Stat Comput
DOI 10.1007/s11222-017-9765-3

On the exact maximum likelihood inference of Fisher–Bingham distributions using an adjusted holonomic gradient method

A. Kume[1] · **T. Sei**[2]

Received: 6 October 2016 / Accepted: 9 July 2017

Abstract Holonomic function theory has been successfully implemented in a series of recent papers to efficiently calculate the normalizing constant and perform likelihood estimation for the Fisher–Bingham distributions. A key ingredient for establishing the standard holonomic gradient algorithms is the calculation of the Pfaffian equations. So far, these papers either calculate these symbolically or apply certain methods to simplify this process. Here we show the explicit form of the Pfaffian equations using the expressions from Laplace inversion methods. This improves on the implementation of the holonomic algorithms for these problems and enables their adjustments for the degenerate cases. As a result, an exact and more dimensionally efficient ODE is implemented for likelihood inference.

Keywords Bingham distributions · Fisher–Bingham distributions · Directional statistics · Holonomic functions

1 Introduction

The Fisher–Bingham distribution is defined as the conditional distribution of a general multivariate normal distribution on a unit sphere. In particular, for a p-dimensional multivariate normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the corresponding density function with respect to $d_{\mathcal{S}^{p-1}}(\boldsymbol{x})$, the uniform measure in the $p-1$- dimensional sphere $\mathcal{S}^{p-1}$, is

$$f(\boldsymbol{x}) = \frac{1}{\mathcal{C}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)} e^{-\frac{\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}}{2} + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} d_{\mathcal{S}^{p-1}}(\boldsymbol{x})$$
$$\propto e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}} \mathbf{1}(\boldsymbol{x}^\top \boldsymbol{x} = 1)$$

where

$$\mathcal{C}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) = \int_{\mathcal{S}^{p-1}} e^{-\frac{\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}}{2} + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} d_{\mathcal{S}^{p-1}}(\boldsymbol{x})$$

is the normalizing constant. Since multiplication by any orthogonal transformation induces isometry in $\mathcal{S}^{p-1}$,

$$\mathcal{C}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) = \mathcal{C}\left(\frac{\boldsymbol{\Delta}^{-1}}{2}, \boldsymbol{\Delta}^{-1}\boldsymbol{O}\boldsymbol{\mu}\right)$$

where $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1^2, \delta_2^2, \ldots, \delta_p^2)$ and the orthogonal matrix $\boldsymbol{O} \in \mathcal{O}(p)$ are obtained from the singular value decomposition of $\boldsymbol{\Sigma} = \boldsymbol{O}^\top \boldsymbol{\Delta} \boldsymbol{O}$. Similarly, we can also choose the particular $\boldsymbol{O}$ such that entries of $\boldsymbol{\Delta}^{-1}\boldsymbol{O}\boldsymbol{\mu}$ are non-negative. Hence, without loss of generality, we can assume that the covariance parameter is diagonal, and therefore, a more efficient parametrization of dimension $2p$ can be used for the normalizing constant

$$\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \int_{\mathcal{S}^{p-1}} e^{\sum_{i=1}^p (-\theta_i x_i^2 + \gamma_i x_i)} d_{\mathcal{S}^{p-1}}(\boldsymbol{x})$$

Electronic supplementary material The online version of this article (doi:10.1007/s11222-017-9765-3) contains supplementary material, which is available to authorized users.

✉ A. Kume
a.kume@kent.ac.uk

T. Sei
sei@mist.i.u-tokyo.ac.jp

[1] SMSAS, University of Kent, Canterbury, UK

[2] Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan

Published online: 09 August 2017

with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p) = \text{diag}(\frac{\boldsymbol{\Delta}^{-1}}{2})$, i.e. $\theta_i = \frac{1}{2\delta_i^2}$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p) = \boldsymbol{\Delta^{-1} O \mu}$. Note the slight inconsistency in notation as we write $\mathcal{C}(\text{diag}(\boldsymbol{\theta}), \boldsymbol{\gamma}) = \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. The special case of $\boldsymbol{\gamma} = 0$ corresponds to the Bingham distributions studied separately in Wood (1993); Kume and Wood (2007); Sei and Kume (2015). Despite the fact that these distributions are part of the exponential family (see e.g. Mardia and Jupp 2000), maximum likelihood estimation ultimately involves numerical routines for approximating the normalizing constant term $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. The method of Kume and Wood (2005) relies on the saddlepoint approximation which is known to be not only very close to the exact value with little computational cost but also numerically stable. However, recently there has been a renewed interest in this problem with the implementation of the holonomic gradient method (HGM), which in theory is exact since the problem of calculating $\mathcal{C}$ is mathematically characterized via a solution of an ODE (see e.g. Nakayama et al. 2011; Hashiguchi et al. 2013; Sei et al. 2013; Koyama 2011; Koyama and Takemura 2016; Koyama et al. 2014 and Koyama et al. 2012).

In particular, the HGM approach generates exact solutions if the corresponding ODE is numerically stable and the dimensionality of the parameters is not extremely large. Please note that, in the relevant literature, numerically unstable ODE's are called stiff (see eg 10.6 in Zarowsky 2004). Koyama et al. (2014) focus on the numerical efficiency of HGM implementation by expressing the corresponding Pfaffian equations (see Sect. 4) in terms of some elementary matrices R_i and Q_i. Note that HGM is applicable not only to $\mathcal{C}$ but also any holonomic function. See Chapter 6 of Hibi (2013) for more details.

The contribution in this paper is threefold. Firstly, by expanding the Laplace transform in Eq. (1) in partial fractions, we obtain the Pfaffian equations explicitly in terms of only two vector parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, each of length p. This makes the differential structure of these functions more transparent and the implementation of the holonomic algorithm more dimensionally and computationally efficient, since at most $2p$ parameters are needed for the normalizing constant and there is no need to use symbolic algebra packages for generating the Pfaffians explicitly.

Secondly, by imposing some constraints on θ_i and γ_i, our approach is easily applied to many important sub-classes within the Fisher–Bingham family (such as the Bingham, Watson and Kent distributions). In fact, the general methodology of HGM algorithms does not automatically apply to these situations because the Pfaffian equations become degenerate. In particular, the corresponding ODE is stiff if some eigenvalues of Δ coalesce. Therefore, special attention for cases with various multiplicities in parameters is practically useful in the model selection process. Our explicit Pfaffian expressions, however, require minimal adjustments for these degenerate cases. The special case of Bingham distribution appearing when all γ_i's are zero is considered separately by Sei and Kume (2015). However, in this paper our approach is more general and accommodates all possible variations in the parameter space. Therefore, we can easily perform model selection within the Fisher–Bingham family based on the standard likelihood ratio tests. If we only need to evaluate the normalizing constant and the first-order derivatives at these degenerate points, the HGM with respect to the radius parameter can be applied as Koyama et al. (2014) and Koyama and Takemura (2016) suggested. However, if we also have to evaluate higher-order derivatives (e.g. standard errors for MLE) or apply the ODE along any general curve and not just as radial rescaling of parameters, the Pfaffian system in our paper is necessary.

Finally, while many papers focus on the normalizing constant, there has not been much interest in the estimation of the orthogonal component $\mathbf{O}$ from the real data. For $p = 3$, this problem is tackled in Kent (1982) where a closed form solution is shown for a very useful family of spherical distributions. However, for general p such a solution is not available. We combine the holonomic gradient method for the normalizing constant with that of a particular solution on orthogonal matrices $\mathbf{O}$ so that a maximum likelihood estimator is evaluated. This method is shown to work well in both simulated and real data examples, but special care is needed in the general setting for the Fisher–Bingham distributions due to multimodality of the likelihood function for these members of the curved exponential family.

The paper is organized as follows. We start with general remarks about the Fisher–Bingham normalizing constant where we provide a simple univariate integral representation. We then give a brief introduction to the holonomic gradient method which characterizes the the evaluation of $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ as a solution of an ordinary differential system of equations. The explicit expressions for the Pfaffian equations needed for such ODE in the case of the Fisher–Bingham integral are given in the next section where degenerate cases with multiplicities on the parameters are specifically addressed. We then focus on the implementation of the proposed MLE approach for both degenerate and non-degenerate cases of Fisher–Bingham distributions so that some log-likelihood ratio test can be used for choosing the appropriate model.

2 Laplace inversion representation

2.1 General case

Based on the key result in Proposition 1 from Kume and Wood (2005), one can easily derive that

$$\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = 2\pi^{p/2} \prod_{i=1}^{p} \theta_i^{-1/2} e^{\sum_{i=1}^{p} \frac{\gamma_i^2}{4\theta_i}} f_r(1)$$

where $f_r(1)$ is the density at point 1 of $r = \sum_{i=1}^p y_i^2$, while y_i are independent normal random variables as $y_i \sim N(\frac{\gamma_i}{2\theta_i}, \frac{1}{2\theta_i})$ with $\theta_i = \frac{1}{2\delta_i^2}$ and $\gamma_i = \frac{\mu_i}{\delta_i^2} > 0$. Since the random variable r takes non-negative values, the Laplace transform of its density is the same as its moment generating function (with a sign switch in its argument) which in our parametrization is:

$$\mathcal{L}(t) = \frac{e^{\sum_{i=1}^p \frac{\gamma_i^2}{4(\theta_i+t)} - \frac{\gamma_i^2}{4\theta_i}}}{\prod_{i=1}^p \sqrt{1+t/\theta_i}}.$$

Applying the inverse Laplace transform

$$f_r(1) = \frac{1}{2\pi \mathrm{i}} \int_{\mathrm{i}\mathbb{R}+t_0} \mathcal{L}(t) e^t \mathrm{d}t = \frac{1}{2\pi \mathrm{i}} \int_{\mathrm{i}\mathbb{R}+t_0} \frac{e^{\sum_{i=1}^p \frac{\gamma_i^2}{4(\theta_i+t)} - \frac{\gamma_i^2}{4\theta_i}}}{\prod_{i=1}^p \sqrt{1+t/\theta_i}} e^t \mathrm{d}t,$$

for any $-t_0 < \min(\boldsymbol{\theta})$, implies

$$\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \int_{\mathcal{S}^{p-1}} e^{\sum_{i=1}^p (-\theta_i x_i^2 + \gamma_i x_i)} d_{\mathcal{S}^{p-1}}(\boldsymbol{x}) = \int_{\mathrm{i}\mathbb{R}+t_0} \mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\theta}) e^t \mathrm{d}t \tag{1}$$

where

$$\mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{2\pi^{p/2}}{2\pi i} \prod_{i=1}^p \frac{e^{\frac{\gamma_i^2}{4(\theta_i+t)}}}{\sqrt{\theta_i + t}}.$$

Equation (1) establishes the general Fisher–Bingham normalizing constant in terms of an univariate complex integration. In particular, it is easily seen from (1) and the definition of $\mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\theta})$ that for any $c \in \mathbb{R}$

$$\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathcal{C}(\boldsymbol{\theta}, |\boldsymbol{\gamma}|) \quad \text{and} \quad \mathcal{C}(\boldsymbol{\theta} - c, \boldsymbol{\gamma}) = \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) e^c \tag{2}$$

where $|\boldsymbol{\gamma}|$ is the vector of absolute values of $\boldsymbol{\gamma}$. Therefore, without loss of generality we can assume that both vector parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ have non-negative entries.

2.2 Degenerate cases

Constraints on the parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ could lead to degeneracy in the corresponding ODE. For statistical inference however, some model constraints in the Fisher–Bingham distributions are necessary for practical use. Such models induce constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ for the corresponding normalizing constants as follows (c.f. Mardia and Jupp 2000, Table 9.2):

- Bingham distribution is generated if $\boldsymbol{\gamma}$ is set to zero.
- Fisher–Watson if $\theta_2 = \theta_3 = \cdots = \theta_p$ and $\gamma_3 = \gamma_4 = \cdots = \gamma_p = 0$
- Kent distributions if $\gamma_2 = \gamma_3 = \cdots = \gamma_p = 0$ and $\sum_{i=1}^p \theta_i = p\theta_1$
- von Mises–Fisher if $\theta_1 = \theta_2 = \cdots = \theta_p$
- Bingham–Mardia if $\theta_2 = \theta_3 = \cdots = \theta_p$ and $\gamma_2 = \gamma_3 = \cdots = \gamma_p = 0$
- Watson if $\theta_2 = \theta_3 = \cdots = \theta_p$ and $\gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$

Note that property (2) implies that θ_i can be assumed strictly positive. Alternatively, this property implies that we can also fix one entry θ_i to a fixed value and hence reduce the dimension by one, but we will not concern ourselves here with that. Of the models mentioned above, degeneracy appears in the corresponding ODE if one or two of the following scenarios occur:

(a) some entries in $\boldsymbol{\theta}$ coincide.
(b) some entries in $\boldsymbol{\gamma}$ are zero.

In order to accommodate scenario (a), let us assume that we have l distinct values such that each θ_i has multiplicity n_i, i.e. $n_1 + n_2 + \ldots + n_l = p$. Let us index the corresponding n_i entries of $\boldsymbol{\gamma}$ as $\gamma_{1,i}, \ldots, \gamma_{n_i,i}$. From the integral representation of

$$\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{2\pi^{p/2}}{2\pi i} \int_{\mathrm{i}\mathbb{R}+t_0} \prod_{i=1}^l \frac{e^{\frac{\sum_{r=1}^{n_i} \gamma_{r,i}^2}{4(\theta_i+t)}}}{(\theta_i+t)^{n_i/2}} e^t \mathrm{d}t,$$

it is clear that its value depends on only the summation terms $\sum_{r=1}^{n_i} \gamma_{r,i}^2$ and not on the particular values $\gamma_{r,i}^2$. This implies that for scenario (b), we can work with $\sum_{r=1}^{n_i} \gamma_{r,i}^2 = \gamma_i^2$ and perform the required differentiation only with respect to this particular $\gamma_{1,i} = \gamma_i = \sqrt{\sum_{r=1}^{n_i} \gamma_{r,i}^2}$, while the other $\gamma_{r,i}^2$ remain zero. As a result,

$$\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \frac{2\pi^{p/2}}{2\pi i} \int_{\mathrm{i}\mathbb{R}+t_0} \prod_{i=1}^l \frac{e^{\frac{\gamma_i^2}{4(\theta_i+t)}}}{(\theta_i+t)^{n_i/2}} e^t \mathrm{d}t, \tag{3}$$

and without loss of generality, we can focus on evaluating (3) with l distinct θ_i, while (1) is derived from above if $n_i = 1$ for all i. In the remainder of the paper, we will focus on evaluating $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ as in (3) where $\boldsymbol{\theta}$ has l distinct values.

3 Holonomic gradient method

In this section, we briefly review the framework of the holonomic gradient methods. See Nakayama et al. (2011),

Hashiguchi et al. (2013), Sei et al. (2013), Koyama (2011), Koyama et al. (2014) and Koyama et al. (2012) for details and further information.

Let Θ be an open subset of the d-dimensional Euclidean space. Denote the partial derivative $\partial/\partial\alpha_i$ by ∂_i. A function $c(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha} \in \Theta$ is called *holonomic* if there exists a finite-dimensional (say r-dimensional) column vector $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{\alpha})$ consisting of (possibly $c(\boldsymbol{\alpha})$ and higher-order) partial derivatives of $c(\boldsymbol{\alpha})$ such that $\boldsymbol{g}$ satisfies

$$\partial_i \boldsymbol{g}(\boldsymbol{\alpha}) = \boldsymbol{P}_i(\boldsymbol{\alpha})\boldsymbol{g}(\boldsymbol{\alpha}), \quad i = 1, \ldots, d, \tag{4}$$

where $\boldsymbol{P}_i(\boldsymbol{\alpha})$ is a $r \times r$-matrix of rational functions of $\boldsymbol{\alpha}$. For example, the trigonometric function $c(\alpha) = \sin\alpha$ is holonomic since it satisfies

$$\partial \begin{pmatrix} c \\ \partial c \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \partial c \end{pmatrix},$$

where $\partial = \partial/\partial\alpha$, $d = 1$ and $r = 2$. It is known that the normalizing constants of the von Mises–Fisher, Bingham and Fisher–Bingham distributions are holonomic.

The Eq. (4) is called *the Pfaffian equation* of $\boldsymbol{g}$. This equation essentially states that higher- order derivatives of $\boldsymbol{g}(\boldsymbol{\alpha})$ are linear combinations of its entries while involving the Pfaffian matrices as rescaling constants. For example, the second-order derivative is

$$\begin{aligned} \partial_i\partial_j \boldsymbol{g} &= \partial_i(\boldsymbol{P}_j \boldsymbol{g}) = (\partial_i \boldsymbol{P}_j)\boldsymbol{g} + \boldsymbol{P}_j(\partial_i \boldsymbol{g}) \\ &= (\partial_i \boldsymbol{P}_j)\boldsymbol{g} + \boldsymbol{P}_j \boldsymbol{P}_i \boldsymbol{g} \\ &= (\partial_i \boldsymbol{P}_j + \boldsymbol{P}_j \boldsymbol{P}_i)\boldsymbol{g}. \end{aligned}$$

Assume that a numerical value of the vector $\boldsymbol{g}(\boldsymbol{\alpha}^{(0)})$ at some point $\boldsymbol{\alpha}^{(0)} \in \Theta$ is given. The *holonomic gradient algorithm* evaluates $\boldsymbol{g}(\boldsymbol{\alpha}^{(1)})$ at any other point $\boldsymbol{\alpha}^{(1)}$. Here the term gradient refers to the gradient of $\boldsymbol{g}(\boldsymbol{\alpha})$.

Let $\bar{\boldsymbol{\alpha}}(\tau)$, $\tau \in [0, 1]$, be a smooth curve in Θ such that $\bar{\boldsymbol{\alpha}}(0) = \boldsymbol{\alpha}^{(0)}$ and $\bar{\boldsymbol{\alpha}}(1) = \boldsymbol{\alpha}^{(1)}$. Denote $\bar{\boldsymbol{g}}(\tau) = \boldsymbol{g}(\bar{\boldsymbol{\alpha}}(\tau))$. Then, it is easily shown that $\bar{\boldsymbol{g}}(\tau)$ is the solution of the ODE

$$\frac{d}{d\tau}\bar{\boldsymbol{g}}(\tau) = \mathbf{K}(\tau)\bar{\boldsymbol{g}}(\tau) \tag{5}$$

where

$$\mathbf{K}(\tau) = \sum_{i=1}^{d} \frac{d\bar{\alpha}_i(\tau)}{d\tau} \boldsymbol{P}_i(\bar{\boldsymbol{\alpha}}(\tau)) \quad \bar{\boldsymbol{g}}(0) = \boldsymbol{g}(\boldsymbol{\alpha}^{(0)})$$

In particular, $\bar{\boldsymbol{g}}(1) = \boldsymbol{g}(\boldsymbol{\alpha}^{(1)})$.

A natural choice of $\bar{\boldsymbol{\alpha}}(\tau)$ is the segment $\bar{\boldsymbol{\alpha}}(\tau) = (1 - \tau)\boldsymbol{\alpha}^{(0)} + \tau\boldsymbol{\alpha}^{(1)}$ connecting $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\alpha}^{(1)}$ with the constant derivative vector $\frac{d\bar{\alpha}_i(\tau)}{d\tau} = \boldsymbol{\alpha}_i^{(1)} - \boldsymbol{\alpha}_i^{(0)}$. The holonomic gradient algorithm is described as follows:

Input $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{g}(\boldsymbol{\alpha}^{(0)})$, $\boldsymbol{\alpha}^{(1)}$ and a sufficiently small number $\delta > 0$.
Output $\boldsymbol{g}(\boldsymbol{\alpha}^{(1)})$.
Algorithm

1. Solve the ODE (5) over $\tau \in [0, 1]$ numerically by a Runge–Kutta method so that the solution is attained within a required accuracy.
2. Return $\bar{\boldsymbol{g}}(1)$.

Note that the standard numerical routines for solving (5) are highly accurate and available in most computer packages. More specifically, the rk function in the deSolve package of R provides the required solution for a given accuracy.

As shown later in Sect. 5, the holonomic gradient method is used for maximum likelihood estimation via some gradient descent scheme, where the orthogonal matrix $\boldsymbol{O}$ can somehow be treated independently from the normalizing constant. As a result, we only need the Pfaffian equations for diagonal covariance matrices when the corresponding ODE has dimension $2l$.

4 Explicit Pfaffians and HGM for Fisher Bingham

The parameters of Sect. 2.2 for the most general Fisher–Bingham case are $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$, i.e. $\dim(\Theta) = 2l$ where l is the number of distinct values of θ_i. Using properties (2), we can assume here that θ_i and γ_i are allowed to vary freely as positive values, while the smallest entry of $\boldsymbol{\theta}$ can be fixed to 0. As a direct consequence of differentiating (1) and the fact that $\sum_{i=1}^{p} x_i^2 = 1$,

$$\sum_{i=1}^{l} \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i} = -\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}). \tag{6}$$

This equation implies that partial derivatives $\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i}$ are sufficient for evaluating $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ where the vector $\boldsymbol{g}$ has length $r = 2l$ and is defined as

$$\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \left(\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_1} \cdots \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_l} \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_1} \cdots \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_l} \right) \tag{7}$$

where the first-order partial derivatives above are easily seen from (1) or (3), to depend on $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ as

$$\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i} = -\int_{i\mathbb{R}+t_0} \left(\frac{n_i}{2(\theta_i + t)} + \frac{\gamma_i^2}{4(\theta_i + t)^2} \right) \mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\theta}) e^t \, dt \tag{8}$$

and for $\gamma_i \neq 0$

$$\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i} = \int_{\mathrm{i}\mathbb{R}+t_0} \frac{\gamma_i}{2(\theta_i + t)} \mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\theta}) e^t \mathrm{d}t. \tag{9}$$

In this case, the corresponding ODE as in (5) is seeking the solution of some vector curve $\boldsymbol{g}(\boldsymbol{\alpha})$ of dimension $2l$ and the required normalizing constant is simply minus the sum of the components of this vector as in (6). The left side of the Pfaffian equations (4) is clearly $\frac{\partial \boldsymbol{g}}{\partial \theta_i}$ and $\frac{\partial \boldsymbol{g}}{\partial \gamma_i}$, which are actually the second-order derivatives $\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i \partial \gamma_j}$. In other words, the Pfaffian equations (4) are stating identities such that these second-order derivatives of $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ are linearly dependent on the first-order ones and Pfaffian entries. Therefore, in order to establish explicitly the Pfaffian equations for $\boldsymbol{g}$ we need to consider such particular relationships between the first- and second- order derivatives of $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. They are stated in the following theorem:

Theorem 1 *If $\theta_i \neq \theta_j$ and $\gamma_i \neq 0 \neq \gamma_j$, the Pfaffian equations (4) for the general Fisher–Bingham distribution are generated by*

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i \partial \theta_j} = -\begin{pmatrix} \frac{n_j}{2(\theta_j - \theta_i)} + \frac{\gamma_j^2}{4(\theta_j - \theta_i)^2} \\ \frac{n_i}{2(\theta_i - \theta_j)} + \frac{\gamma_i^2}{4(\theta_i - \theta_j)^2} \\ \frac{n_j \gamma_i}{4(\theta_j - \theta_i)^2} + \frac{\gamma_i \gamma_j^2}{4(\theta_j - \theta_i)^3} \\ \frac{n_i \gamma_j}{4(\theta_i - \theta_j)^2} + \frac{\gamma_i^2 \gamma_j}{4(\theta_i - \theta_j)^3} \end{pmatrix}^T \begin{pmatrix} \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i} \\ \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_j} \\ \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i} \\ \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_j} \end{pmatrix} \tag{10}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_j \partial \gamma_i} = \begin{pmatrix} \frac{\gamma_i}{2(\theta_i - \theta_j)} \\ -\frac{n_j}{2(\theta_j - \theta_i)} - \frac{\gamma_j^2}{4(\theta_j - \theta_i)^2} \\ \frac{\gamma_i \gamma_j}{4(\theta_i - \theta_j)^2} \end{pmatrix}^T \begin{pmatrix} \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_j} \\ \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i} \\ \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_j} \end{pmatrix} \tag{11}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i \partial \gamma_j} = \begin{pmatrix} \frac{\gamma_j}{2(\theta_j - \theta_i)} \\ -\frac{\gamma_i}{2(\theta_j - \theta_i)} \end{pmatrix}^T \begin{pmatrix} \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i} \\ \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_j} \end{pmatrix} \tag{12}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i \partial \theta_i} = -\sum_{i \neq j = 1}^{l} \frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_j \partial \gamma_i} - \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i} \tag{13}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial^2 \theta_i} = -\sum_{i \neq j = 1}^{l} \frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i \partial \theta_j} - \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i} \tag{14}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial^2 \gamma_i} = -\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i} - \frac{n_i - 1}{\gamma_i} \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i} \tag{15}$$

where $\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_j \partial \gamma_i}$ and $\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i \partial \theta_j}$ in (13) and (14) can be given in terms of first-order derivatives using (11) and (10).

The proofs of these identities which are in "Appendix" rely on results from partial fractions.

Note also that as the Pfaffian matrices are defined in terms of the pairwise differences $\theta_i - \theta_j$, one can easily see that the ODE solution for $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ satisfies properties (2).

As a corollary to the theorem, the differential equation for the well-known von Mises–Fisher distribution is derived from (6) and (15) as:

$$\frac{\partial^2 \mathcal{C}(0, \gamma_1)}{\partial^2 \gamma_1} + \frac{p-1}{\gamma_1} \frac{\partial \mathcal{C}(0, \gamma_1)}{\partial \gamma_1} - \mathcal{C}(0, \gamma_1) = 0.$$

where $l = 1$, $n_1 = p$ and $\boldsymbol{\theta} = \mathbf{0}$. The expression $\gamma_1^{\frac{p}{2}-1} \mathcal{C}(0, \gamma_1)$ satisfies equation 9.6.1 in Abramowitz and Stegun (1972) for the modified Bessel functions and is consistent with the known expression for these cases (see 9.3.4 in Mardia and Jupp 2000).

Two types of Pfaffians

The Pfaffian matrices will be of two types: $\mathbf{P}_i$ and $\mathbf{P}_{i+l}$ for $i = 1, 2, \ldots, l$ since the vector $\boldsymbol{g}$ in (7) with parameters $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$ implies $\boldsymbol{g}_i = \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \theta_i}$ and $\boldsymbol{g}_{i+l} = \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \gamma_i}$. Each $\mathbf{P}_i$ or $\mathbf{P}_{i+l}$ will be of dimension $2l \times 2l$, with all but two rows having at most 4 nonzero entries. The explicit expressions are found in "Appendix"

For a curve $\bar{\boldsymbol{\alpha}}$ with constant derivative, the matrix function $\mathbf{K}$ of (5) will be a linear combination of the $2l$ Pfaffian matrices.

This implies that for situations where some θ_i and θ_j coalesce, the matrix $\mathbf{K}$ will have intolerably large entries due to the presence of $\frac{1}{(\theta_j - \theta_i)^r}$ for $r = 1, 2, 3$ in the Pfaffian matrices. In these cases, stiffness in the corresponding ODE could appear. These situations are generally addressed by reparametrizing or changing the integrating curve

$\bar{\boldsymbol{\alpha}}(\tau)$ along which $\mathbf{K}$ remains manageable. For example, the choose of integrating path along some radial direction as suggested in Koyama et al. (2014) and Koyama and Takemura (2016) seems to work well. The default setting in our implementation is based on the same path so that

$$\bar{\theta}_i(\tau) = \tau \theta_i \quad \bar{\gamma}_i(\tau) = \sqrt{\tau} \gamma_i$$

starting from a small τ_0 so that $\boldsymbol{g}(\boldsymbol{\alpha}(\tau_0))$ is accurately evaluated as a starting point for the ODE. For example, using the curve above for a choice of close entries for $\boldsymbol{\theta} = (1, 2.9999, 3, 3.0001)$ and $\boldsymbol{\gamma} = (1, 1, 1, 1)$ the method works well by providing within 0.53 seconds a value for $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = 2.9753553$. Note that the saddlepoint approximation provides the value 2.942742 in 0.001 seconds. This is not surprising since, while SPA is very fast, the proposed method relies on a potentially computationally expensive step of evaluating the starting value of $\boldsymbol{g}$ at a sufficiently small τ_0 so that to guarantee the required accuracy at the target value $\tau = 1$. In general, our method could require a careful choice of both

the starting values and the integration path so that the ODE is does not have numerical problems. However, our implementation with the radial curve described as not failed in the examples that we have considered.

One can easily see that the Pfaffian values do not become degenerate even if all γ_i become zero (except cases when $n_i > 1$); therefore, a possible starting value for carrying out the numerical evaluation for general $\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ or $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ could be the corresponding derivatives of the Bingham normalizing constant at $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\gamma} = \mathbf{0})$ (evaluated as in Sei and Kume 2015), and then stemming from this point in R^{2l}, a second integration curve can be defined ending at the required $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. In cases of $n_i > 1$, we can use the power series derived by Kume and Walker (2009) and Koyama et al. (2014) as a starting value of $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\gamma})$.

5 MLE optimization using the gradient approach

If the observed data are collected in a matrix $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ of dimension $p \times n$, such that $\mathbf{A} = \frac{\sum_{i=1}^n x_i x_i^\top}{n}$ and $\mathbf{B} = \frac{\sum_{i=1}^n x_i}{n}$, the corresponding likelihood function is

$$\begin{aligned}
&\log \mathcal{L}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\mu, \mathbf{X}\right) \\
&= -n \log \mathcal{C}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\mu\right) \\
&\quad - \sum_{i=1}^{n}\left(x_i^\top \frac{\boldsymbol{\Sigma}^{-1}}{2} x_i - x_i^\top \boldsymbol{\Sigma}^{-1}\mu\right) \\
&= -n \log \mathcal{C}\left(\frac{\boldsymbol{\Delta}^{-1}}{2}, \boldsymbol{\gamma}\right) - ntr\left(\mathbf{A}\mathbf{O}^\top \frac{\boldsymbol{\Delta}^{-1}}{2}\mathbf{O} - \mathbf{O}\mathbf{B}\boldsymbol{\gamma}^\top\right) \\
&= -n\left(\log \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) + tr(\mathbf{A}\mathbf{O}^\top \mathrm{diag}(\boldsymbol{\theta})\mathbf{O} + \mathbf{O}\mathbf{B}\boldsymbol{\gamma}^\top)\right)
\end{aligned}$$

with $\boldsymbol{\Sigma}^{-1} = \mathbf{O}^\top \boldsymbol{\Delta}^{-1}\mathbf{O}$, $\mathcal{C}(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\mu) = \mathcal{C}(\frac{\boldsymbol{\Delta}^{-1}}{2}, \boldsymbol{\gamma})$, $\boldsymbol{\gamma} = \boldsymbol{\Delta}^{-1}\mathbf{O}\boldsymbol{\mu}$ and $\frac{\boldsymbol{\Delta}^{-1}}{2} = \mathrm{diag}(\boldsymbol{\theta})$, while without loss of generality one can replace $\boldsymbol{O}$ with $-\boldsymbol{O}$ as this does not affect $\mathbf{O}^\top \boldsymbol{\Delta}^{-1}\mathbf{O}$ but switches the sign of $\mathbf{O}\mathbf{B}\boldsymbol{\gamma}^\top$.

Therefore, maximizing $\log \mathcal{L}(\frac{\boldsymbol{\Sigma}^{-1}}{2}, \boldsymbol{\Sigma}^{-1}\mu, \mathbf{X})$ is equivalent to minimizing

$$\log \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}) + tr(\mathbf{A}\mathbf{O}^\top \mathrm{diag}(\boldsymbol{\theta})\mathbf{O} + \mathbf{O}\mathbf{B}\boldsymbol{\gamma}^\top). \tag{16}$$

Since values of $\boldsymbol{\theta}$ can be shifted so that its smallest value becomes 0 and $\mathbf{O}$ can allow for $\boldsymbol{\gamma}$ to have non-negative entries, we can optimize (16) on $\boldsymbol{\theta} \geq 0$ with $\min(\boldsymbol{\theta}) = 0$ and $\boldsymbol{\gamma} \geq 0$ by iteratively updating the parameters which increase the likelihood value such that:

1. for a fixed $\mathbf{O}$, consider the optimization problem on $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ which is performed in $2l$ dimensions including the ODE for the HGM implementation.
2. by keeping these values $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ fixed, we can then find the optimal $\mathbf{O}$ by minimizing or decreasing only the quadratic part of the likelihood $tr(\mathbf{A}\mathbf{O}^\top \mathrm{diag}(\boldsymbol{\theta})\mathbf{O} + \mathbf{O}\mathbf{B}\boldsymbol{\gamma}^\top)$.

In order to establish a gradient descent approach for the first step, we only need the partial derivatives of $\log L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{O})$ as follows:

$$\frac{\partial \log L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{O})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}} \frac{1}{\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})} + \mathrm{diag}(\mathbf{O}\mathbf{A}\mathbf{O}^\top) \tag{17}$$

$$\frac{\partial \log L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{O})}{\partial \boldsymbol{\gamma}} = \frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \frac{1}{\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})} + \mathbf{B}^\top \mathbf{O}^\top \tag{18}$$

where $\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\theta}} \frac{1}{\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}$ and $\frac{\partial \mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \frac{1}{\mathcal{C}(\boldsymbol{\theta}, \boldsymbol{\gamma})}$ are the output of our holonomic gradient algorithm implementation for the Pfaffian equations shown earlier. In its general form, the second optimization needs special care as it is a non standard optimization problem in $\mathcal{O}(p)$. We show below an adopted gradient method which addresses this problem and therefore completes the MLE optimization. In fact, two special cases that do not require our optimization in $\mathcal{O}(p)$ are:

- Bingham distribution, i.e. $\boldsymbol{\gamma} = \mathbf{0}$, here the orthogonal component of the SVD decomposition of $\mathbf{A}$ is optimal
- Kent distributions for $p = 3$ where approximate MLE is used and the problem is conveniently reduced to an optimization in $\mathcal{O}(2)$ after the third column vector of $\mathbf{O}$ is chosen independently such that the 3-dimensional vector $\mathbf{B}$ coincides with a fixed axis (see Kent 1982, Sect. 4).

5.1 Optimization in O

In particular, we need to find the optimal $\hat{\mathbf{O}}$ such that

$$\hat{\mathbf{O}} = \underset{\mathbf{O} \in \mathcal{O}(p)}{\mathrm{argmin}}\, tr(\mathbf{A}\mathbf{O}^\top \mathrm{diag}(\boldsymbol{\theta})\mathbf{O} + \mathbf{O}\mathbf{B}\boldsymbol{\gamma}^\top)$$

In fact, this problem is equivalent to

$$\hat{\mathbf{O}} = \underset{\mathbf{O} \in \mathcal{O}(p)}{\mathrm{argmin}} \left\| \mathrm{diag}(\sqrt{\boldsymbol{\theta}})\mathbf{O}\mathbf{A}^{1/2} + \mathbf{A}^{-1/2}\mathbf{B}\boldsymbol{\gamma}^\top \mathrm{diag}\left(\frac{1}{\sqrt{\boldsymbol{\theta}}}\right) \right\|^2$$

This is the weighted Procrustes optimization problem considered in Chu and Trendafilov (1998). The authors there adopt an ODE approach to this problem as a simple adaption of continuous gradient optimization. We show in the following the gradient descent version in discrete time which can be immediately implemented within a unified MLE

optimization procedure for the Fisher–Bingham family of distributions. Note that, provided we allow in the likelihood optimization the sign of one of the components in $\boldsymbol{\gamma}$ to vary, the optimal matrix $\mathbf{O}$ can be allowed to be a rotation matrix, i.e. $\mathbf{O} = e^{\boldsymbol{v}}$ where $\boldsymbol{v}$ is skew symmetric, i.e. $\boldsymbol{v}+\boldsymbol{v}^\top = 0$.

Proposition 1 *A necessary condition for* $\mathbf{O}$ *to be an optimal orthogonal matrix is that*

$$\mathcal{A} = \text{diag}(\boldsymbol{\theta})\mathbf{O}\mathbf{A}\mathbf{O}^\top - \mathbf{O}\mathbf{A}\mathbf{O}^\top\text{diag}(\boldsymbol{\theta}) + \boldsymbol{\gamma}\mathbf{B}^\top\mathbf{O}^\top$$

is symmetric.

Proof (see "Appendix"). □

In our case however, we can implement the gradient approach in the orthogonal group by taking as a possible new update for $\mathbf{O}$ some rotation along the curve

$$\mathbf{O}e^{\hat{\boldsymbol{v}}t} \quad \text{where} \quad \hat{\boldsymbol{v}} = \mathcal{A} - \mathcal{A}^\top \tag{19}$$

Clearly, this curve reduces to a single point only if $\mathcal{A}$ is symmetric, i.e. $\mathbf{O}$ is a critical point. We can use this fact as a stopping criterion in our gradient optimization. We proceed in a similar way to obtain the second derivative, and it can be shown that a necessary condition that a particular critical $\mathbf{O}$ is a local minimum is

$$tr(\mathcal{A} + 2\mathbf{O}\mathbf{A}\mathbf{O}^\top\text{diag}(\boldsymbol{\theta}))(\boldsymbol{v}^2)^\top + 2tr(\mathbf{A}\mathbf{O}^\top\boldsymbol{v}^\top\text{diag}(\boldsymbol{\theta})\boldsymbol{v}\mathbf{O}) \geq 0$$
$$\forall \boldsymbol{v} = -\boldsymbol{v}^\top$$

5.2 Algorithm for finding the MLE

We now have all the ingredients to establish our gradient approach for the MLE of the Fisher–Bingham distributions.

Algorithm

For a given initial set of estimates $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and $\mathbf{O}$, we perform the updates as follows:

1. $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \dfrac{\partial \log L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{O})}{\partial \boldsymbol{\theta}}\delta_{\boldsymbol{\theta}}$

   where $\frac{\partial \log L(\boldsymbol{\theta},\boldsymbol{\gamma},\mathbf{O})}{\partial \boldsymbol{\theta}}$ is as in (17) and $\delta_{\boldsymbol{\theta}}$ is a real number such that $\log L(\hat{\boldsymbol{\theta}}, \boldsymbol{\gamma}, \mathbf{O}) > \log L(\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{O})$.
2. $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma} + \dfrac{\partial \log L(\hat{\boldsymbol{\theta}}, \boldsymbol{\gamma}, \mathbf{O})}{\partial \boldsymbol{\gamma}}\delta_{\boldsymbol{\gamma}}$

   where $\frac{\partial \log L(\hat{\boldsymbol{\theta}},\boldsymbol{\gamma},\mathbf{O})}{\partial \boldsymbol{\gamma}}$ is as in (18) and $\delta_{\boldsymbol{\gamma}}$ is a real number such that $\log L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \mathbf{O}) > \log L(\hat{\boldsymbol{\theta}}, \boldsymbol{\gamma}, \mathbf{O})$.
3. $\hat{\mathbf{O}} = e^{\hat{\boldsymbol{v}}t_0}\mathbf{O}$

   where $\boldsymbol{v}$ is calculated as in (19) and t_0 is chosen such that $\log L(\hat{\boldsymbol{\theta}}, \hat{\gamma}, \hat{\mathbf{O}}) > \log L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \mathbf{O})$.
4. We stop when the derivatives in steps 1 and 2 and $\hat{\boldsymbol{v}}$ are practically zero.

Note, however, that if we wanted to fit the Bingham distribution, i.e. $\boldsymbol{\gamma}$ is assumed to be zero, then there is no need to implement step 3 above as the optimal $\mathbf{O}$ in this case is simply the one for which $\mathcal{A} = 0$ that is $\mathbf{O}\mathbf{A}\mathbf{O}^\top$ is diagonal.

5.3 Numerical evidence

In Nakayama et al. (2011), the authors illustrate the general methodology of holonomic gradient method by focusing on two data sets: one from the area of astronomy and the other one from the magnetism. We revisit the first data set in order to confirm that our method gives the same MLE results. We also want to make use of our different parametrization which deals with the sub-classes of Fisher–Bingham family to perform statistical inference to choose the most appropriate model. The second data set considered is previously used in the paper of Arnold and Jupp (2013) where a statistical model of orthogonal frames is introduced. Particular recordings of three orthogonal axis related to individual earthquake events in New Zealand are grouped in three data sets. Each triplet of orthogonal axes in $\mathbb{R}^3$ related to a particular earthquake event gives rise to a direction orthogonal to the horizontal plane. Observations of these directions can allow modelling by Bingham distributions c.f Arnold and Jupp (2013). So we have three classes of directional data where Bingham distributions are considered appropriate. In particular, a Bayesian modelling approach to fitting Bingham distributions to such data is also considered in Fallaize and Kypraios (2014). We will show below that in fact the best modelling choice among the sub-classes of Fisher–Bingham family is indeed the Bingham distribution.

Astronomy data

For this data set in our parametrization, the components A and B are as follows:

$$\mathbf{A} = \begin{pmatrix} 0.312 & 0.029 & 0.071 \\ 0.029 & 0.360 & 0.046 \\ 0.071 & 0.046 & 0.327 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0.006 \\ 0.005 \\ 0.076 \end{pmatrix}$$

Fitting the Fisher–Bingham distribution to these data, we get the following MLE values

$$\hat{\boldsymbol{\theta}}_{\text{FB}} = \begin{pmatrix} 0 \\ 0.708(0.576) \\ 1.416(1.469) \end{pmatrix} \quad \hat{\boldsymbol{\gamma}}_{\text{FB}} = \begin{pmatrix} 0.122(0.124) \\ 0.087(0.087) \\ 0.197(0.196) \end{pmatrix}$$

$$\hat{\mathbf{O}}_{\text{FB}} = \begin{pmatrix} -0.511(-0.510) & -0.612(-0.613) & -0.605(-0.604) \\ -0.490(-0.489) & 0.785(0.784) & -0.380(-0.383) \\ 0.706(0.708) & 0.102(0.100) & -0.700(-0.699) \end{pmatrix}$$

where the values in brackets are the MLE estimates using the saddlepoint approximation for the normalizing constant.

The optimal likelihood value rescaled by $-n$ as defined in (16) is $2.457746 = \log 11.67846$ which is same as the value reported in Nakayama et al. (2011). The corresponding quantity for the saddlepoint approximation is 2.463414. We fitted to this data set the Kent distribution and the MLE of the corresponding quantities using HGM (and saddlepoint approximation) are

$$\hat{\boldsymbol{\theta}}_K = \begin{pmatrix} 0 \\ -0.703(-0.783) \\ 0.703(0.783) \end{pmatrix} \quad \hat{\boldsymbol{\gamma}}_K = \begin{pmatrix} 0.099(0.098) \\ 0 \\ 0 \end{pmatrix}$$

$$\hat{\mathbf{O}}_K = \begin{pmatrix} -0.461(-0.463) & 0.774(0.775) & -0.435(-0.429) \\ 0.493(0.495) & 0.631(0.628) & 0.599(0.601) \\ -0.738(-0.735) & -0.062(-0.066) & 0.672(0.674) \end{pmatrix}$$

with 2.465478 and 2.471299 being the corresponding values of function (16) at these optimal points. Since the difference in the number of parameters between these models is $8-5 = 3$, we can apply the log-likelihood ratio test, under the null hypothesis of the Kent model

$$H_0 : 2n \left(\log L(\boldsymbol{\Theta}_K) - \log L(\boldsymbol{\Theta}_{\mathrm{FB}})\right) \sim \chi^2_3$$

where $\log L$ is (the rescaled log-likelihood by $-n$) defined in (16) and $\boldsymbol{\Theta}_{\mathrm{FB}}$ and $\boldsymbol{\Theta}_K$ represent the MLE estimates for the full Fisher–Bingham and Kent, respectively. The sample size is $n = 168$, and the value of likelihood ratio statistic is therefore $2 * 168 * (2.465478 - 2.457746) = 2.597952$ which suggests that there is not enough evidence supporting the full Fisher–Bingham distribution model here. The same conclusion holds for the saddlepoint approximation quantities.

Earthquake data

The three axes of interest for an earthquake event are the directions of compressional axis **P**; tensional axis **T**, while the null axis **A** is defined as $\mathbf{A} = \mathbf{P} \times \mathbf{T}$. For these data sets, the first two axes tend to be horizontal, and therefore, the third axis points vertically. These axial data are shown in Fig. 1 and are split into three groups of particular interest. The key assumption in modelling these data sets is that observed directions of axis **A** follow a Bingham distribution on the sphere of dimension 2.

The MLEs for the Bingham distributions parameters fitted to the three data sets of directions of axis **A** shown in Fig. 1 are as follows:

$$\hat{\boldsymbol{\theta}}_B = \begin{pmatrix} 0 \\ 5.059 \\ 3.804 \end{pmatrix} \quad \hat{\mathbf{O}}_B = \begin{pmatrix} 0.008 & 0.054 & -0.999 \\ 0.428 & 0.902 & 0.052 \\ 0.904 & -0.428 & -0.016 \end{pmatrix}$$

$$\hat{\boldsymbol{\theta}}_B = \begin{pmatrix} 0 \\ 5.094 \\ 2.941 \end{pmatrix} \quad \hat{\mathbf{O}}_B = \begin{pmatrix} 0.044 & 0.012 & -0.999 \\ 0.522 & 0.852 & 0.033 \\ 0.852 & -0.523 & 0.031 \end{pmatrix}$$

$$\hat{\boldsymbol{\theta}}_B = \begin{pmatrix} 0 \\ 0.784 \\ -1.025 \end{pmatrix} \quad \hat{\mathbf{O}}_B = \begin{pmatrix} 0.583 & 0.808 & -0.087 \\ -0.750 & 0.494 & -0.439 \\ -0.312 & 0.321 & 0.894 \end{pmatrix}$$

We also fitted Fisher–Bingham distributions to these three clusters of directions, and the corresponding values of the corresponding χ^2_3 test statistics and their p values are 0.4889717(0.9213075), 3.885764(0.2740667) and 1.630983 (0.6523852). These results suggest that the Bingham distribution assumption is reasonable for these data sets. One of the referees mentioned correctly that the Bingham distribution is in fact appropriate for axial data. This means that if the data points undergo independently some axial rearrangement, (namely independent sign changes to each individual coordinates), the likelihood will not change under Bingham but will do so for the Fisher–Bingham case. Therefore, the model choice that we perform here is to only illustrate numerically that our HGM implementation here works for a given axial arrangement of these data points, and alternative models like the matrix Fisher distributions as suggested by the referee could be better modelling strategies for these orthogonal frames.

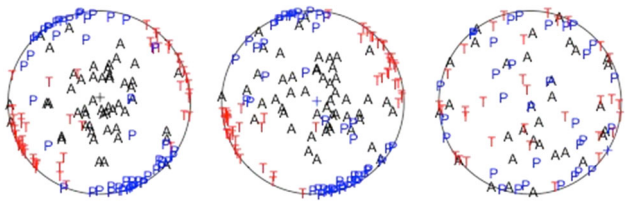

Fig. 1 The planar projections on the horizontal plane of the frame of orthogonal axes (**P A T**) related to earthquake records. The *left plot* shows the data for earthquakes in Christchurch prior to 22 February 2001, the *middle plot* those recorded post 22 February 2001, and the third plot refers to earthquake records in South Island. The point + in each *plot* denotes the mode of the Bingham distribution fitted to directions of axis **A**

6 Concluding remarks

In this paper, we provide explicitly the Pfaffian system for the normalizing constant of the Fisher–Bingham distributions including the degenerate cases. Such explicit expressions have not only theoretical interest but also improve on the implementation of the current methods used for the MLE of these models. We reduce the dimensionality of the ODE equation as we need to operate at a dimension not more than twice the number of distinct values of θ_i. The standard HGM so far does not account for multiplicities among θ_i's or $\gamma_i = 0$. We can also perform exact MLE inference by using gradient optimization methods for the optimal orthogonal component **O** as in weighted Procrustes optimization. Note, however, that optimization in **O** shown in Sect. 5.1 is

only local and rank$(\mathbf{B}^\top \boldsymbol{\gamma}) = 1$ might imply many optimal solutions. For the Bingham distribution, namely $\boldsymbol{\gamma} = 0$ case, the optimal matrix $\mathbf{O}$ does not depend on $\boldsymbol{\theta}$ as that is defined such that $\sum_{i=1}^n \mathbf{O}x_i(\mathbf{O}x_i)^\top$ is diagonal. The numerical examples indicate that when carefully implemented, the method is highly accurate and performs well in real applications. Its implementation can fail sometimes since the corresponding ODE does not perform well numerically. This can be addressed by changing the ODE, namely, by altering either the starting point and/or the integrating path as discussed in the last paragraph of Sect. 4. The default choice of the curve which is used in our implementation in R works well in many tests. As indicated in our first real data example, the MLE using the saddle point approximation is with some exceptions, not far from the our MLE. One can start the HGM from this solution. This hybrid approach could in principle reduce the regions of the numerical search and could be seen as a way of calibrating the saddle point approximation. Our proposed method clearly generalizes that given in Koyama et al. (2014) since it offers explicit expressions for the Pfaffian equations for all Fisher–Bingham distributions including those with degeneracies in the parameters. Finally, since the saddle point approximation method is numerically stable, practically accurate and immediately available, the HGM could be used as a refinement to this approximation.

Acknowledgements The authors are very grateful to Richard Arnold and Peter Jupp for providing the earthquake data and the two anonymous referees for their helpful comments. This work was partially supported by JSPS KAKENHI Grant No. JP26108003 and JP26540013. Our special thanks go to Andy Wood for general discussions and encouragement.

7 Appendix

The results of Theorem 1 rely heavily on Lemma 1 which is stated after some initial remarks.

Remark 1 One can easily notice that

$$\int_{\mathrm{i}\mathbb{R}+t_0} \frac{1}{(\theta_i+t)}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t = \frac{2}{\gamma_i}\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \gamma_i}$$

and

$$\int_{\mathrm{i}\mathbb{R}+t_0} \frac{1}{(\theta_i+t)^2}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t = -\frac{4n_i}{\gamma_i^3}\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \gamma_i} - \frac{4}{\gamma_i^2}\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \theta_i},$$

and therefore, these elementary functions

$$\int_{\mathrm{i}\mathbb{R}+t_0} \frac{1}{(\theta_i+t)^r}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \quad r = 1, 2$$

are actually representing the first-order derivatives $\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \theta_i}$ and $\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \gamma_i}$.

In what follows, we will show that based on the theory of partial fractions, $\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \theta_i}$ and $\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial \gamma_i}$ can be used to express the integrals above even for $r = 3, 4$ which will then derive the expressions for the second-order derivatives. This is the basis of the following methodology for obtaining the Pfaffian equations.

For example, using (8) and (9) the second-order derivatives generate these expressions:

For $i \neq j$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i\partial\theta_j} = \int_{\mathrm{i}\mathbb{R}+t_0} \left(\frac{n_i}{2(\theta_i+t)} + \frac{\gamma_i^2}{4(\theta_i+t)^2}\right) \times \left(\frac{n_j}{2(\theta_j+t)} + \frac{\gamma_j^2}{4(\theta_j+t)^2}\right)\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \tag{20}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\gamma_j} = \int_{\mathrm{i}\mathbb{R}+t_0} \frac{\gamma_i\gamma_j}{4(\theta_i+t)(\theta_j+t)}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \tag{21}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\theta_j} = \int_{\mathrm{i}\mathbb{R}+t_0} \left(\frac{n_j}{2(\theta_j+t)} + \frac{\gamma_j^2}{4(\theta_j+t)^2}\right) \times \frac{-\gamma_i}{2(\theta_i+t)}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \tag{22}$$

and for $i = j$,

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial^2\theta_i} = \int_{\mathrm{i}\mathbb{R}+t_0} \left(\frac{2n_i+n_i^2}{4(\theta_i+t)^2} + \frac{(2+n_i)\gamma_i^2}{4(\theta_i+t)^3} + \frac{\gamma_i^4}{16(\theta_i+t)^4}\right) \times \mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \tag{23}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial^2\gamma_i} = -\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i} - \frac{n_i-1}{\gamma_i}\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \tag{24}$$

$$\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\theta_i} = \int_{\mathrm{i}\mathbb{R}+t_0} -\left(\frac{(n_i+2)\gamma_i}{4(\theta_i+t)^2} + \frac{\gamma_i^3}{8(\theta_i+t)^3}\right) \times \mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \tag{25}$$

Remark 2 If $i = j$ it is clear, however, that nonzero terms γ_i give rise to

$$\int_{\mathrm{i}\mathbb{R}+t_0} \frac{1}{(\theta_i+t)^r}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t \quad r = 3, 4$$

in the second-order derivatives, while for $\gamma_i = 0$ such terms vanish.

Remark 3 The second-order derivatives $\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i\partial\theta_j}$, $\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\gamma_j}$ and $\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\theta_j}$ for $i \neq j$ can be given in terms of only this pair of basis functions

$$\int\limits_{\mathrm{i}\mathbb{R}+t_0} \frac{1}{(\theta_i+t)^r}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t \quad r=1,2$$

which from Remark 1 are obtained in terms of $\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i}$ and $\frac{\partial \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}$. This is easily seen if applying the integration with respect to $\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t$ and by using the following basis decomposition:

Lemma 1 *If $\theta_i \neq \theta_j$ and $\gamma_i \neq 0 \neq \gamma_j$, then*

$$\begin{aligned}&\left(\frac{A}{\theta_i+t}+\frac{B}{(\theta_i+t)^2}\right)\left(\frac{C}{\theta_j+t}+\frac{D}{(\theta_j+t)^2}\right)\\&\quad=\frac{a}{\theta_i+t}+\frac{b}{(\theta_i+t)^2}+\frac{c}{\theta_j+t}+\frac{d}{(\theta_j+t)^2}\end{aligned}\tag{26}$$

and

$$\begin{aligned}&\int\limits_{i\mathbb{R}+t_0}\left(\frac{A}{\theta_i+t}+\frac{B}{(\theta_i+t)^2}\right)\left(\frac{C}{\theta_j+t}+\frac{D}{(\theta_j+t)^2}\right)\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t\\&\quad=-\frac{4b}{\gamma_i^2}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i}-\frac{4d}{\gamma_j^2}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j}+\\&\qquad+\left(a\frac{2}{\gamma_i}-b\frac{4n_i}{\gamma_i^3}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}+\left(c\frac{2}{\gamma_j}-d\frac{4n_j}{\gamma_j^3}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j}\end{aligned}$$

where

$$\begin{aligned}b&=\frac{B\left(C(\theta_j-\theta_i)+D\right)}{(\theta_j-\theta_i)^2}=\frac{BC}{\theta_j-\theta_i}+\frac{BD}{(\theta_j-\theta_i)^2}\\d&=\frac{D\left(A(\theta_i-\theta_j)+B\right)}{(\theta_i-\theta_j)^2}=\frac{AD}{\theta_i-\theta_j}+\frac{DB}{(\theta_i-\theta_j)^2}\\a&=\frac{A\left(C(\theta_j-\theta_i)+D\right)+CB-2b(\theta_j-\theta_i)}{(\theta_j-\theta_i)^2}\\&=\frac{AC}{\theta_j-\theta_i}+\frac{AD-BC}{(\theta_j-\theta_i)^2}-2\frac{BD}{(\theta_j-\theta_i)^3}\\&=\frac{A}{B}b+\frac{BC}{(\theta_j-\theta_i)^2}-\frac{2b}{(\theta_j-\theta_i)}\end{aligned}\tag{27}$$

and

$$\begin{aligned}c&=\frac{C\left(A(\theta_i-\theta_j)+B\right)+AD-2d(\theta_i-\theta_j)}{(\theta_i-\theta_j)^2}\\&=\frac{AC}{\theta_i-\theta_j}+\frac{BC-AD}{(\theta_i-\theta_j)^2}-2\frac{BD}{(\theta_i-\theta_j)^3}\\&=\frac{C}{D}d+\frac{AD}{(\theta_i-\theta_j)^2}-\frac{2d}{(\theta_i-\theta_j)},\end{aligned}\tag{28}$$

i.e.

$$a=-c$$

Note that expressions on the right-hand side for a and b in the statement of Lemma are valid if $B \neq 0 \neq D$.

Proof of Lemma 1 The identity (26) is a direct consequence of the theory of partial fractions. From (26), we see that

$$\begin{aligned}&\frac{A(\theta_i+t)+B}{(\theta_i+t)^2}\frac{C(\theta_j+t)+D}{(\theta_j+t)^2}\\&=\frac{a(\theta_i+t)(\theta_j+t)^2+b(\theta_j+t)^2+c(\theta_i+t)^2(\theta_j+t)+d(\theta_i+t)^2}{(\theta_i+t)^2(\theta_j+t)^2}\end{aligned}$$

or

$$\begin{aligned}&(A(\theta_i+t)+B)\left(C(\theta_j+t)+D\right)\\&\quad=a(\theta_i+t)(\theta_j+t)^2+b(\theta_j+t)^2\\&\qquad+c(\theta_i+t)^2(\theta_j+t)+d(\theta_i+t)^2\end{aligned}\tag{29}$$

and, applying this equation for $t=-\theta_i$, $t=-\theta_j$, we have

$$\begin{aligned}b(\theta_j-\theta_i)^2&=B\left(C(\theta_j-\theta_i)+D\right)\\d(\theta_i-\theta_j)^2&=D\left(A(\theta_i-\theta_j)+B\right)\end{aligned}$$

which establish the explicit expressions for b and d. After differentiating with respect to t both sides of (29) and then substituting $t=-\theta_i$, $t=-\theta_j$ consecutively, we have the following pair of equations

$$\begin{aligned}-a(\theta_j-\theta_i)^2-2b(\theta_j-\theta_i)&=-A\left(C(\theta_j-\theta_i)+D\right)-CB\\-c(\theta_i-\theta_j)^2-2d(\theta_i-\theta_j)&=-C\left(A(\theta_i-\theta_j)+B\right)-AD\end{aligned}$$

which confirm the remaining expressions for a and b of the lemma including the identity $a=-c$.

The second result of the lemma is direct consequence of the first, while $\int\limits_{\mathrm{i}\mathbb{R}+t_0}\frac{1}{(\theta_i+t)}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t=\frac{2}{\gamma_i}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}$ and $\int\limits_{\mathrm{i}\mathbb{R}+t_0}\frac{1}{(\theta_i+t)^2}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t=-\frac{4n_i}{\gamma_i^3}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}-\frac{4}{\gamma_i^2}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i}$ □

Proof of Theorem 1 Applying Lemma 1 to Eqs. (20), (21) and (22), we obtain the following three identities:

$$\begin{aligned}
\frac{\partial^2 \mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i\partial\theta_j} &= \int_{\mathrm{i}\mathbb{R}+t_0} \left(\frac{n_i}{2(\theta_i+t)}+\frac{\gamma_i^2}{4(\theta_i+t)^2}\right) \\
&\quad\times\left(\frac{n_j}{2(\theta_j+t)}+\frac{\gamma_j^2}{4(\theta_j+t)^2}\right)\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t \\
&= -\left(\frac{n_j}{2(\theta_j-\theta_i)}+\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i} \\
&\quad-\left(\frac{n_i}{2(\theta_i-\theta_j)}+\frac{\gamma_i^2}{4(\theta_i-\theta_j)^2}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j} \\
&\quad-\left(\frac{n_j\gamma_i}{4(\theta_j-\theta_i)^2}+\frac{\gamma_i\gamma_j^2}{4(\theta_j-\theta_i)^3}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \\
&\quad-\left(\frac{n_i\gamma_j}{4(\theta_i-\theta_j)^2}+\frac{\gamma_i^2\gamma_j}{4(\theta_i-\theta_j)^3}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \qquad (30)
\end{aligned}$$

$$= -\begin{pmatrix} \frac{n_j}{2(\theta_j-\theta_i)}+\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2} \\ \frac{n_i}{2(\theta_i-\theta_j)}+\frac{\gamma_i^2}{4(\theta_i-\theta_j)^2} \\ \frac{n_j\gamma_i}{4(\theta_j-\theta_i)^2}+\frac{\gamma_i\gamma_j^2}{4(\theta_j-\theta_i)^3} \\ \frac{n_i\gamma_j}{4(\theta_i-\theta_j)^2}+\frac{\gamma_i^2\gamma_j}{4(\theta_i-\theta_j)^3} \end{pmatrix}^T \begin{pmatrix} \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i} \\ \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j} \\ \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \\ \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \end{pmatrix} \qquad (31)$$

since the corresponding terms are

$$a = \frac{n_i n_j}{4(\theta_j-\theta_i)}+\frac{n_i\gamma_j^2-n_j\gamma_i^2}{8(\theta_j-\theta_i)^2}-\frac{\gamma_i^2\gamma_j^2}{8(\theta_j-\theta_i)^3}$$

with

$$b = \frac{n_j\gamma_i^2}{8(\theta_j-\theta_i)}+\frac{\gamma_i^2\gamma_j^2}{16(\theta_j-\theta_i)^2}$$

and

$$c = -a \quad \text{with} \quad d = \frac{n_i\gamma_j^2}{8(\theta_i-\theta_j)}+\frac{\gamma_i^2\gamma_j^2}{16(\theta_i-\theta_j)^2}$$

$$\begin{aligned}
\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j\partial\gamma_i} &= -\int_{\mathrm{i}\mathbb{R}+t_0}\frac{\gamma_i}{2(\theta_i+t)} \\
&\quad\times\left(\frac{n_j}{2(\theta_j+t)}+\frac{\gamma_j^2}{4(\theta_j+t)^2}\right)\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t \\
&= \frac{\gamma_i}{2(\theta_i-\theta_j)}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j}-\left(\frac{n_j}{2(\theta_j-\theta_i)}+\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2}\right) \\
&\quad\times\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \\
&\quad+\frac{\gamma_i\gamma_j}{4(\theta_i-\theta_j)^2}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \qquad (32)
\end{aligned}$$

$$= \begin{pmatrix} \frac{\gamma_i}{2(\theta_i-\theta_j)} \\ -\frac{n_j}{2(\theta_j-\theta_i)}-\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2} \\ \frac{\gamma_i\gamma_j}{4(\theta_i-\theta_j)^2} \end{pmatrix}^T \begin{pmatrix} \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j} \\ \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \\ \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \end{pmatrix} \qquad (33)$$

since

$$b = 0 \quad d = -\frac{\gamma_i\gamma_j^2}{8(\theta_i-\theta_j)}$$

and

$$a = -c = -\frac{n_j\gamma_i}{4(\theta_j-\theta_i)}-\frac{\gamma_i\gamma_j^2}{8(\theta_j-\theta_i)^2}$$

$$\begin{aligned}
\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\gamma_j} &= \int_{\mathrm{i}\mathbb{R}+t_0}\frac{\gamma_i\gamma_j}{4(\theta_i+t)(\theta_j+t)}\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\mathrm{d}t \quad i\neq j \\
&= \frac{\gamma_j}{2(\theta_j-\theta_i)}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}-\frac{\gamma_i}{2(\theta_j-\theta_i)}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \qquad (34) \\
&= \begin{pmatrix} \frac{\gamma_j}{2(\theta_j-\theta_i)} \\ -\frac{\gamma_i}{2(\theta_j-\theta_i)} \end{pmatrix}^T \begin{pmatrix} \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \\ \frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \end{pmatrix} \qquad (35)
\end{aligned}$$

The corresponding cases of $i = j$, are obtained after applying $\frac{\partial}{\partial\gamma_i}$ on both sides of (6) and separating the term $\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i\partial\gamma_i}$ while using (32)

$$\begin{aligned}
\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i\partial\theta_i} &= -\sum_{i\neq j=1}^{l}\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j\partial\gamma_i}-\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i} \\
&= -\sum_{i\neq j=1}^{l}\frac{\gamma_i}{2(\theta_i-\theta_j)}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j} \\
&\quad-\sum_{i\neq j=1}^{l}\frac{\gamma_i\gamma_j}{4(\theta_i-\theta_j)^2}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j} \\
&\quad+\left(\sum_{i\neq j=1}^{l}\left(\frac{n_j}{2(\theta_j-\theta_i)}+\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2}\right)-1\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}
\end{aligned}$$

Similarly, after applying $\frac{\partial}{\partial\theta_j}$ on both sides of (6) and using (31) we have

$$\begin{aligned}
\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial^2\theta_i} &= -\sum_{i\neq j=1}^{l}\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i\partial\theta_j}-\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i} \\
&= \left(\sum_{i\neq j=1}^{l}\left(\frac{n_j}{2(\theta_j-\theta_i)}+\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2}\right)-1\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i}
\end{aligned}$$

$$+\sum_{i\neq j=1}^{l}\left(\frac{n_i}{2(\theta_i-\theta_j)}+\frac{\gamma_i^2}{4(\theta_i-\theta_j)^2}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_j}$$
$$+\sum_{i\neq j=1}^{l}\left(\frac{\gamma_i n_j}{4(\theta_j-\theta_i)^2}+\frac{\gamma_i\gamma_j^2}{4(\theta_j-\theta_i)^3}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i}$$
$$+\left(\frac{\gamma_j n_i}{4(\theta_i-\theta_j)^2}+\frac{\gamma_i^2\gamma_j}{4(\theta_i-\theta_j)^3}\right)\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_j}$$

Finally, the equation for $\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial^2\gamma_i^2}$ is

$$\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial^2\gamma_i^2}=\int_{\mathrm{i}\mathbb{R}+t_0}\left(\frac{1}{2(\theta_i+t)}+\frac{\gamma_i^2}{4(\theta_i+t)^2}\right)\mathcal{A}(\boldsymbol{\gamma},\boldsymbol{\theta})e^t\,\mathrm{d}t$$
$$=-\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i}-\frac{n_i-1}{\gamma_i}\frac{\partial\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\gamma_i},$$

with singularity at $\gamma_i=0$ if $n_i\geq 2$.

Explicit expressions for the Pfaffians For $\mathbf{P}_i$ only the rows i and $i+l$ will have $2l$ nonzero entries, while the remaining $2(l-1)$ rows indexed by $j\in\{1,2,\ldots,l|j\neq i\}$ and $j+l$ will have at most 4 nonzero entries as indicated in (10):

$$\mathbf{P}_i(j,i)=-\frac{n_j}{2(\theta_j-\theta_i)}-\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2}$$
$$\mathbf{P}_i(j,j)=-\frac{n_i}{2(\theta_i-\theta_j)}-\frac{\gamma_i^2}{4(\theta_i-\theta_j)^2}$$
$$\mathbf{P}_i(j,i+l)=-\frac{n_j\gamma_i}{4(\theta_j-\theta_i)^2}-\frac{\gamma_i\gamma_j^2}{4(\theta_j-\theta_i)^3}$$
$$\mathbf{P}_i(j,j+l)=-\frac{n_i\gamma_j}{4(\theta_i-\theta_j)^2}-\frac{\gamma_i^2\gamma_j}{4(\theta_i-\theta_j)^3},$$

and for the $l-1$ rows $j+l$ using (11) (with i and j interchanged) we have only 3 nonzero entries:

$$\mathbf{P}_i(j+l,i)=\frac{\gamma_j}{2(\theta_j-\theta_i)}$$
$$\mathbf{P}_i(j+l,j+l)=-\frac{n_i}{2(\theta_i-\theta_j)}-\frac{\gamma_i^2}{4(\theta_i-\theta_j)^2}$$
$$\mathbf{P}_i(j+l,i+l)=\frac{\gamma_i\gamma_j}{4(\theta_i-\theta_j)^2}$$

The ith row of $\mathbf{P}_i$ can be obtained by rewriting (14):

$$\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial^2\theta_i}=-\left(1+\sum_{i\neq j=1}^{l}\mathbf{P}_i(j,i)\right)\boldsymbol{g}_i$$
$$-\sum_{i\neq j=1}^{l}\mathbf{P}_i(j,j)\boldsymbol{g}_j-$$
$$-\sum_{i\neq j=1}^{l}\mathbf{P}_i(j,i+l)\boldsymbol{g}_{i+l}-\sum_{i\neq j=1}^{l}\mathbf{P}_i(j,j+l)\boldsymbol{g}_{j+l},$$

and therefore, the nonzero entries of $\mathbf{P}_i(i,:)$ are:

$$\mathbf{P}_i(i,i)=-\left(1+\sum_{i\neq j=1}^{l}\mathbf{P}_i(j,i)\right)$$
$$\mathbf{P}_i(i,i+l)=-\sum_{i\neq j=1}^{l}\mathbf{P}_i(j,i+l)$$
$$\mathbf{P}_i(i,j)=-\mathbf{P}_i(j,j),\mathbf{P}_i(i,j+l)=-\mathbf{P}_i(j,j+l)$$

$j\in\{1,2,\ldots,l|j\neq i\}$ and for the $i+l$th row. Please note that Eq. (13) implies that

$$\frac{\partial^2\mathcal{C}(\boldsymbol{\theta},\boldsymbol{\gamma})}{\partial\theta_i\partial\gamma_i}=-\sum_{i\neq j=1}^{l}\mathbf{P}_j(i+l,j)\boldsymbol{g}_j$$
$$-\left(1+\sum_{i\neq j=1}^{l}\mathbf{P}_j(i+l,i+l)\right)\boldsymbol{g}_{i+l}$$
$$-\sum_{i\neq j=1}^{l}\mathbf{P}_j(i+l,j+l)\boldsymbol{g}_{j+l}$$

and

$$\mathbf{P}_i(i+l,j)=-\sum_{i\neq j=1}^{l}\mathbf{P}_j(i+l,j)$$
$$\mathbf{P}_i(i+l,j+l)=-\mathbf{P}_j(i+l,j+l)$$
$$\mathbf{P}_i(i+l,i+l)=-\left(1+\sum_{i\neq j=1}^{l}\mathbf{P}_j(i+l,i+l)\right)$$

Similarly, one can show that for the second type $\mathbf{P}_{i+l}$ the only nonzero elements in the rows j and $j+l$, for all $j\neq i$ are

$$\mathbf{P}_{i+l}(j,j)=\frac{\gamma_i}{2(\theta_i-\theta_j)}$$
$$\mathbf{P}_{i+l}(j,i+l)=-\frac{n_j}{2(\theta_j-\theta_i)}-\frac{\gamma_j^2}{4(\theta_j-\theta_i)^2}$$
$$\mathbf{P}_{i+l}(j,j+l)=\frac{\gamma_i\gamma_j}{4(\theta_i-\theta_j)^2}$$
$$\mathbf{P}_{i+l}(j+l,i+l)=\frac{\gamma_j}{2(\theta_j-\theta_i)}$$
$$\mathbf{P}_{i+l}(j+l,j+l)=-\frac{\gamma_i}{2(\theta_j-\theta_i)}$$

as seen from (11). For the ith row

$$\mathbf{P}_{i+l}(i, i+l) = -1 - \sum_{i \neq j=1}^{l} \mathbf{P}_{i+l}(j, i+l)$$
$$\mathbf{P}_{i+l}(i, j) = -\mathbf{P}_{i+l}(j, j) \quad \mathbf{P}_{i+l}(i, j+l) = -\mathbf{P}_{i+l}(j, j+l))$$

and for the $(i+l)$th row

$$\mathbf{P}_{i+l}(i+l, i) = -1 \quad \mathbf{P}_{i+l}(i+l, i+l) = -\frac{n_i - 1}{\gamma_i}$$

Proof of Proposition 1 Now, for a given $\mathbf{O}$ and some direction $\boldsymbol{v}$, we can define a curve in the space of orthogonal matrices which start from $\mathbf{O}$: $\mathbf{O}(t) = e^{\boldsymbol{v}t}\mathbf{O}$ where $\boldsymbol{v}$ is skew symmetric. Such curves clearly start from $\mathbf{O}$ since $e^{\boldsymbol{v}t}|_{t=0} = \mathbf{I}$. Since $\frac{\log L(\boldsymbol{\theta}, \mathbf{O}(t), \boldsymbol{\gamma})}{\partial t} = -\frac{tr(\mathbf{A}\mathbf{O}^\top \mathrm{diag}(\boldsymbol{\theta})\mathbf{O} + \boldsymbol{\gamma}\mathbf{B}^\top\mathbf{O}^\top)}{\partial t}$ and $\frac{\partial \mathbf{O}(t)}{\partial t}|_{t=0} = \boldsymbol{v}\mathbf{O}$, $-\boldsymbol{v}^\top = \boldsymbol{v}$ we obtain

$$\begin{aligned}
&\frac{\partial \log L(\boldsymbol{\theta}, \mathbf{O}(t), \boldsymbol{\gamma})}{\partial t}\Big|_{t=0} \\
&= -\frac{\partial tr(\mathbf{A}\mathbf{O}^\top e^{\boldsymbol{v}^\top t}\mathrm{diag}(\boldsymbol{\theta})e^{\boldsymbol{v}t}\mathbf{O} + \mathbf{B}^\top\mathbf{O}^\top e^{\boldsymbol{v}^\top t}\boldsymbol{\gamma})}{\partial t}\Big|_{t=0} \\
&= -tr(\mathbf{A}\mathbf{O}^\top \boldsymbol{v}^\top e^{\boldsymbol{v}^\top t}\mathrm{diag}(\boldsymbol{\theta})e^{\boldsymbol{v}t}\mathbf{O} \\
&\quad + \mathbf{A}\mathbf{O}^\top e^{\boldsymbol{v}^\top t}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{v}e^{\boldsymbol{v}t}\mathbf{O} \\
&\quad + \mathbf{B}^\top\mathbf{O}^\top\boldsymbol{v}^\top e^{\boldsymbol{v}^\top t}\boldsymbol{\gamma})|_{t=0} \\
&= -tr(\mathbf{A}\mathbf{O}^\top \boldsymbol{v}^\top \mathrm{diag}(\boldsymbol{\theta})\mathbf{O} + \mathbf{A}\mathbf{O}^\top \mathrm{diag}(\boldsymbol{\theta})\boldsymbol{v}\mathbf{O} \\
&\quad + \mathbf{B}^\top\mathbf{O}^\top\boldsymbol{v}^\top\boldsymbol{\gamma}) \\
&= tr\left(\boldsymbol{v}(\mathrm{diag}(\boldsymbol{\theta})\mathbf{O}\mathbf{A}\mathbf{O}^\top - \mathbf{O}\mathbf{A}\mathbf{O}^\top\mathrm{diag}(\boldsymbol{\theta}) + \boldsymbol{\gamma}\mathbf{B}^\top\mathbf{O}^\top)\right)
\end{aligned}$$

This derivative is zero for any skew symmetric matrix $\boldsymbol{v}$ only if

$$\mathcal{A} = \mathrm{diag}(\boldsymbol{\theta})\mathbf{O}\mathbf{A}\mathbf{O}^\top - \mathbf{O}\mathbf{A}\mathbf{O}^\top\mathrm{diag}(\boldsymbol{\theta}) + \boldsymbol{\gamma}\mathbf{B}^\top\mathbf{O}^\top$$

is symmetric, i.e. $\mathcal{A} = \mathcal{A}^\top$.

References

Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas. Graphs and Mathematical Functions. Dover, New York (1972)

Arnold, R., Jupp, P.E.: Statistics of orthogonal axial frames. Biometrika **100**, 571–586 (2013)

Chu, M.T., Trendafilov, N.T.: On a differential equation approach to the weighted orthogonal Procrustes problem. Stat. Comput. **8**, 125–133 (1998)

Fallaize, C.J., Kypraios, T.: Exact Bayesian Inference for the Bingham Distribution Statistics and Computing. Springer, Berlin (2014)

Hashiguchi, H., Numata, Y., Takayama, N., Takemura, A.: Holonomic gradient method for the distribution function of the largest root of a wishart matrix. J. Multivar. Anal. **117**, 296–312 (2013)

Hibi, T. (ed.): Gröbner Bases: Statistics and Software Systems. Springer, New York (2013)

Kent, J.T.: The Fisher–Bingham distribution on the sphere. J. R. Stat. Soc. Ser. B **44**, 71–80 (1982)

Koyama, T.: A Holonomic Ideal Annihilating the Fisher-Bingham Integral. arxiv:1104.1411 (2011)

Koyama, T., Takemura, A.: Holonomic gradient method for distribution function of a weighted sum of noncentral chi-square random variables. Comput. Stat. **31**, 1645–1659 (2016)

Koyama, T., Nakayama, H., Nishiyama, K., Takayama, N.: The Holonomic Rank of the Fisher–Bingham System of Differential Equations. arxiv:1205.6144 (2012)

Koyama, T., Nakayama, H., Nishiyama, K., Takayama, N.: Holonomic gradient descent for the Fisher–Bingham distribution on the n-dimensional sphere. Comput. Stat. **29**, 661–683 (2014)

Kume, A., Walker, S.G.: On the Fisher–Bingham distribution. Stat. Comput. **19**, 167–172 (2009)

Kume, A., Wood, A.T.A.: Saddlepoint approximations for the Bingham and Fisher–Bingham normalising constants. Biometrika **92**, 465–476 (2005)

Kume, A., Wood, A.T.A.: On the derivatives of the normalising constant of the Bingham distribution. Stat. Probab. Lett. **77**, 832–837 (2007)

Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley Series in Probability and Statistics. Wiley, Chichester (2000)

Nakayama, H., Nishiyama, K., Noro, M., Ohara, K., Sei, T., Takayama, N., Takemura, A.: Holonomic gradient descent and its application to the Fisher–Bingham integral. Adv. Appl. Math. **47**, 639–658 (2011)

Sei, T., Kume, A.: Calculating the normalising constant of the Bingham distribution on the sphere using the holonomic gradient method. Stat. Comput. **25**, 321–332 (2015)

Sei, T., Shibata, H., Takemura, A., Ohara, K., Takayama, N.: Properties and applications of Fisher distribution on the rotation group. J. Multivar. Anal. **116**, 440–455 (2013)

Wood, A.T.A.: Estimation of the concentration parameters of the Fisher matrix distribution on $SO(3)$ and the Bingham distribution on S_q, $q \geq 2$. Aust. J. Stat. **35**, 69–79 (1993). (**Wiley Online Library**)

Zarowsky, C.J.: An Introduction to Numerical Analysis for Electrical and Computer Engineers. Wiley, London (2004)