1):113–133, 03 2017. doi: 10.1214/15-BA989. URL http://dx.doi.org/10.1214/15-BA989.

C. Herzet, N. Noels, V. Lottici, H. Wymeersch, M. Luise, M. Moeneclaey, and L. Vandendorpe. Code-Aided Turbo Synchronization. *Proceedings of the IEEE*, 95(6):1255–1271, jun 2007. doi: 10.1109/JPROC.2007.896518.

G. E. Hinton. Deterministic boltzmann learning performs steepest descent in weight-space. *Neural Comput.*, 1(1):143–150, Mar. 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.1.143. URL http://dx.doi.org/10.1162/neco.1989.1.1.143.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018. URL http://dx.doi.org/10.1162/089976602760128018.

G. E. Hinton. A practical guide to training restricted boltzmann machines. Technical Report UTML TR 2010–003, University of Toronto, Toronto, Canada, Aug. 2010. URL http://www.csri.utoronto.ca/~hinton/absps/guideTR.pdf.

M. W. Hoffman and B. Shahriari. Modular mechanisms for bayesian optimization. In *the NIPS workshop on Bayesian optimization*, 2014.

D. N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, School of Mathematics,, Princeton, NJ, 1979.

J. Hubbard. Calculation of Partition Functions. *Phys. Rev. Lett.*, 3(2):77–78, jul 1959. URL http://link.aps.org/doi/10.1103/PhysRevLett.3.77.

E. Jang, S. Gu, and B. Poole. Categorical Reparameterization with Gumbel-Softmax. 2016. ISSN 1611.01144. URL https://arxiv.org/pdf/1611.01144.pdfhttp://arxiv.org/abs/1611.01144.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957a. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

E. T. Jaynes. Information theory and statistical mechanics. ii. *Phys. Rev.*, 108:171–190, Oct 1957b. doi: 10.1103/PhysRev.108.171. URL https://link.aps.org/doi/10.1103/PhysRev.108.171.

E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, Sept 1968. ISSN 0536-1567. doi: 10.1109/TSSC.1968.300117.

K. N. C. D. T. H. I. V. L. T. R. N. Joe, H. *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC New York, 1997. ISBN 9781466581432.

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

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

M. C. Jones. Perlman and wellner's circular and transformed circular copulas are particular beta and t copulas. *Symmetry*, 5(1):81–85, 2013. ISSN 2073-8994. doi: 10.3390/sym5010081. URL http://www.mdpi.com/2073-8994/5/1/81.

M. C. Jones and A. Pewsey. A family of symmetric distributions on the circle. *Journal of the American Statistical Association*, 100(472):1422–1428, 2005.

M. C. Jones, A. Pewsey, and S. Kato. On a class of circulas: copulas for circular distributions. *Annals of the Institute of Statistical Mathematics*, 67(5):843–862, Oct 2015. ISSN 1572–9052. doi: 10.1007/s10463-014-0493-6. URL https://doi.org/10.1007/s10463-014-0493-6.

C. G. Khatri and K. V. Mardia. The von mises-fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):95–106, 1977. ISSN 00359246. URL http://www.jstor.org/stable/2984884.

M. J. Kirby and R. Miranda. Circular nodes in neural networks. *Neural Computation*, 8(2):390–402, 1996. doi: 10.1162/neco.1996.8.2.390. URL http://dx.doi.org/10.1162/neco.1996.8.2.390.

A. N. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, volume 2 of *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. Springer Berlin Heidelberg, 1933. doi: 10.1007/978-3-642-49888-6.

A. Kume, S. P. Preston, and A. T. A. Wood. Saddlepoint approximations for the normalizing constant of fisher–bingham distributions on products of spheres and stiefel manifolds. *Biometrika*, 100(4):971, 2013. doi: 10.1093/biomet/ast021. URL +http://dx.doi.org/10.1093/biomet/ast021.

B. Lakshminarayanan, D. M. Roy, and Y. W. Teh. Mondrian forests: Efficient online random forests. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3140–3148. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5234-mondrian-forests-efficient-online-random-forests.pdf.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 329–336. MIT Press, 2004. URL http://papers.nips.cc/paper/2540-gaussian-process-latent-variable-models-for-visualisation-of-high-dimensional-data.pdf.

B.-J. Lee, J. Lee, and K.-E. Kim. Hierarchically-partitioned Gaussian Process Approximation. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 822–831, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/lee17a.html.

S. Li, F. Hany, and M. M. A. Modeling three-dimensional morphological structures using spherical harmonics. *Evolution*, 63(4):1003–1016, 2008. doi: 10.1111/j.1558-5646.2008.00557.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2008.00557.x.

J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 829–837, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969332.

H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. J. C. H. Watkins. Text classification using string kernels. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 563–569. MIT Press, 2001. URL http://papers.nips.cc/paper/1869-text-classification-using-string-kernels.pdf.

D. Lopez-Paz, J. M. Hernández-Lobato, and G. Zoubin. Gaussian process vine copulas for multivariate dependence. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings*

of Machine Learning Research, pages 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/lopez-paz13.html.

D. J. C. Mackay. Probable networks and plausible predictions — a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995. doi: 10.1088/0954-898X_6_3_011. URL http://www.tandfonline.com/doi/abs/10.1088/0954-898X_6_3_011.

D. J. C. MacKay. Introduction to Gaussian processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *NATO ASI Series*, pages 133–165. Springer, 1998.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2 edition, sep 2003. ISBN 978-0521642989.

C. J. Maddison, A. Mnih, and Y. W. Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. 2016. URL https://arxiv.org/pdf/1611.00712.pdfhttp://arxiv.org/abs/1611.00712.

V. M. Maksimov. Necessary and sufficient statistics for the family of shifts of probability distributions on continuous bicompact groups. *Theory of Probability & Its Applications*, 12(2):267–280, 1967. doi: 10.1137/1112029. URL http://dx.doi.org/10.1137/1112029.

K. V. Mardia. Distribution theory for the von Mises-Fisher distribution and its application. In *A Modern Course on Statistical Distributions in Scientific Work*, pages 113–130. Springer, 1975a.

K. V. Mardia. Statistics of Directional Data. *Journal of the Royal Statistical Society*, 37(3):52, jan 1975b.

K. V. Mardia. Directional statistics and shape analysis. *Journal of applied Statistics*, 26(8):949–957, 1999.

K. V. Mardia. On some recent advancements in applied shape analysis and directional statistics. *Systems Biology & Statistical Bioinformatics*, pages 9–17, 2007.

K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh. A Multivariate Von Mises Distribution with Applications to Bioinformatics. *Canadian Journal of Statistics*, 36 (1):99–109, mar 2008.

J. Martens and I. Sutskever. Parallelizable sampling of markov random fields. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 517–524, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/martens10a.html.

D. McFarlane and K. Glover. A loop-shaping design procedure using h infin; synthesis. *IEEE Transactions on Automatic Control*, 37(6):759–769, Jun 1992. ISSN 0018-9286. doi: 10.1109/9.256330.

N. Metropolis and S. M. Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335–341, Sept. 1949. ISSN 01621459. doi: 10.2307/2280232. URL http://dx.doi.org/10.2307/2280232.

T. P. Minka. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems 13*, volume 13, pages 598–604, 2000.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020, 9780262018029.

R. M. Neal. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

R. M. Neal. Density Modelling and Clustering Using Dirichlet Diffusion Trees. In J. O. B. A. P. D. J. M. Bernardo, M. J. Bayarri, editor, *Bayesian Statistics 7*, pages 619–629. Oxford University Press, 2003. URL http://www.cs.toronto.edu/~radford/ftp/dft-val7.pdf.

R. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, 2007. ISBN 9780387286785.

M. Opper and O. Winther. Expectation Consistent Approximate Inference. *Journal of Machine Learning Research*, 6:2177–2204, 12 2005.

A. Pakman and L. Paninski. Auxiliary-variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributions. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2490–2498. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5045-auxiliary-variable-exact-hamiltonian-monte-carlo-samplers-for-binary-distributions.pdf.

C. Pasarica and A. Gelman. Adaptively scaling the metropolis algorithm using expected square jumped distance. *Statistica Sinica*, 20:343–364, 2010. URL http://www3.stat.sinica.edu.tw/sstest/oldpdf/A20n113.pdf.

M. D. Perlman and J. A. Wellner. Squaring the circle and cubing the sphere: Circular and spherical copulas. *Symmetry*, 3(3):574–599, 2011. ISSN 2073-8994. doi: 10.3390/sym3030574. URL http://www.mdpi.com/2073-8994/3/3/574.

J. Pitman. Combinatorial stochastic processes. Notes for Saint Flour Summer School 621, University of California - Berkeley, 2002.

G. Pólya. Elementarer Beweis einer Thetaformel. *S-B. Akad. Wiss. Berl. (Phys.-Math. Kl.)*, pages 158–161, 1927.

J. Quiñonero-Candela and C. E. Rasmussen. A Unifying view of Sparse Approximate Gaussian Process Regression. *Journal of Machine Learning Research*, 6:1939–1959, dec 2005. ISSN 1532-4435.

J. R. Ragazzini and L. A. Zadeh. The analysis of sampled-data systems. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry*, 71(5):225–234, Nov 1952. ISSN 0097-2185. doi: 10.1109/TAI.1952.6371274.

B. Rajaratnam and D. Sparks. Mcmc-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains, 2015.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, Cambridge, Mass., 2006. ISBN 026218253X 9780262182539.

N. Razavian, H. Kamisetty, and C. J. Langmead. The von mises graphical model: Regularized structure and parameter learning. *Tech Rep CMU-CS-11-108, Carnegie Mellon University, Department of Computer Science*, 2011.

N. S. Razavian, H. Kamisetty, and C. J. Langmead. Learning generative models of molecular dynamics. *BMC genomics*, 13(Suppl 1):S5, 2012.

L. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*. Cambridge Mathematical Library. Cambridge University Press, 2000. ISBN 9780521775946.

D. M. Roy and Y. W. Teh. The mondrian process. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1377–1384. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3622-the-mondrian-process.pdf.

S. Rüping and T. Scheffer. Proceedings of the ICML 2015 Workshop on Learning with Multiple Views. http://www-ai.cs.uni-dortmund.de/MULTIVIEW2005/MultipleViews.pdf, August 2005.

C. Ryll-Nardzewski. On stationary sequences of random variables and the de finetti's equivalence. *Colloquium Mathematicae*, 4(2):149–156, 1957. URL http://eudml.org/doc/210023.

R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/salakhutdinov09a.html.

A. Santos, R. Wernersson, and L. J. Jensen. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research*, 43(D1):D1140–D1144, 2015. doi: 10.1093/nar/gku1092. URL http://nar.oxfordjournals.org/content/43/D1/D1140.abstract.

M. Scholz. Analysing periodic phenomena by circular pca. In S. Hochreiter and R. Wagner, editors, *Proceedings of the Conference on Bioinformatics Research and Development (BIRD'07)*, volume 4414 of *LNCS/LNBI*, pages 38–47, Berlin, 2007. Springer-Verlag.

S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control: Analysis and Design*. Wiley, Hoboken, NJ, 2 edition, 2005. ISBN 978-0-470-01167-6.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006a. URL http://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf.

E. Snelson and Z. Ghahramani. Variable noise and dimensionality reduction for sparse gaussian processes. In R. Dechter and T. S. Richardson, editors, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006b.

S. Sra. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of Is(x). *Computational Statistics*, 27(1):177–190, 2012.

R. L. Stratonovich. On a Method of Calculating Quantum Distribution Functions. *Soviet Physics Doklady*, 2:416, jul 1957.

J. Taghia, Z. Ma, and A. Leijon. On von-mises fisher mixture model in text-independent speaker identification. In *INTERSPEECH*, pages 2499–2503, 2013.

H. Tang, S. M. Chu, and T. S. Huang. Generative model-based speaker clustering via mixture of von Mises-Fisher distributions. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4101–4104. IEEE, 2009.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. Technical Report NCRG/97/10, Aston University, Birmingham, UK, 9 1997.

M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. ISSN 1467-9868. doi: 10.1111/1467-9868.00196.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL http://proceedings.mlr.press/v5/titsias09a.html.

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

R. D. Turner. *Gaussian Processes for State Space Models and Change Point Detection*. PhD thesis, University of Cambridge, Cambridge, UK, July 2011.

R. E. Turner, P. Berkes, and M. Sahani. Two problems with variational expectation maximisation for time-series models. In *Workshop on Inference and Estimation in Probabilistic Time-Series Models*, volume 2, 2008.

R. E. von Mises. Über die 'Ganzzahligkeit' der Atomgewichte und verwandte Fragen. *Physikalische Zeitschrift*, 19:490–500, 1918.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, Jan. 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001.

Z. Wang, S. Mohamed, and N. Freitas. Adaptive hamiltonian and riemann manifold monte carlo. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1462–1470, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/wang13e.html.

C. K. I. Williams. Computing with infinite networks. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 295–301. MIT Press, 1997. URL http://papers.nips.cc/paper/1197-computing-with-infinite-networks.pdf.

E. A. Yfantis and L. E. Borgman. An extension of the von Mises distribution. *Communications in Statistics - Theory and Methods*, 11(15):1695–1706, jan 1982. ISSN 0361-0926. doi: 10.1080/03610928208828342.

R. S. Zemel, C. K. I. Williams, and M. C. Mozer. Directional-unit boltzmann machines. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 172–179. Morgan-Kaufmann, 1993. URL http://papers.nips.cc/paper/674-directional-unit-boltzmann-machines.pdf.

Y. Zhang, Z. Ghahramani, A. J. Storkey, and C. A. Sutton. Continuous Relaxations for Discrete Hamiltonian Monte Carlo. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3194–3202. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4652-continuous-relaxations-for-discrete-hamiltonian-monte-carlo.pdf.

Appendix A

Sparse models using the augmented representations

A sparse version of the augmented representation of the $m\mathcal{G}v\mathcal{M}$ can be generated drawing on sparse Gaussian Process representations. In particular, we can apply the exchangeability property to the augmented states to reparametrise the augmenting states by a low-dimension set of variables $\mathbf{u}$ that induce the distribution over the augmented state $\mathbf{f}$, i.e.

$$p(\mathbf{u}) = \mathcal{N}\Big(\mathbf{u}; 0, \mathbf{K}_{\mathsf{u},\mathsf{u}}\Big) \tag{A.1}$$

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}\Big(\mathbf{f}; \boldsymbol{m}, \mathbf{Q}_{\mathsf{f},\mathsf{f}}\Big) \tag{A.2}$$

$$p(\boldsymbol{\phi}|\mathbf{f}, \mathbf{g}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} v\mathcal{M}(\phi_{\mathrm{n}}; \alpha_{\mathrm{n}}(\mathbf{f}), \beta_{\mathrm{n}}(\mathbf{f})) \tag{A.3}$$

where $\boldsymbol{m}_{\mathsf{f}} = \mathbf{K}_{\mathsf{fu}}\mathbf{K}_{\mathsf{u},\mathsf{u}}^{-1}\mathbf{u}$, and the covariance $\mathbf{Q}_{\mathsf{ff}}$ is a function of $\mathbf{K}_{\mathsf{ff}} - \mathbf{K}_{\mathsf{fu}}\mathbf{K}_{\mathsf{uu}}^{-1}\mathbf{K}_{\mathsf{uf}}$.

The function to be employed in obtaining $\mathbf{Q}_{\mathsf{ff}}$ determines the sparsity pattern of the model as is the case for sparse Gaussian Processes. For example, Quiñonero-Candela and Rasmussen (2005) show that the Partially Independent Training Conditional (PITC) approximation for Gaussian process assumes that

$$\mathbf{Q}_{\mathsf{ff}} = \underset{\mathcal{S}_1,\ldots,\mathcal{S}_{\mathrm{D}}}{\mathrm{blkdiag}}\Big(\mathbf{K}_{\mathsf{fu}}\mathbf{K}_{\mathsf{uu}}^{-1}\mathbf{K}_{\mathsf{uf}}\Big) \tag{A.4}$$

where $\mathrm{blkdiag}_{\mathcal{S}_1,\ldots,\mathcal{S}_{\mathrm{D}}}$ extracts a the block diagonal matrix over D index subsets $\mathcal{S}_1, \ldots, \mathcal{S}_{\mathrm{D}}$. A particualr case of the PITC model is the Fully-Independent Training Conditional (FITC) model, where each index subset contains only a single element,

i.e.,

$$\mathbf{Q_{ff}} = \text{diag}\Big(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}\Big) \tag{A.5}$$

where diag extracts only the diagonal elements of its matrix argument. Another sparse approximation closely linked to the FITC approximation is the Variational Free Energy model of Titsias (2009). In both cases, the covariance $\mathbf{K_{uu}}$ is dense and the overall complexity of the approximation is of order $\mathcal{O}(\mathrm{M}^2 \times \mathrm{N})$ computations, where M is the number of inducing points u.

More recently, sparse models have considered special covariate structure for the inducing points $\mathbf{u}$ to alleviate this constraint. Bui and Turner (2014) proposed structuring the inducing points covariance as a block-diagonal matrix, which has $\mathcal{O}(\mathrm{D}^2\mathrm{N})$ complexity, where D is the average number of observations per block. Lee et al. (2017) have expanded over this work and considered a tree-structured approximation that can be summarized as nesting PITC approximations over the inducing inputs.

A graphical summary for the models for FITC, PITC and tree-based approximations are presented in Figure A.1.

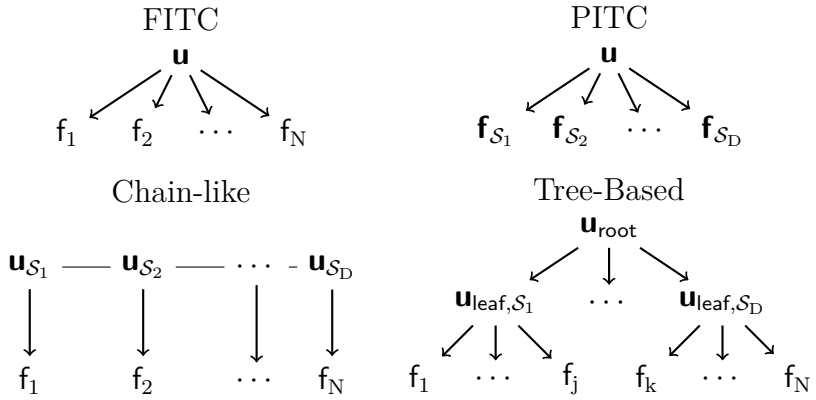


Fig. A.1 The graphical model for the PITC, FITC, Chain-like and Tree-based approximations.

Appendix B

Variational inference for the augmented $m\mathcal{G}v\mathcal{M}$ representation

In this section, we outline a variational free energy approximation based on Titsias (2009) for the augmented model.

Recall that the augmented representation for a $m\mathcal{G}v\mathcal{M}(\boldsymbol{\phi}, \boldsymbol{\kappa}, \boldsymbol{\mu}, \mathbf{K})$ was presented in Chapter 3 as the joint model over the circular variables $\boldsymbol{\phi}$ and the augmentation states $\mathbf{f}$ given by

$$p(\boldsymbol{\phi}, \mathbf{f}) = \underbrace{\mathcal{N}\left(\mathbf{f}; \boldsymbol{m}, \mathbf{A}(\sigma^2\mathbf{I} - \mathbf{K}^{-1})\mathbf{A}^\top\right)}_{p(\mathbf{f})} \times \underbrace{\prod_{\mathrm{n}=1}^{\mathrm{N}} \left[v\mathcal{M}(\phi_\mathrm{n}; \alpha_\mathrm{n}(\mathbf{f}), \beta_\mathrm{n}(\mathbf{f}))\right]}_{p(\boldsymbol{\phi}|\mathbf{f})} \tag{B.1}$$

with parameters $\mathbf{A}$ an arbitrary real-valued invertible matrix, $\boldsymbol{m}$ an arbitrary mean vector, σ^2 a constant chosen so that the matrix $\sigma^2\mathbf{I} - \mathbf{K}^{-1}$ is positive-definite, where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ defined as

$$\begin{bmatrix} \boldsymbol{\alpha} \odot \cos\boldsymbol{\beta} \\ \boldsymbol{\alpha} \odot \sin\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\kappa} \odot \cos\boldsymbol{\mu} \\ \boldsymbol{\kappa} \odot \sin\boldsymbol{\mu} \end{bmatrix} + \mathbf{A}^{-1}(\mathbf{f} - \boldsymbol{m}) \tag{B.2}$$

with $\odot$ denotes point-wise multiplication.

Titsias (2009) suggested constructing the variational approximation distributions to produce cancellations between difficult terms arising from likelihoods in Gaussian Process models. Using the same principle to form an approximation for the augmented model, the approximating distribution takes the form

$$q(\boldsymbol{\phi}, \mathbf{f}) = p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f}). \tag{B.3}$$

This approximation results in a simplification of the free energy so that

$$\mathcal{F}(q) = \left\langle \log \frac{p(\boldsymbol{\psi}, \boldsymbol{\phi}, \mathbf{f})}{q(\boldsymbol{\phi}, \mathbf{f})} \right\rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})} \tag{B.4}$$

$$= \left\langle \log p(\boldsymbol{\psi}|\boldsymbol{\phi}) + \log \frac{\cancel{p(\boldsymbol{\phi}|\mathbf{f})}p(\mathbf{f})}{\cancel{p(\boldsymbol{\phi}|\mathbf{f})}q(\mathbf{f})} \right\rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})} \tag{B.5}$$

$$= \langle \log p(\boldsymbol{\psi}|\boldsymbol{\phi}) \rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})} - \mathrm{KL}(q(\mathbf{f}) \| p(\mathbf{f})). \tag{B.6}$$

Expanding the likelihood term of Equation (B.6) reveals the mean sine and cosine relationship,

$$\begin{aligned} \mathcal{F}(q) = &-\mathrm{KL}(q(\mathbf{f}) \| p(\mathbf{f})) - \mathrm{N} \log \mathcal{I}_0(\kappa) + \mathrm{N} \log 2\pi \\ &+ \kappa \cos \boldsymbol{\psi}^\top \langle \cos \boldsymbol{\phi} \rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})} + \kappa \sin \boldsymbol{\psi}^\top \langle \sin \boldsymbol{\phi} \rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})}, \end{aligned} \tag{B.7}$$

which admit further simplification. In particular, the sine and cosine expectations can be analytically solved with respect to $p(\boldsymbol{\phi}|\mathbf{f})$ to yield

$$\langle \cos \phi_\mathrm{n} \rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})} = \left\langle \frac{\mathcal{I}_1(\alpha_\mathrm{n}(\mathbf{f}))}{\mathcal{I}_0(\alpha_\mathrm{n}(\mathbf{f}))} \cos \beta_\mathrm{n}(\mathbf{f}) \right\rangle_{q(\mathbf{f})} \tag{B.8}$$

$$\langle \sin \phi_\mathrm{n} \rangle_{p(\boldsymbol{\phi}|\mathbf{f})q(\mathbf{f})} = \left\langle \frac{\mathcal{I}_1(\alpha_\mathrm{n}(\mathbf{f}))}{\mathcal{I}_0(\alpha_\mathrm{n}(\mathbf{f}))} \sin \beta_\mathrm{n}(\mathbf{f}) \right\rangle_{q(\mathbf{f})}, \tag{B.9}$$

Despite not admitting closed-form expressions to evaluate Equation (B.8) and Equation (B.9), these expectations can be computed numerically with quadrature schemes or approximated through simple Monte Carlo.

To complete the variational specification, we define define $q(\mathbf{f}) = \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$. Substituting Equation (B.8) Equation (B.9) and the variational distribution chosen for $\mathbf{f}$ in Equation (B.7), yields the free energy expression

$$\begin{aligned} \mathcal{F}(q) = &\mathrm{N} \left(\log 2\pi + 1 - \log \mathcal{I}_0(\kappa) \right) + \kappa \cos \boldsymbol{\psi}^\top \left\langle \frac{\mathcal{I}_1(\alpha_\mathrm{n}(\mathbf{f}))}{\mathcal{I}_0(\alpha_\mathrm{n}(\mathbf{f}))} \cos \beta_\mathrm{n}(\mathbf{f}) \right\rangle_{q(\mathbf{f})} \\ &+ \kappa \sin \boldsymbol{\psi}^\top \left\langle \frac{\mathcal{I}_1(\alpha_\mathrm{n}(\mathbf{f}))}{\mathcal{I}_0(\alpha_\mathrm{n}(\mathbf{f}))} \sin \beta_\mathrm{n}(\mathbf{f}) \right\rangle_{q(\mathbf{f})} + \frac{1}{2} \log \frac{\det \boldsymbol{\Sigma}}{\det \mathbf{A}(\sigma^2 \mathbf{I} - \mathbf{K}^{-1})\mathbf{A}^\top} \\ &- \frac{1}{2} \mathrm{Tr} \left[\mathbf{A}^{-\top} (\sigma^2 \mathbf{I} - \mathbf{K}^{-1})^{-1} \mathbf{A}^{-1} \left(\boldsymbol{\Sigma} + (\boldsymbol{m} - \boldsymbol{\nu})(\boldsymbol{m} - \boldsymbol{\nu})^\top \right) \right] \end{aligned} \tag{B.10}$$

which can be maximised using gradient-based methods. We note however, that up to the submission of this thesis, automatic differentiation packages for machine learning such as Tensorflow and Theano offer no support for exponentially-weighted implementations

of modified Bessel functions of first kind required for numerically stable implementations of ratio of Bessel functions.

Appendix C

Improving numerical stability of the Generalised von Mises moment calculations

Gatto and Jammalamadaka (2007) derived the moments for the Generalised von Mises distribution using series approximations. During this work, it was verified that the formulas provided needed a correction and are given as

$$G_{\mathrm{n}}(\mu_1-\mu_2,\kappa_1,\kappa_2)=\mathcal{I}_0(\kappa_2)\mathcal{I}_{\mathrm{n}}(\kappa_1)+\sum_{\mathrm{j}=1}^{\infty}\cos(2\mathrm{j}\mu_1-\mu_2)\mathcal{I}_{\mathrm{j}}(\kappa_2)\Big[\mathcal{I}_{2\mathrm{j}+\mathrm{n}}(\kappa_1)+\mathcal{I}_{|2\mathrm{j}-\mathrm{n}|}(\kappa_1)\Big] \tag{C.1}$$

$$H_{\mathrm{n}}(\mu_1-\mu_2,\kappa_1,\kappa_2)=\sum_{\mathrm{j}=1}^{\infty}\sin(2\mathrm{j}\mu_1-\mu_2)\mathcal{I}_{\mathrm{j}}(\kappa_2)\Big[\mathcal{I}_{2\mathrm{j}+\mathrm{n}}(\kappa_1)-\mathcal{I}_{|2\mathrm{j}-\mathrm{n}|}(\kappa_1)\Big]. \tag{C.2}$$

Naive implementation of these series suffer from two major problems: the fact that modified Bessel functions grow exponentially and the difficulty in evaluating oscillating series.

The overflow problem can be overcome by using exponentially-weighted implementations of modified Bessel functions of first kind and appropriate weighting. In particular, for non-zero κ_1, we can form the alternative series

$$\frac{G_{\mathrm{n}}(\mu_1-\mu_2,\kappa_1,\kappa_2)}{\mathcal{I}_0(\kappa_2)\mathcal{I}_0(\kappa_1)}=\mathcal{R}_{\mathrm{j},0}(\kappa_1)+\sum_{\mathrm{j}=1}^{\infty}\cos(2\mathrm{j}\mu_1-\mu_2)\mathcal{R}_{\mathrm{j},0}(\kappa_2)\Big[\mathcal{R}_{2\mathrm{j}+\mathrm{n},0}(\kappa_1)+\mathcal{R}_{|2\mathrm{j}-\mathrm{n}|,0}(\kappa_1)\Big] \tag{C.3}$$

$$\frac{H_{\mathrm{n}}(\mu_1-\mu_2,\kappa_1,\kappa_2)}{\mathcal{I}_0(\kappa_2)\mathcal{I}_0(\kappa_1)}=\sum_{\mathrm{j}=1}^{\infty}\sin(2\mathrm{j}\mu_1-\mu_2)\mathcal{R}_{\mathrm{j},0}(\kappa_2)\Big[\mathcal{R}_{2\mathrm{j}+\mathrm{n},0}(\kappa_1)-\mathcal{R}_{|2\mathrm{j}-\mathrm{n}|,0}(\kappa_1)\Big]. \tag{C.4}$$

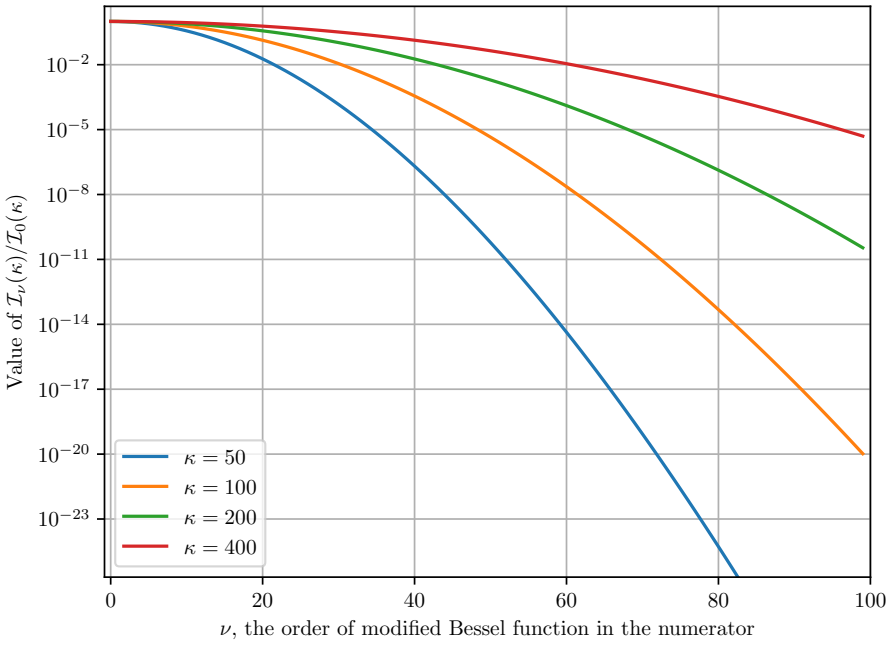


Fig. C.1 Modified Bessel function ratio value decrease with the order of the numerator Bessel function.

The series in Equation (C.3) and Equation (C.4) are substantially more stable than the forms presented in Equation (C.1) and Equation (C.2). The principal reason for this lies in the rate at which the ratio of modified Bessel functions decrease value with their order, illustrated in Figure C.1. We also found empirically that numerically safe implementation of the moments also require limiting the concentration parameters to 500 to avoid overflow issues.

The second numerical issue arising from the use of Equation (C.1) and Equation (C.2) is the non-decreasing series. The source of the non-decrease lies in the trigonometric terms of the series as all other terms are monotonically decreasing. Hence, a simple way to side-step this issue is to we aggregate terms of the series to produce strictly decreasing terms, i.e. series of the form

$$\frac{G_{\mathrm{n}}(\mu_1 - \mu_2, \kappa_1, \kappa_2)}{\mathcal{I}_0(\kappa_2)\mathcal{I}_0(\kappa_1)} = \mathcal{R}_{\mathrm{j},0}(\kappa_1) + \sum_{\mathrm{m}=1}^{\infty} \mathcal{S}_G(\mathrm{m}, \mathrm{n}, \mu_1 - \mu_2, \kappa_1, \kappa_2) \tag{C.5}$$

$$\frac{H_{\mathrm{n}}(\mu_1 - \mu_2, \kappa_1, \kappa_2)}{\mathcal{I}_0(\kappa_2)\mathcal{I}_0(\kappa_1)} = \sum_{\mathrm{m}=1}^{\infty} \mathcal{S}_H(\mathrm{m}, \mathrm{n}, \mu_1 - \mu_2, \kappa_1, \kappa_2) \tag{C.6}$$

where the aggregated terms $\mathcal{S}_G$ and $\mathcal{S}_H$ are strictly decreasing. We can construct $\mathcal{S}_G$ and $\mathcal{S}_H$ leveraging on the periodic nature of sines and cosines. In particular, by finding the period T associated with $\mu_1 - \mu_2$, we can form the updates

$$\begin{aligned}\mathcal{S}_G(\mathrm{m}, \mathrm{n}, \mu_1 - \mu_2, \kappa_1, \kappa_2) = \sum_{\mathrm{j}=1}^{T} \cos(2\mu_1 - \mu_2(\mathrm{m} \cdot T + \mathrm{j}))\mathcal{R}_{(\mathrm{m}\cdot T+\mathrm{j}),0}(\kappa_2)\times \\ \Big[\mathcal{R}_{2(\mathrm{m}\cdot T+\mathrm{j})+\mathrm{n},0}(\kappa_1) + \mathcal{R}_{|2(\mathrm{m}\cdot T+\mathrm{j})-\mathrm{n}|,0}(\kappa_1)\Big]\end{aligned} \tag{C.7}$$

$$\begin{aligned}\mathcal{S}_H(\mathrm{m}, \mathrm{n}, \mu_1 - \mu_2, \kappa_1, \kappa_2) = \sum_{\mathrm{j}=1}^{T} \sin(2\mu_1 - \mu_2(\mathrm{m} \cdot T + \mathrm{j}))\mathcal{R}_{(\mathrm{m}\cdot T+\mathrm{j}),0}(\kappa_2)\times \\ \Big[\mathcal{R}_{2(\mathrm{m}\cdot T+\mathrm{j})+\mathrm{n},0}(\kappa_1) - \mathcal{R}_{|2(\mathrm{m}\cdot T+\mathrm{j})-\mathrm{n}|,0}(\kappa_1)\Big].\end{aligned} \tag{C.8}$$

which are stricly decreasing.