Scaled von Mises-Fisher distributions and regression models for paleomagnetic directional data

J. L. Scealy[1,*] and Andrew T. A. Wood[2]

[1] *Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra ACT 2601, Australia*

[2] *School of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, UK*

[*] *Corresponding author's email: janice.scealy@anu.edu.au.*

Abstract

We propose a new distribution for analysing paleomagnetic directional data that is a novel transformation of the von Mises-Fisher distribution. The new distribution has ellipse-like symmetry, as does the Kent distribution; however, unlike the Kent distribution the normalising constant in the new density is easy to compute and estimation of the shape parameters is straightforward. To accommodate outliers, the model also incorporates an additional shape parameter which controls the tail-weight of the distribution. We also develop a general regression model framework that allows both the mean direction and the shape parameters of the error distribution to depend on covariates. The proposed regression procedure is shown to be equivariant with respect to the choice of coordinate system for the directional response. To illustrate, we analyse paleomagnetic directional data from the GEOMAGIA50.v3 database (Brown et al. 2015). We predict the mean direction at various geological

time points and show that there is significant heteroscedasticity present. It is envisaged that the regression structures and error distribution proposed here will also prove useful when covariate information is available with (i) other types of directional response data; and (ii) square-root transformed compositional data of general dimension.

Keywords: heteroscedasticity; regression; spherical data; t-distribution.

1 Introduction

1.1 Background: paleomagnetic directional data

Spherical data are frequently encountered in the earth and environmental sciences (e.g. Schuenemeyer and Drew, 2011; Borradaile, 2003). A common example is paleomagnetic data consisting of observations on the direction of magnetism in either rocks, sediment or in archeological specimens, measured at various geological time points and spatial locations. The directions are usually measured as declination and inclination angles based on strike and dip coordinates (see Schuenemeyer and Drew (2011, p. 379) for a full definition). Often it is of interest to calculate a sample mean and standard error estimate of the direction at a particular spatial location and in small geological time ranges (e.g. Acton et al., 2000, p. 166). In other cases, depending on the data available, it is of interest to explore the relationships between the directions versus geological time and/or space to understand how the Earth's magnetic field has evolved. In this case, to account for the highly non-linear relationships between the geomagnetic field directions and the covariates, in the geophysics literature, the geomagnetic field is usually expressed in terms of spherical harmonics, and the temporal evolution of the process is modelled using cubic B-splines. The residuals in these models are then assumed to have either an approximate Gaussian or Laplace

distribution (e.g. Walker and Jackson, 2000; Panovska et al., 2015). Paleomagnetic data is typically heavy-tailed and contains outliers (e.g. Acton et al., 2000; Panovska et al., 2015).

In this paper we focus on analysing archeomagnetic data in the GEOMAGIA50.v3 database (GMAG; Brown et al., 2015), extracted in February 2017. GMAG is a very detailed online database providing access to a large amount of published paleomagnetic, rock magnetic, and chronological data from a variety of materials that record Earth's magnetic field over the past 50,000 years. For simplicity we restrict our analysis to a single spatial location which is the Eifel maars (EIF) lakes in Germany. Similarly to Panovska et al. (2015), we relocate nearby archeomagnetic data (latitudes in the range $[40°, 60°]$ and longitudes in the range $[-3°, 17°]$) to the EIF location using an axial dipole correction as defined at equation (1) in Noel and Batt (1990). Our archeomagnetic data is therefore equivalent or close to equivalent to Panovska et al. (2015), Figure 10, top two plots (we exclude the sediment data). These plots are given here in Figure 1 and they show that the angles may be heavy-tailed and there is some evidence of non constant variability (heteroscedasticity) across time. Before we analyse the data, we convert these angles to Cartesian coordinates defined on $\mathcal{S}^2$, where $\mathcal{S}^{p-1}$ denotes the unit sphere $\{\boldsymbol{y} \in \mathbb{R}^p : ||\boldsymbol{y}|| = 1\}$. In the conversion we use the following reference frame: $y_1 = \sin I$, $y_2 = \cos I \cos D$ and $y_3 = \cos I \sin D$, where I represents inclination defined on $[-90°, 90°]$ and D represents declination defined on $[0°, 360°]$.

Historically both the Kent distribution and von Mises-Fisher have been used to a limited extent to summarise paleomagnetic data samples (e.g. Fisher et al., 1987; Tauxe, 2010). One major issue with the Kent distribution is that the normalising constant does not exist in closed form and involves multidimensional integrals that are difficult to compute. This has led to the use of either a high concentration or

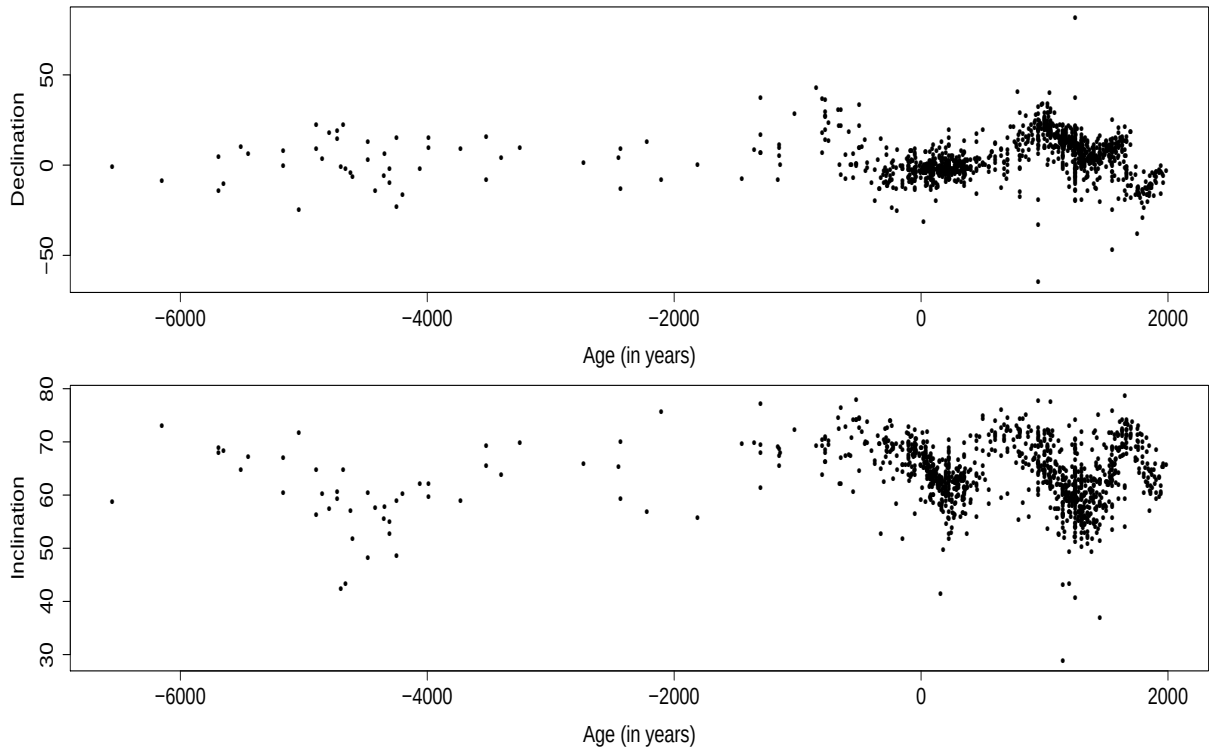

Figure 1: EIF GMAG data. Top: Declination versus Age (in years) scatterplot; Bottom: Inclination versus Age (in years) scatterplot.

saddlepoint density approximation (Kent, 1982; Kume and Wood, 2005) to estimate the shape parameters. However, often the residual variability in applications is not small and these methods can lead to biased estimates especially when the shape parameters are spread and the ellipticity is high (Scealy and Welsh, 2014). More recently Scealy and Welsh (2017) avoided the issue of estimating the shape parameters by instead modelling and then estimating the first- and second-order moments of the Kent distribution. They based inference for the moments on a nonparametric bootstrap method, but this has the disadvantage of being computationally intensive and is cumbersome to apply.

1.2 Main contributions of the paper

We propose a new family of flexible yet tractable error distributions for directional response data, which we call the Scaled von Mises-Fisher (SvMF) family. The SvMF family is generated by applying a bijective transformation of the sphere to itself which is defined in Section 2. The SvMF family has the same symmetry properties as the Kent (1982) distribution and the new density has virtually identical contours to the Kent density for certain ranges of the shape parameters. However, unlike the Kent distribution, the normalising constant in a SvMF distribution is essentially that of the underlying von Mises-Fisher distribution and is therefore highly tractable. A further interesting property of the SvMF family is an extra parameter, which should be thought of as a tuning parameter, which allows some control of the tail-weight of the distribution for a given level of concentration of the distribution. We demonstrate that the shape parameters can be estimated in a computationally convenient way using standard maximum likelihood estimation methods. Moreover, we show how the new model can be used in the regression setting, allowing both the mean direction and the shape parameters to be modelled directly as functions of a general covariate

vector. Simulation from the new model is also straightforward, as it just involves a simple transformation of a von Mises-Fisher random variable. We use this new modelling approach to analyse GMAG paleomagnetic data (see Brown et al., 2015).

1.3 Relevant literature

We briefly mention some other families of distributions on $\mathcal{S}^p$ of interest but do not consider them further in this paper. Jones and Pewsey (2005) point out that the family they consider on the circle $\mathcal{S}^1$ has an extension to $\mathcal{S}^p$ where $p > 1$. However, this family necessarily exhibits rotational symmetry, unlike the families considered here and mentioned below. Second, Paine et al. (2018) consider a subfamily of the angular Gaussian family (see Mardia and Jupp, 2000 for the definition) whose distributions are Kent-like, i.e. they have contours of constant density which exhibit ellipse-like symmetry. This angular Gaussian subfamily has some similar features to the family proposed here, though the mathematical form of the density is somewhat different. Downs and Mardia (2002) and Kato and Jones (2010) have considered families of distributions on the unit circle $\mathcal{S}^1$ generated by the Möbius transformation, while Kato and McCullagh (2015) propose a Cauchy family of distributions on the unit sphere $\mathcal{S}^p$, $p \geq 1$, which is based on a Möbius transformation on $\mathcal{S}^p$; see Section 3 of their paper. However, although these constructions are similar in spirit to the construction proposed here, the resulting families of distributions are quite different.

Jupp and Kent (1987) and Di Marzio et al. (2014) propose nonparametric regression approaches on the sphere, but restricted to the case of a scalar covariate or a unit vector covariate (in the latter case only), and an isotropic error structure appears to be assumed in both papers. In contrast, our goal here is to develop a general flexible regression framework on the sphere which can handle general vector covariate structures and can accommodate heavy-tails and heteroscedasticity, without assuming a

priori that the error distribution is rotationally symmetric. Finally, we mention that Rivest et al. (2016) suggest some interesting ideas for regression modelling on the circle; some of these ideas may prove useful for regression modelling on the sphere.

1.4 Structure of the paper

The rest of this paper is organised as follows. In Section 2 we specify the family of transformations of $\mathcal{S}^{p-1}$ used to create the SvMF family, which is presented in Section 3. In Section 4 we propose iterative estimation schemes for the parameters, first considering the independent and identically distributed (IID) case and then focusing on the regression case. In Section 5 we describe our analysis of the GMAG data discussed in Section 1.1. In Section 6 we present simulation results which provide information about the properties of the parameter estimators for the SvMF model in the IID case. Conclusions are briefly summarised in Section 7 and proofs are given in appendices. Although the paleomagnetic directional data is defined on $\mathcal{S}^2$, throughout most of the paper we keep the dimension $p \geq 3$ quite general.

2 A group of transformations on $\mathcal{S}^{p-1}$

The best-known transformation group on the unit sphere is of course the group of isometries. In a given Cartesian coordinate system, such an isometry may be represented by $\boldsymbol{y} \mapsto \boldsymbol{\Gamma}\boldsymbol{y}$ where $\boldsymbol{\Gamma}$ is a $p \times p$ orthogonal matrix satisfying $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top = \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. We now consider a second type of transformation. In words, we consider a bijection of $\mathcal{S}^{p-1}$ onto itself obtained by rescaling the coordinate axes in the ambient space $\mathbb{R}^p$, and then projecting each point in the image of $\mathcal{S}^{p-1}$ (under the linear transformation of the ambient space) back onto the unit sphere.

To make this more mathematically explicit, define $\mathbb{R}_+$ to be the set of strictly positive real numbers. For each $\boldsymbol{a} = (a_1, \ldots, a_p)^\top \in \mathbb{R}_+^p$, define the transformation $\mathcal{T}_{\boldsymbol{a}} : \mathcal{S}^{p-1} \to \mathcal{S}^{p-1}$ by

$$\boldsymbol{z} = \mathcal{T}_{\boldsymbol{a}}(\boldsymbol{y}) = \frac{1}{\{\sum_{i=1}^p (y_i/a_i)^2\}^{1/2}} \left(\frac{y_1}{a_1}, \ldots, \frac{y_p}{a_p}\right)^\top, \quad \boldsymbol{y} \in \mathcal{S}^{p-1}. \tag{1}$$

By construction, $\boldsymbol{z} \in \mathcal{S}^{p-1}$, and it is clear that $\mathcal{T}_{\boldsymbol{a}}$ is a bijection from $\mathcal{S}^{p-1}$ to itself. Moreover, the set of transformations $\{\mathcal{T}_{\boldsymbol{a}} : \boldsymbol{a} \in \mathbb{R}_+^p\}$ forms a group with group operation $\mathcal{T}_{\boldsymbol{a}} \circ \mathcal{T}_{\boldsymbol{b}} = \mathcal{T}_{\boldsymbol{a} \circ \boldsymbol{b}}$ where, abusing notation slightly, we have used the same symbol for the group operation and for the Hadamard product of two vectors; here $\boldsymbol{a} \circ \boldsymbol{b} = (a_1 b_1, \ldots, a_p b_p)^\top$, where $\boldsymbol{b} = (b_1, \ldots, b_p)^\top$. Note that the inverse transformation $\mathcal{T}_{\boldsymbol{a}}^{-1}$ is given by $\mathcal{T}_{\boldsymbol{b}}$ where $\boldsymbol{b} = (1/a_1, \ldots, 1/a_p)^\top$.

Let $d\mathcal{S}^{p-1}$ denote the standard geometric measure on the unit sphere. Let $\boldsymbol{Z}$ denote a random unit vector in $\mathcal{S}^{p-1}$ with probability density function $f_Z(\boldsymbol{z})$ with respect to the surface area measure $d\mathcal{S}^{p-1}$. Then, since for each $\boldsymbol{a} \in \mathbb{R}_+^p$, $\mathcal{T}_{\boldsymbol{a}}$ defines a smooth bijection, it follows that if $\boldsymbol{Z} = \mathcal{T}_{\boldsymbol{a}}(\boldsymbol{Y})$ then the random unit vector $\boldsymbol{Y} \in \mathcal{S}^{p-1}$ has probability density function which satisfies

$$f_Y(\boldsymbol{y}) = f_Z\{\mathcal{T}_{\boldsymbol{a}}(\boldsymbol{y})\} J_a(\boldsymbol{y}). \tag{2}$$

The Jacobian function $J_a(\boldsymbol{y})$ is determined in the following lemma whose proof is given in Appendix A.1.

Lemma 2.1. *For $\boldsymbol{a} \in \mathbb{R}_+^p$ and $\boldsymbol{y} \in \mathcal{S}^{p-1}$, the function $J_a(\boldsymbol{y})$ is given by*

$$J_a(\boldsymbol{y}) = \left(\prod_{i=1}^p a_i\right)^{-1} \left\{\sum_{i=1}^p \left(\frac{y_i}{a_i}\right)^2\right\}^{-(p-1)/2}. \tag{3}$$

It is interesting to note that, when we take $f_Z(\boldsymbol{z})$ to be the probability density

function of the uniform distribution on $\mathcal{S}^{p-1}$, the resulting distribution of $\boldsymbol{y}$ turns out to be the angular central Gaussian distribution; see Watson (1983, p. 110) and Mardia and Jupp (2000).

3 Construction of a Kent-like distribution

When we take $f_Z(\boldsymbol{z})$ to be the von Mises-Fisher distribution and apply the transformation $\mathcal{T}_{\boldsymbol{a}}$, we obtain a useful and seemingly new family of distributions, referred to as the SvMF family in the Introduction. Suppose that the components of $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ in $\mathcal{T}_{\boldsymbol{a}}$ satisfy

$$\prod_{j=2}^{p} a_j = 1, \tag{4}$$

and let $f_Z(\boldsymbol{z})$ denote the probability density function of the von Mises-Fisher distribution with respect to geometric measure $d\mathcal{S}^{p-1}(\boldsymbol{z})$ on $\mathcal{S}^{p-1}$, and given by

$$f_Z(\boldsymbol{z}) = \{c_p(\kappa)\}^{-1} \exp(\kappa \boldsymbol{e}_1^\top \boldsymbol{z}) = \{c_p(\kappa)\}^{-1} \exp(\kappa z_1), \tag{5}$$

where $\boldsymbol{z} = (z_1, \ldots, z_p)^\top$, $\boldsymbol{e}_j$ is the p-vector with component j equal to 1 and all other components zero, $j = 1, \ldots, p$, $c_p(\kappa) = (2\pi)^{p/2} I_{(p/2)-1}(\kappa)/\kappa^{(p/2)-1}$ is the normalising constant, and I_ν denotes the modified Bessel function of the first kind of order ν. When $p = 3$, the normalising constant takes the simple form $c_p(\kappa) = 2\pi(e^\kappa - e^{-\kappa})/\kappa$.

Substituting $\boldsymbol{z} = \mathcal{T}_{\boldsymbol{a}}(\boldsymbol{y})$, where $\boldsymbol{a}$ satisfies (4), and using (2) and (3), leads to the probability density function

$$f_Y(\boldsymbol{y}) = \{c_p(\kappa) a_1\}^{-1} \left\{ \sum_{j=1}^{p} (y_j/a_j)^2 \right\}^{-(p-1)/2} \exp\left\{ \frac{\kappa\,(y_1/a_1)}{\left\{ \sum_{j=1}^{p} (y_j/a_j)^2 \right\}^{1/2}} \right\}. \tag{6}$$

In (6), the coordinate axes play a special role. We may write $y_j = \boldsymbol{y}^\top \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the jth coordinate axis defined above, and generalising from $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\}$ to a general orthonormal basis $\{\boldsymbol{\mu}, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p\}$, we obtain the general density

$$f_Y(\boldsymbol{y}) = \{c_p(\kappa)a_1\}^{-1} \left\{ (\boldsymbol{y}^\top \boldsymbol{\mu}/a_1)^2 + \sum_{j=2}^{p} \left(\boldsymbol{y}^\top \boldsymbol{\gamma}_j/a_j\right)^2 \right\}^{-(p-1)/2} \times \exp\left\{ \frac{\kappa \boldsymbol{y}^\top \boldsymbol{\mu}/a_1}{\left\{ (\boldsymbol{y}^\top \boldsymbol{\mu}/a_1)^2 + \sum_{j=2}^{p} \left(\boldsymbol{y}^\top \boldsymbol{\gamma}_j/a_j\right)^2 \right\}^{1/2}} \right\}. \tag{7}$$

Three theoretical results are now presented. The first result gives a sufficient condition for the density to be unimodal. The proof is given in Appendix A.2.

Proposition 1. *Consider the density (7) on $\mathcal{S}^{p-1}$ where (4) is satisfied, and without loss of generality assume that $a_2 = \max(a_2, \ldots, a_p)$ and $a_1 \geq 1$. Then (7) is unimodal and has a unique mode $\boldsymbol{y} = \boldsymbol{\mu}$ if*

$$\kappa \geq a_1(p-1)\left((a_2/a_1)^2 - 1\right) \text{ and } a_1 \leq a_2.$$

If $\kappa > 0$ and $a_1 > a_2$ then, on the other hand, the density has a global maximum at $\boldsymbol{y} = \boldsymbol{\mu}$ (but is not necessarily unimodal).

Our second result shows that in the high-concentration limit, i.e. when $\kappa \to \infty$, the density is asymptotically Gaussian. The proof is given in Appendix A.3.

Proposition 2. *Let $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_p^*)^\top \in \mathcal{S}^{p-1}$ be a random variable with density (6) and define $\boldsymbol{y}_L^* = (y_2^*, y_3^*, \ldots, y_p^*)^\top$. Then with $a_1, a_2, \ldots, a_p$ held fixed and $\kappa \to \infty$,*

$$\kappa^{1/2} \boldsymbol{y}_L^* \xrightarrow{d} N_{p-1}\left(\boldsymbol{0}_{p-1}, \text{Diag}\left((a_2/a_1)^2, (a_3/a_1)^2 \ldots, (a_p/a_1)^2\right)\right).$$

Our third result shows that the mean direction is $\boldsymbol{\mu}$ and the columns of $\boldsymbol{\Gamma} = \{\boldsymbol{\mu}, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p\}$ are the eigenvectors corresponding to the second-order moment

matrix. The proof is given in Appendix A.4.

Proposition 3. *Let $\boldsymbol{y} \in \mathcal{S}^{p-1}$ be a random variable with density (7), let $\boldsymbol{y}^* = (y_1^*, y_2^*, \ldots, y_p^*)^\top \in \mathcal{S}^{p-1}$ be a random variable with density (6). Then with $\kappa > 0$ and $a_1 > 0$,*

$$\mathrm{E}\left(\boldsymbol{y}\right) = \mathrm{E}\left(y_1^*\right)\boldsymbol{\mu} \quad \text{and} \quad \mathrm{E}\left(\boldsymbol{y}\boldsymbol{y}^\top\right) = \boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\Gamma}^\top,$$

where $\boldsymbol{D}$ is a diagonal $p \times p$ matrix and is a non-linear function of the shape parameters $\kappa, a_1, a_2, \ldots, a_p$.

Later, for estimation it will also prove useful to consider the following alternative parameterisation of $\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p$ and $a_2, a_3, \ldots, a_p$. Following Scealy and Welsh (2014), define the $p \times p$ orthogonal matrix

$$\boldsymbol{H}(\boldsymbol{\mu}) = \begin{pmatrix} \mu_1 & \boldsymbol{\mu}_L^\top \\ \boldsymbol{\mu}_L & \frac{1}{1+\mu_1}\boldsymbol{\mu}_L\boldsymbol{\mu}_L^\top - \boldsymbol{I}_{p-1} \end{pmatrix} = \left\{\boldsymbol{\mu}, \boldsymbol{H}^*(\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu}_L = (\mu_2, \mu_3, ..., \mu_p)^\top$ and $\boldsymbol{H}^*(\boldsymbol{\mu})$ is a $p \times (p-1)$ matrix whose columns are orthogonal to $\boldsymbol{\mu}$. Let $\boldsymbol{K}^*$ be a general $(p-1) \times (p-1)$ orthogonal matrix defined such that

$$\boldsymbol{\Gamma} = \left\{\boldsymbol{\mu}, \boldsymbol{H}^*(\boldsymbol{\mu})\boldsymbol{K}^*\right\} \tag{8}$$

holds. Then let

$$\boldsymbol{V} = \boldsymbol{K}^* \mathrm{Diag}(a_2^2, a_3^2, \ldots, a_p^2) \boldsymbol{K}^{*\top}, \tag{9}$$

where $\boldsymbol{V}$ is a $(p-1) \times (p-1)$ dimensional symmetric positive definite matrix with the constraint $\det(\boldsymbol{V}) = 1$, which corresponds to condition (4); and assume $a_2 > a_3 > \ldots > a_p$. In general the lower $p-1$ elements on the diagonal of $\boldsymbol{D}$ in Proposition 3 do not correspond to the eigenvalues of $\boldsymbol{V}$ except under high concentration (see Proposition 2).

To obtain a Kent-like distribution it is convenient, for many practical purposes, to set $a_1 = 1$. However, numerical investigations indicate that as a_1 increases the density becomes heavier tailed with a higher probability of outliers and the shapes of the densities in the tangent space are more similar to those of a multivariate t-distribution of dimension $p - 1$. The model is a $\mathscr{Q}$-symmetric model as defined by Rivest (1984) and Rivest showed that the information matrix in such models is block diagonal (with the first block associated with location parameters and the second block associated with the shape parameters). In this context the shape parameters are $\kappa, a_1, a_2, \ldots, a_p$ and the location parameters are $\{\boldsymbol{\mu}, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p\}$. It is also straightforward to prove that, under condition (4), both a_1 and κ are information orthogonal to all the shape parameters $\{a_2, a_3, \ldots, a_p\}$, but a_1 and κ are not information orthogonal to each other. In fact, as discussed later, a_1 and κ are not jointly estimable by maximum likelihood.

4 Models and estimators

In this section we assume that a_1 is fixed and not estimated. By default we suggest setting $a_1 = 1$ unless a heavier-tailed density is required.

4.1 Independent and identically distributed data case

Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be an independent and identically distributed sample from the distribution with density (7) and let $\boldsymbol{y}_i = (y_{1,i}, y_{2,i}, \ldots, y_{p,i})^\top$ for $i = 1, 2, \ldots, n$ denote the observed values of these random variables.

4.1.1 Moment and M-estimators of location parameters

The location parameters in $\boldsymbol{\Gamma}$ can be estimated straightforwardly by using the Kent (1982) moment estimators. The moment estimator of $\boldsymbol{\mu}$ is the sample mean direction

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \boldsymbol{y}_i}{\| \sum_{i=1}^{n} \boldsymbol{y}_i \|}, \tag{10}$$

and the moment estimators of $\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_p$ denoted by $\tilde{\boldsymbol{\gamma}}_2, \tilde{\boldsymbol{\gamma}}_3, \ldots, \tilde{\boldsymbol{\gamma}}_p$ respectively, are the unit eigenvectors corresponding, in decreasing order, to the $p-1$ strictly positive eigenvalues of

$$\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top\right) \left(\sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^\top \right) \left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top\right).$$

Note that there is some non-uniqueness in the definition of the $\tilde{\boldsymbol{\Gamma}}$, in that any choice of the form $\{\tilde{\boldsymbol{\mu}}, \pm\tilde{\boldsymbol{\gamma}}_2, \pm\tilde{\boldsymbol{\gamma}}_3, \ldots, \pm\tilde{\boldsymbol{\gamma}}_p\}$ will suffice. If we wish to specify the signs of the $\tilde{\boldsymbol{\gamma}}_j$ uniquely, we can do this with probability one by choosing, for example, the first component of each $\tilde{\boldsymbol{\gamma}}_j$ to be positive.

The sample mean direction may not be efficient for heavy-tailed distributions (as seen in the simulation experiment in Section 6). In this case, the normalised spatial median estimator or the spherical median estimator of location available for the von Mises-Fisher distribution can be used (e.g. Ko and Chang, 1993). These M-estimators are consistent under the model due to symmetry.

4.1.2 Maximum likelihood estimation of all parameters

If $\boldsymbol{\mu}$ and κ are known, then the log-likelihood for $\boldsymbol{V}$ is

$$-\frac{(p-1)}{2} \sum_{i=1}^{n} \log\left(y_{1,i}^{**2} a_1^{-2} + \boldsymbol{y}_{L,i}^{**\top} \boldsymbol{V}^{-1} \boldsymbol{y}_{L,i}^{**}\right) + \sum_{i=1}^{n} \frac{\kappa a_1^{-1} y_{1,i}^{**}}{\left(y_{1,i}^{**2} a_1^{-2} + \boldsymbol{y}_{L,i}^{**\top} \boldsymbol{V}^{-1} \boldsymbol{y}_{L,i}^{**}\right)^{1/2}}, \tag{11}$$

where $y_{1,i}^{**} = \boldsymbol{\mu}^\top \boldsymbol{y}_i$ and $\boldsymbol{y}_{L,i}^{**} = \boldsymbol{H}^*(\boldsymbol{\mu})^\top \boldsymbol{y}_i$. To estimate $\boldsymbol{V}$ we maximise (11) with respect to $\boldsymbol{V}$ subject to $\det(\boldsymbol{V}) = 1$ or equivalently subject to $\log\{\det(\boldsymbol{V})\} = 0$. This constrained optimisation problem can be solved by using the method of Lagrange multipliers and the resulting Lagrangian function has a similar form to the log-likelihood for a general scatter matrix for an elliptically symmetric distribution defined on $\mathbb{R}^{p-1}$. Similar to Maronna (1976, pp. 51-52), maximising the Lagrangian function leads to the following estimating equation:

$$\boldsymbol{V} \propto \sum_{i=1}^{n} \left((p-1) \left(\check{s}_i^2 \right)^{-1} + \kappa a_1^{-1} y_{1,i}^{**} \left(\check{s}_i^2 \right)^{-3/2} \right) \boldsymbol{y}_{L,i}^{**} \boldsymbol{y}_{L,i}^{**\top},$$

where $\check{s}_i^2 = y_{1,i}^{**2} a_1^{-2} + \boldsymbol{y}_{L,i}^{**\top} \boldsymbol{V}^{-1} \boldsymbol{y}_{L,i}^{**}$. This estimating equation can be solved by applying the following iterative reweighting algorithm

$$\hat{\boldsymbol{V}}_{m+1} = \frac{\sum_{i=1}^{n} \left((p-1) \left(\hat{s}_{i(m)}^2 \right)^{-1} + \kappa a_1^{-1} y_{1,i}^{**} \left(\hat{s}_{i(m)}^2 \right)^{-3/2} \right) \boldsymbol{y}_{L,i}^{**} \boldsymbol{y}_{L,i}^{**\top}}{\left(\det \left(\sum_{i=1}^{n} \left((p-1) \left(\hat{s}_{i(m)}^2 \right)^{-1} + \kappa a_1^{-1} y_{1,i}^{**} \left(\hat{s}_{i(m)}^2 \right)^{-3/2} \right) \boldsymbol{y}_{L,i}^{**} \boldsymbol{y}_{L,i}^{**\top} \right) \right)^{1/(p-1)}},$$

for some suitable starting value such as $\hat{\boldsymbol{V}}_0 = \boldsymbol{I}_{p-1}$, where $\hat{s}_{i(m)}^2 = y_{1,i}^{**2} a_1^{-2} + \boldsymbol{y}_{L,i}^{**\top} \hat{\boldsymbol{V}}_m^{-1} \boldsymbol{y}_{L,i}^{**}$. If $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are known, then the log-likelihood for κ is

$$-n \log\left(c_p(\kappa)\right) + \kappa a_1^{-1} \sum_{i=1}^{n} \boldsymbol{y}_i^\top \boldsymbol{\mu} \left(\left(\boldsymbol{y}_i^\top \boldsymbol{\mu} \right)^2 a_1^{-2} + \boldsymbol{y}_i^\top \boldsymbol{H}^*(\boldsymbol{\mu}) \boldsymbol{V}^{-1} \boldsymbol{H}^*(\boldsymbol{\mu})^\top \boldsymbol{y}_i \right)^{-1/2}. \quad (12)$$

The modified Bessel function of the first kind is available in many software packages including in R and therefore the above log-likelihood is straightforward to maximise by applying one dimensional derivative free interval search methods. Given $\boldsymbol{\mu}$, to compute joint estimates of κ and $\boldsymbol{V}$ we suggest iterating between maximising (11) and (12), where the most recent update of $\boldsymbol{V}$ is used in (12) and the most recent update of κ is used in (11), until convergence of both sets of parameters.

In practice $\boldsymbol{\mu}$, κ and $\boldsymbol{V}$ are all unknown so we suggest first calculating a preliminary estimate of $\boldsymbol{\mu}$ using the sample mean direction (10) or the normalised spatial median estimator and then maximising the log-likelihood conditional on the preliminary estimate of $\boldsymbol{\mu}$ to update κ and $\boldsymbol{V}$. Then, given the κ and $\boldsymbol{V}$ estimates, we suggest maximising the log-likelihood for $\boldsymbol{\mu}$ to obtain a second, but more efficient estimate of $\boldsymbol{\mu}$. This second estimator of $\boldsymbol{\mu}$ can be calculated using the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) when the dimension p is low.

The parameter a_1 is set to a fixed value because it was not possible to jointly estimate both κ and a_1 at the same time using the method of maximum likelihood estimation. Specifically, κ and a_1 are not jointly identifiable. We simulated lots of datasets, both heavy-tailed and not heavy-tailed and we observed that in all of these cases the log-likelihood function increased as $a_1 \to 0$ and $\kappa \to \infty$. This phenomenon of parameters approaching boundary points is not unusual when modelling error distributions with an extra shape parameter; for example, see the comment in Taylor (1992, p. 41). Even for the t-distribution, often the degrees of freedom parameter is treated as a tuning constant rather than estimated because maximum likelihood estimation can sometimes give unsatisfactory results (e.g. Lange et al., 1989).

4.1.3 Preliminary transformation

For computational convenience, prior to estimation we suggest applying the following orthogonal transformation to the response data

$$\boldsymbol{y}_i = (y_{1,i}, y_{2,i}, \ldots, y_{p,i})^\top = \tilde{\boldsymbol{\Gamma}}^\top \tilde{\boldsymbol{y}}_i, \quad i = 1, 2, \ldots, n, \tag{13}$$

where the $\tilde{\boldsymbol{y}}_i \in \mathcal{S}^{p-1}$ are the original data in Cartesian coordinates and $\tilde{\boldsymbol{\Gamma}}$ is the moment estimator of $\boldsymbol{\Gamma}$ based on the original data. This preliminary transformation is needed to ensure the final estimates of $\boldsymbol{\mu}$ and $\boldsymbol{K}^*$ are not too far from the north

pole $(1, 0, \ldots, 0)^\top$ and the identity matrix, respectively. This initial transformation will lead to approximate information orthogonality of $\boldsymbol{\mu}$ and $\boldsymbol{K}^*$ and will be exact in the large sample limit case. Note that Kent et al. (2006, pp. 758-759) applied a similar idea in estimation for the complex Bingham quartic distribution which is the analog of the Kent distribution in landmark-based shape analysis for $2D$ objects.

4.2 Regression case

Assume we have vector responses $\{\boldsymbol{Y}_i \in \mathcal{S}^{p-1} : i = 1, 2, \ldots, n\}$ associated with a set of covariates $\{\boldsymbol{X}_i \in \mathbb{R}^q : i = 1, 2, \ldots, n\}$ and the responses are assumed to be conditionally independent given the covariates. Scealy and Welsh (2011) modelled the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i = \boldsymbol{x}_i$ as having a Kent distribution (Kent, 1982). In this model the location parameters were modelled as a function of $\boldsymbol{x}_i$ and the shape parameters were assumed to be constant. We now describe a tractable way to also model shape parameters as functions of $\boldsymbol{x}_i$. We assume that the density of each $\boldsymbol{Y}_i$ conditional on $\boldsymbol{X}_i = \boldsymbol{x}_i$ is given by (7), where all the parameters in the model are now functions of $\boldsymbol{x}_i$ except a_1 which is assumed fixed at some value e.g. $a_1 = 1$.

4.2.1 Preliminary transformation

Prior to estimation we suggest first replacing the observations $\boldsymbol{y}_i$ by $\tilde{\boldsymbol{T}}\boldsymbol{y}_i$ for $i = 1, \ldots, n$, where $\tilde{\boldsymbol{T}} = \boldsymbol{H}(p^{-1/2}\mathbf{1}_p)\tilde{\boldsymbol{\Gamma}}^\top$, and $\tilde{\boldsymbol{\Gamma}}$ is defined in Section 4.1.1. Note that, under mild conditions, $\tilde{\boldsymbol{T}}$ is a consistent estimator of its population analogue $\boldsymbol{T} = \boldsymbol{H}(p^{-1/2}\mathbf{1}_p)\boldsymbol{\Gamma}^\top$ and consequently, in large samples, the columns of $\boldsymbol{T}$ play an important role in the specification of the regression model if consistency holds. There is also the option of estimating $\boldsymbol{T}$ and $\boldsymbol{\Gamma}$ using maximum likelihood estimation, a possibility that deserves further investigation, but in this paper we have opted to use a simpler

approach, specifically the moment estimator for $\boldsymbol{T}$ indicated in Section 4.1.3. This transformation using $\tilde{\boldsymbol{T}}$ results in estimators and fitted values which are equivariant and is also convenient from a computational point of view, as typically $\boldsymbol{K}^*$ is not too far from the identity and the response data is centred as near as possible to the middle of the positive orthant. Centring the responses in this way helps to avoid any of the regression coefficients getting too close to infinity points on the link function scale (see below for further details).

4.2.2 Link functions

There are many different choices of link functions available to model the parameters in density (7). For convenience and comparative purposes we choose the same link functions as Scealy and Welsh (2017). However, an interesting topic for further research is the construction and exploration of other link functions. Let $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{x}_i)$ where

$$\mu_k(\boldsymbol{x}_i) = \begin{cases} \left(1 + \sum_{m=1}^{p-1} \exp\left(\boldsymbol{\beta}_m^\top \boldsymbol{x}_i\right)\right)^{-\frac{1}{2}} & k = 1 \\ \exp\left(\frac{\boldsymbol{\beta}_{k-1}^\top \boldsymbol{x}_i}{2}\right)\left(1 + \sum_{m=1}^{p-1} \exp\left(\boldsymbol{\beta}_m^\top \boldsymbol{x}_i\right)\right)^{-\frac{1}{2}} & k = 2, 3, ..., p, \end{cases}$$

where $\mu_k(\boldsymbol{x}_i)$ is the kth component of $\boldsymbol{\mu}(\boldsymbol{x}_i)$ and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \ldots, \boldsymbol{\beta}_{p-1}^\top\right)^\top \in \mathbb{R}^{q(p-1)}$ is a vector of regression coefficients. This model assumes that the mean direction is in the positive orthant. In many applications, including the paleomagnetic data example discussed in Section 5, it is reasonable to assume that the conditional mean direction $\boldsymbol{\mu}(\boldsymbol{x}_i)$ is not highly variable across the range of $\boldsymbol{x}_i$ and is contained well within the positive orthant after re-centring the data using the preliminary transformation discussed in Section 4.2.1.

For $p = 3$ let

$$\boldsymbol{V}(\boldsymbol{x}_i) = \begin{pmatrix} \sigma_3(1-c_1^2)^{-0.5}v_i^{2\delta_4} & c_1(1-c_1^2)^{-0.5} \\ c_1(1-c_1^2)^{-0.5} & \sigma_3^{-1}(1-c_1^2)^{-0.5}v_i^{-2\delta_4} \end{pmatrix} \quad \text{and} \quad \kappa(\boldsymbol{x}_i) = \sigma_4^{-1}v_i^{-2\delta_3}, \tag{14}$$

where $v_i = g(\boldsymbol{x}_i) \in \mathbb{R}$ is a known function and $\sigma_3 > 0$, $\sigma_4 > 0$, $\delta_3 \in \mathbb{R}$, $\delta_4 \in \mathbb{R}$ and $c_1 \in (-1, 1)$ are five variance component parameters. The above parameterisation (14) implies that

$$\frac{\boldsymbol{V}(\boldsymbol{x}_i)}{\kappa(\boldsymbol{x}_i)} = \sigma_1^2 \begin{pmatrix} v_i^{\delta_1} & 0 \\ 0 & \sigma_2 v_i^{\delta_2} \end{pmatrix} \begin{pmatrix} 1 & c_1^* \\ c_1^* & 1 \end{pmatrix} \begin{pmatrix} v_i^{\delta_1} & 0 \\ 0 & \sigma_2 v_i^{\delta_2} \end{pmatrix}, \tag{15}$$

where $\sigma_1 > 0$, $\sigma_2 > 0$, $\delta_1 \in \mathbb{R}$, $\delta_2 \in \mathbb{R}$ and $c_1^* \in (-1, 1)$ are five variance component parameters which satisfy $\delta_1 = \delta_3 + \delta_4$, $\delta_2 = \delta_3 - \delta_4$, $\sigma_2 = \sigma_3^{-1}$, $c_1^* = c_1$ and $\sigma_1^2 = \sigma_4\sigma_3(1-c_1^2)^{-0.5}$. The right hand side of (15) is the same covariance matrix structure used by Scealy and Welsh (2017) to model their Kent distribution second-order moment matrix. This is a standard general flexible heteroscedastic variance-covariance structure (e.g. Pinheiro and Bates, 2000, p. 205) and it can easily be extended into higher dimensions.

4.2.3 Estimation

The regression model parameters can be estimated directly by maximising the log-likelihood. The log-likelihood is given by

$$-n\log(a_1) - \sum_{i=1}^{n}\log c_p(\kappa(\boldsymbol{x}_i)) - \frac{(p-1)}{2}\sum_{i=1}^{n}\log\left(s_i^2\right) + \sum_{i=1}^{n}\frac{\kappa(\boldsymbol{x}_i)}{a_1}\boldsymbol{y}_i^\top\boldsymbol{\mu}(\boldsymbol{x}_i)\left(s_i^2\right)^{-1/2}, \tag{16}$$

where $s_i^2 = \left(\boldsymbol{y}_i^T\boldsymbol{\mu}(\boldsymbol{x}_i)\right)^2 a_1^{-2} + \boldsymbol{y}_i^T\boldsymbol{H}^*(\boldsymbol{\mu}(\boldsymbol{x}_i))\boldsymbol{V}(\boldsymbol{x}_i)^{-1}\boldsymbol{H}^*(\boldsymbol{\mu}(\boldsymbol{x}_i))^\top\boldsymbol{y}_i$. We suggest a two step iterative algorithm to maximise the above log-likelihood. First, calculate

a preliminary estimate of the regression coefficients $\boldsymbol{\beta}$ by solving for example the estimating equation (17) in Scealy and Welsh (2011). Then repeat the following two step algorithm until convergence of the parameters.

Step 1: Given $\boldsymbol{\beta}$, update the variance component parameters in $\kappa(\boldsymbol{x}_i)$ and $\boldsymbol{V}(\boldsymbol{x}_i)$ by maximising (16) with respect to these variance components.

Step 2: Given the variance components update from step 1, update $\boldsymbol{\beta}$ by maximising (16) with respect to $\boldsymbol{\beta}$.

A standard second derivative Newton-Raphson algorithm can be applied in each step to do the optimisations. Note that the derivatives of the modified Bessel function of the first kind can be calculated straightforwardly from known recurrence relations. Approximate standard errors for $\boldsymbol{\beta}$ can also be estimated directly by using the observed information matrix obtained from the second derivative matrix for $\boldsymbol{\beta}$ conditional on the other parameters (treating the variance components as fixed). Or alternatively, a bootstrap can be employed to calculate estimated standard errors by resampling the $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ pairs.

4.2.4 Equivariance

An important property of our new regression model is that the estimators are equivariant to orthogonal transformations. This is proved in Proposition 4 below.

Firstly, denote the original sample data by $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ and let $\tilde{\boldsymbol{\Gamma}}_y$ be the moment estimator of $\boldsymbol{\Gamma}$ for this data defined in Section 4.1.1. Also define

$$\tilde{\boldsymbol{Q}}_y = \boldsymbol{H}(p^{-1/2}1_p)\tilde{\boldsymbol{\Gamma}}_y^{\top}, \tag{17}$$

which is the orthogonal matrix given in Section 4.2.1. Now define $\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{Q}}_y\boldsymbol{y}_i$, $i = 1, \ldots, n$. We apply the regression modelling to the $\tilde{\boldsymbol{y}}_i$, not the $\boldsymbol{y}_i$. Suppose

that, after doing the regression modelling we end up with fitted mean directions $\hat{\boldsymbol{\mu}}(\boldsymbol{x}_1), \ldots, \hat{\boldsymbol{\mu}}(\boldsymbol{x}_n)$ for $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n$, respectively. If we wish to find the corresponding fitted mean directions in the original coordinate system for the $\boldsymbol{y}_i$, we calculate

$$\hat{\boldsymbol{y}}_1 = \tilde{\boldsymbol{Q}}_y^\top \hat{\boldsymbol{\mu}}(\boldsymbol{x}_1), \quad \hat{\boldsymbol{y}}_2 = \tilde{\boldsymbol{Q}}_y^\top \hat{\boldsymbol{\mu}}(\boldsymbol{x}_2), \quad \ldots, \quad \hat{\boldsymbol{y}}_n = \tilde{\boldsymbol{Q}}_y^\top \hat{\boldsymbol{\mu}}(\boldsymbol{x}_n).$$

We now state the equivariance result. The proof is given in Appendix A.5. In what follows the subscript y indicates quantities based on the $\boldsymbol{y}_i$, and a subscript w indicates quantities based on the $\boldsymbol{w}_i$, defined in the proposition below.

Proposition 4. *Suppose that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are unit p-vectors which span $\mathbb{R}^p$ and have a non-zero vector sum. Let $\boldsymbol{A}$ denote an arbitrary orthogonal $p \times p$ matrix and define $\boldsymbol{w}_i = \boldsymbol{A}\boldsymbol{y}_i$, $i = 1, \ldots, n$. Then there exists a choice $\tilde{\boldsymbol{\Gamma}}_w$ such that*

$$\tilde{\boldsymbol{\mu}}_w = \frac{\sum_{i=1}^n \boldsymbol{w}_i}{\| \sum_{i=1}^n \boldsymbol{w}_i \|};$$

the jth column of $\tilde{\boldsymbol{\Gamma}}_w$, $j = 2, \ldots, p$, are eigenvectors of

$$\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_w \tilde{\boldsymbol{\mu}}_w^\top\right) \left(\sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i^\top \right) \left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_w \tilde{\boldsymbol{\mu}}_w^\top\right),$$

corresponding to positive descending eigenvalues; and also

$$\tilde{\boldsymbol{\Gamma}}_w = \boldsymbol{A}\tilde{\boldsymbol{\Gamma}}_y.$$

Moreover, $\tilde{\boldsymbol{Q}}_w = \tilde{\boldsymbol{Q}}_y \boldsymbol{A}^\top$; the $\hat{\boldsymbol{\mu}}(\boldsymbol{x}_i)$ based on the $\tilde{\boldsymbol{w}}_i = \tilde{\boldsymbol{Q}}_w \boldsymbol{w}_i$, $i = 1, \ldots, n$, are invariant; and we have

$$\hat{\boldsymbol{w}}_i \equiv \tilde{\boldsymbol{Q}}_w^\top \hat{\boldsymbol{\mu}}(\boldsymbol{x}_i) = \boldsymbol{A}\hat{\boldsymbol{y}}_i, \quad i = 1, \ldots, n,$$

and consequently the fitted mean directions are equivariant with respect to orthogonal transformation $\boldsymbol{A}$.

There is a finite number of possible choices of $\tilde{\boldsymbol{\Gamma}}_w$, as of $\tilde{\boldsymbol{\Gamma}}_y$. In both cases, this number is 2^{p-1}, corresponding to sign changes of the $\tilde{\boldsymbol{\gamma}}_{j,w}$ and $\tilde{\boldsymbol{\gamma}}_{j,y}$, respectively and assuming distinct eigenvalues which occurs with probability one when $n \geq p$. However, if we require that $\tilde{\boldsymbol{\Gamma}}_w$ is continuous in $\boldsymbol{A}$ as $\boldsymbol{A}$ ranges over the $p \times p$ orthogonal matrices, then uniqueness in the choice of $\tilde{\boldsymbol{\Gamma}}_w$ is recovered, and this leads to the equivariance claimed in the proposition.

5 Analysis of paleomagnetic directional data

We now describe our analysis of the GMAG data discussed in the Introduction. For illustrative purposes we considered three further subsets of the data. Case 1 refers to a single time point, where the geological time variable Age (in years) is set equal to 1250; this is the time point with the most data, leading to a sample size of $n = 50$. Case 2 covers the Age range 0 to 1500, giving a sample of size $n = 788$; and Case 3 covers the Age range 1500 to 1900, giving a sample of size $n = 150$. We fitted the independent and indentically distributed model to Case 1 and the regression models to Case 2 and Case 3, with Age as the covariate. As a first step we calculated moment estimates for each of the three cases separately and then transformed the samples so that they were centred at the north pole using (13).

We now discuss our analysis of the Case 1 data. The top two plots and bottom left plot in Figure 2 contain kernel density estimates of the components $y_{2,i}$, $y_{3,i}$ and $y_{1,i}$, respectively. The top left plot shows that a model with heavy-tails may be needed. We interpret the bottom right scatterplot of $y_{3,i}$ versus $y_{2,i}$ as providing evidence that the contours of the underlying density are elliptical in shape. As a first

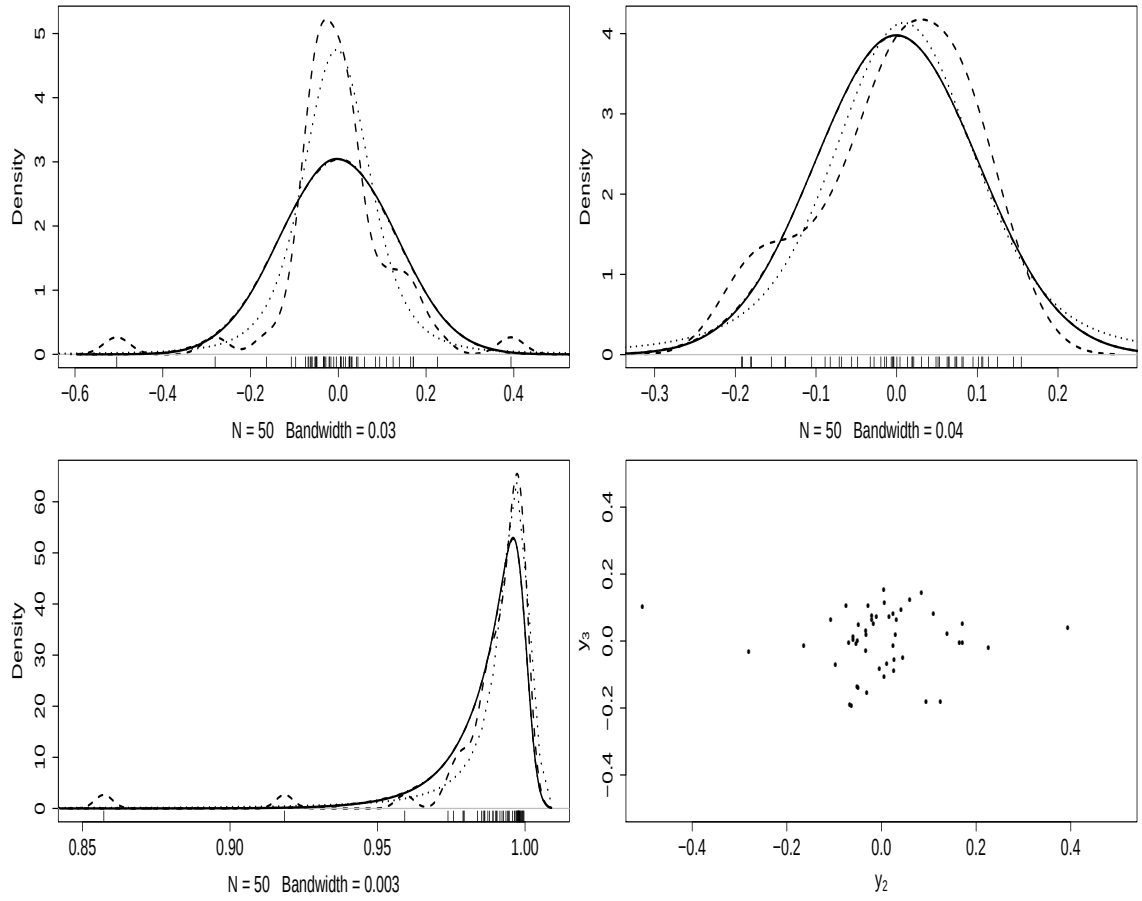

Figure 2: Case 1 data. Top left: $y_{2,i}$; top right: $y_{3,i}$; bottom left: $y_{1,i}$. Small vertical lines = data values, dashed line = kernel density estimate, solid line = fitted Kent density, dotted line = fitted $a_1 = 6$ model density and dot-dash line = fitted $a_1 = 1$ model density (close to the solid line). Bottom right: $y_{3,i}$ versus $y_{2,i}$ scatterplot.

step we fitted the Kent distribution to the data using maximum likelihood estimation coupled with a saddlepoint approximation for the shape parameters as in Scealy and Welsh (2014). We then simulated a sample of size $n = 100,000$ from the fitted Kent model and plotted the resulting nonparametric kernel density estimate (solid black line in Figure 2). We then fitted the distribution defined in Section 3 with $a_1 = 1$ and then $a_1 = 6$ using the estimators defined in Section 4.1. The parameter estimates for these models are the true values in Tables 3 and 4 used in the simulation experiment described in Section 6 and the standard errors in these tables can be considered as parametric bootstrap estimates. We simulated large samples from these two fitted models and in Figure 2 we plotted the resulting kernel density estimates. The value $a_1 = 6$ was chosen to give densities as close as possible to the observed sample marginal distributions of $y_{2,i}$ and $y_{3,i}$ based on making the Kolmogorov-Smirnov (KS) two sample test statistics small. The $p-$values for the KS test statistics for $y_{2,i}$ when $a_1 = 1$ and $a_1 = 6$ were 0.07 and 0.79 respectively. This gives evidence that the sample could have been generated from the Kent or $a_1 = 1$ distribution since the test statistic was borderline significant at the 5% level, but the heavy-tailed distribution with $a_1 = 6$ provided a better fit. For the component $y_{3,i}$, all of the KS test statistics were similar and non-significant, and there was little difference between the fitted distributions.

We now describe the analysis of the Case 2 and Case 3 data. Let $\tilde{x}_i$ represent the ith value of Age. For convenience we rescaled the covariate for each of the two cases separately as

$$x_i = \frac{\tilde{x}_i - \min \tilde{x}_i}{\max \tilde{x}_i - \min \tilde{x}_i} + 1.$$

In both Figure 3 and Figure 4, the top left, top right and middle left panels are plots of $y_{1,i}$ vs x_i, $y_{2,i}$ vs x_i and $y_{3,i}$ vs x_i, respectively. It is seen that there are non-linear relationships between $\boldsymbol{y}_i$ and x_i and the variability appears roughly to increase with

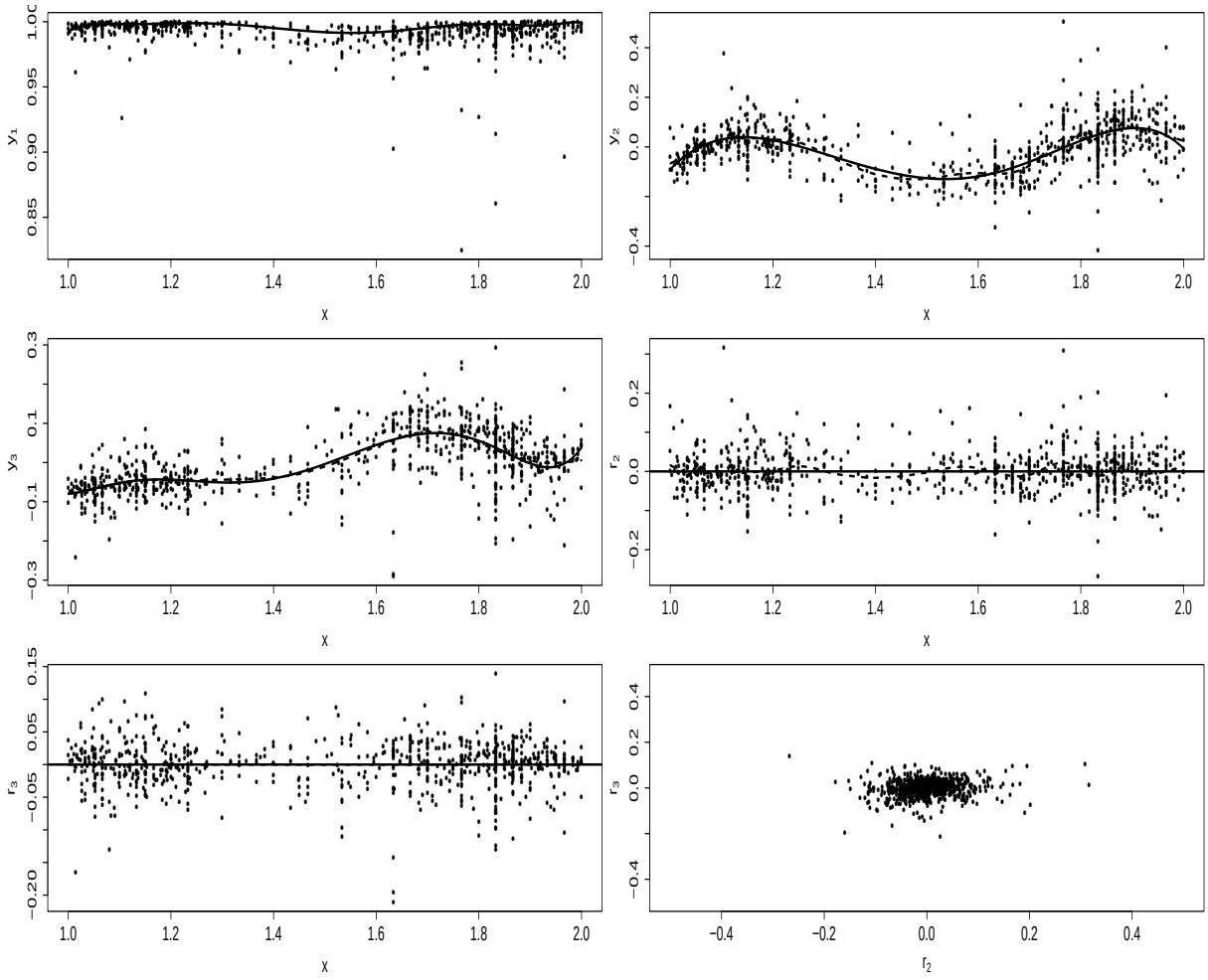

Figure 3: Case 2 data. Top left: $y_{1,i}$ versus x_i scatterplot; top right: $y_{2,i}$ versus x_i scatterplot; middle left: $y_{3,i}$ vs x_i scatterplot. Solid line is $a_1 = 1$ fitted values and dashed line is the cubic smoothing spline. Middle right: $r_{2,i}$ vs x_i scatterplot; bottom left: $r_{3,i}$ vs x_i scatterplot. Solid line is through the origin and dashed line is the cubic smoothing spline. Bottom right: $r_{3,i}$ vs $r_{2,i}$ scatterplot.

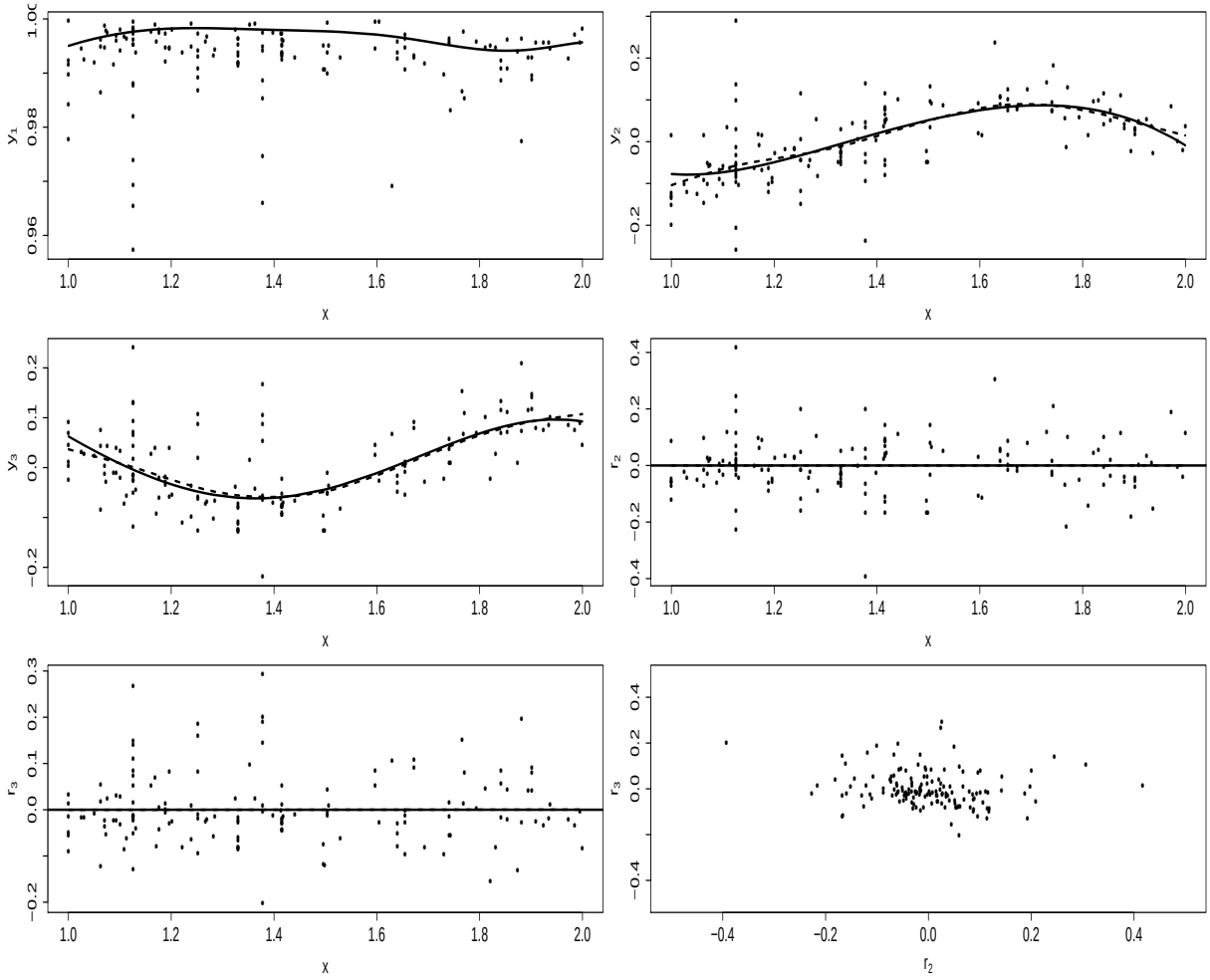

Figure 4: Case 3 data. This figure has the same structure as Figure 3 except that it is based on Case 3 rather than Case 2 data.

x_i for Case 2 and decrease with x_i for Case 3. Before fitting the new regression models, we further transformed the two samples so that they were centred in the middle of the positive orthant as follows

$$\boldsymbol{y}_i = (y_{1,i}, y_{2,i}, y_{3,i})^\top = \boldsymbol{H}(3^{-1/2}\mathbf{1}_3)\check{\boldsymbol{y}}_i, \quad i = 1, 2, \ldots, n,$$

where $\check{\boldsymbol{y}}_i$ are the samples centred at the north pole based on the preliminary transformation given at (13). We modelled the mean direction using a 6th degree polynomial and so the covariate vector is $\boldsymbol{x}_i = (1, x_i, x_i^2, x_i^3, x_i^4, x_i^5, x_i^6)^\top$ in $\boldsymbol{\mu}(\boldsymbol{x}_i)$. Some of the regression coefficients were not significant based on the size of the estimated standard errors and these were removed from the models (i.e. these regression coefficients were set to zero and are omitted in Tables 1 and 2).

Next, we fitted three different models to both the Case 2 and Case 3 data separately. First we fitted the Kent model defined in Scealy and Welsh (2017), but with the random effects omitted (this is equivalent to a fixed-effects only regression model). In this model there are two regression coefficient vectors $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{17})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}, \beta_{26}, \beta_{27})^\top$ and five variance component parameters. Similarly to Section 5 of Scealy and Welsh (2011), we obtained approximate maximum likelihood estimates of these parameters. In summary, to update the regression coefficients we maximised the Kent log-likelihood and to update the shape parameters we maximised the approximate Gaussian log-likelihood. We repeated these two steps until convergence. In this Kent model we parameterised $\mathrm{E}\left(\boldsymbol{H}^*(\boldsymbol{\mu}(\boldsymbol{x}_i))^T\boldsymbol{Y}_i\boldsymbol{Y}_i^\top\boldsymbol{H}^*(\boldsymbol{\mu}(\boldsymbol{x}_i))\right)$ using the right hand side of equation (15) with $v_i = x_i$, and from the large concentration asymptotics for the Kent distribution this expectation is approximately equal to

$$\boldsymbol{K}^*(\boldsymbol{x}_i)\mathrm{Diag}\left(\left(\check{\kappa}(\boldsymbol{x}_i) - 2\check{\beta}(\boldsymbol{x}_i)\right)^{-1}, \left(\check{\kappa}(\boldsymbol{x}_i) + 2\check{\beta}(\boldsymbol{x}_i)\right)^{-1}\right)\boldsymbol{K}^*(\boldsymbol{x}_i)^\top, \tag{18}$$

where $\check{\kappa}(\boldsymbol{x}_i)$ and $\check{\beta}(\boldsymbol{x}_i)$ are the Kent shape parameters for the ith unit. The parameter estimates we obtained for this model are given in Tables 1 and 2. We also include approximate standard errors for the regression coefficients based on the observed Kent information matrix conditional on the asymptotic approximations for the shape parameters as well as standard error estimates from the nonparametric bootstrap with 1000 resamples. With the bootstrap standard errors account is taken of the preliminary transformation but this is not the case with the standard errors based on observed information; nevertheless, these two types of standard errors are often in reasonable agreement.

We also fitted the new regression model defined in Section 4.2 with $a_1 = 1$ and $a_1 = 6$ to both the Case 2 and Case 3 data with the same covariate vector as the Kent model and with $\boldsymbol{V}(\boldsymbol{x}_i)$ and $\kappa(\boldsymbol{x}_i)$ parameterised by (14) with $v_i = x_i$. In Tables 1 and 2, covering Case 2 and Case 3 respectively, we present the following: parameter estimates; approximate standard errors for the regression coefficients based on the observed information matrix conditional on the variance component estimates; and standard error estimates obtained from the nonparametric bootstrap with 1000 resamples.

Table 1 shows that the Case 2 parameter estimates for $a_1 = 1$ and the Kent model are virtually identical. In all 3 models the bootstrap standard errors are similar in size to the observed Fisher information standard errors. The asymptotic Gaussian approximation appears to be working reasonably well here for the Kent distribution. All of the retained regression coefficients are significantly different from zero based on the size of the estimated standard errors. The estimates of the regression coefficients for $a_1 = 6$ are not the same as the $a_1 = 1$ case, although they are not significantly different based on the size of the estimated standard errors. The standard error estimates are smaller for the $a_1 = 6$ model and this is not surprising because it has

Table 1: Parameter and standard error estimates for Case 2

	estimates			standard errors					
				bootstrap			observed information		
	$a_1 = 1$	Kent	$a_1 = 6$	$a_1 = 1$	Kent	$a_1 = 6$	$a_1 = 1$	Kent	$a_1 = 6$
β_{11}	197	197	196	13.6	13.6	11.2	12.9	13.1	11.8
β_{12}	-554	-553	-550	38.6	38.5	31.7	36.7	37.1	33.4
β_{13}	572	572	568	40.3	40.2	33.0	38.4	38.8	34.9
β_{14}	-257	-257	-255	18.4	18.3	15.0	17.6	17.8	15.9
β_{15}	42.4	42.4	42.0	3.10	3.08	2.52	2.96	3.00	2.68
β_{21}	-776	-776	-803	246	246	215	260	264	236
β_{22}	3650	3650	3760	1040	1040	909	1110	1120	998
β_{23}	-6980	-6980	-7180	1800	1800	1590	1940	1970	1740
β_{24}	6950	6960	7140	1650	1650	1460	1790	1820	1610
β_{25}	-3820	-3820	-3920	843	844	748	922	938	823
β_{26}	1100	1100	1120	227	227	202	251	255	223
β_{27}	-128	-128	-132	25.2	25.2	22.5	28.1	28.6	24.9
σ_3	1.47	1.46	1.41	0.184	0.180	0.140			
c_1	0.117	0.113	0.136	0.0556	0.0559	0.0431			
δ_4	-0.0659	-0.0660	-0.0116	0.125	0.123	0.0999			
σ_4	0.00197	0.00197	0.0569	0.000240	0.000234	0.00489			
δ_3	0.959	0.943	0.626	0.151	0.147	0.0934			

heavier tails and accounts better for outliers. Note that the t-distribution with small degrees of freedom often gives smaller standard errors in models when compared with the Gaussian distribution (Lange et al., 1989). Based on the size of the bootstrap standard errors, there is evidence that σ_3 in (14) satisfies $\sigma_3 > 1$, implying that an elliptically symmetric model is needed. There is also evidence that δ_3 in (14) satisfies $\delta_3 > 0$, implying a heteroscedastic model is needed.

Table 2 shows that the Case 3 parameter estimates for $a_1 = 1$ and the Kent model are a lot more different than in Case 2. The bootstrap standard errors and the observed Fisher information standard errors for the $a_1 = 1$ model are similar in size, but for the Kent model the observed Fisher information standard errors are sometimes much larger than the bootstrap ones. We suspect that the observed information method is grossly overestimating the standard errors for the Kent model

Table 2: Parameter and standard error estimates for Case 3

	estimates			standard errors					
				bootstrap			observed information		
	$a_1 = 1$	Kent	$a_1 = 6$	$a_1 = 1$	Kent	$a_1 = 6$	$a_1 = 1$	Kent	$a_1 = 6$
β_{11}	-13.6	-8.54	-11.7	3.78	4.29	3.33	4.01	7.55	3.50
β_{12}	32.1	21.4	28.6	7.78	9.09	6.90	8.21	16.2	7.24
β_{13}	-23.8	-16.4	-21.7	5.23	6.27	4.67	5.50	11.3	4.90
β_{14}	5.54	3.91	5.14	1.15	1.41	1.03	1.21	2.59	1.08
β_{22}	-15.6	-10.4	-16.1	2.84	2.88	2.31	2.83	5.45	2.47
β_{23}	31.8	20.8	33.1	5.73	6.08	4.74	5.73	11.6	5.03
β_{24}	-20.6	-13.2	-21.7	3.78	4.18	3.17	3.79	8.04	3.35
β_{25}	4.32	2.67	4.55	0.816	0.937	0.694	0.823	1.83	0.731
σ_3	1.30	1.35	1.21	0.345	0.355	0.250			
c_1	-0.164	-0.144	-0.206	0.101	0.102	0.0977			
δ_4	-0.297	-0.367	-0.200	0.305	0.317	0.262			
σ_4	0.00781	0.00726	0.139	0.00213	0.00183	0.0280			
δ_3	-1.08	-0.959	-0.675	0.336	0.306	0.254			

because the shape parameter estimates are biased due to the asymptotic Gaussian approximation breaking down. Being able to approximate the standard errors well in both the $a_1 = 1$ and $a_1 = 6$ models using the observed information matrix is very useful because the bootstrap method is much more computationally intensive. Model selection for the terms in $\boldsymbol{\beta}$ can also be based on the approximate observed Fisher information standard errors, simplifying the analysis. Based on the size of the bootstrap standard errors, there is evidence that σ_3 in (14) satisfies $\sigma_3 = 1$, implying that a rotationally symmetric model is reasonable in this case. Similarly to Case 2, there is evidence that $\delta_3 \neq 0$, implying that a heteroscedastic model is needed. In both Cases 2 and 3 there is evidence that $\delta_4 = 0$ which implies that $\boldsymbol{V}$ may not depend on Age.

The solid lines in the top two plots and the middle left plot in Figures 3 and 4 were obtained by plotting $\boldsymbol{H}(p^{-1/2}\mathbf{1}_p)^\top \hat{\boldsymbol{\mu}}(\boldsymbol{x}_i)$ vs x_i (we transformed from the centre of the positive orthant back to the north pole), where $\hat{\boldsymbol{\mu}}(\boldsymbol{x}_i)$ is the predicted value from the fitted $a_1 = 1$ models (we also included a cubic smoothing spline for comparison

in the y_2 and y_3 plots). The model for the mean direction appears to fit the data reasonably well in both cases. To further check the fit of the $a_1 = 1$ model we also calculated the following standardised residuals:

$$\boldsymbol{r}_i = (r_{2,i}, r_{3,i})^\top = \text{Diag}(x_i^{-(\hat{\delta}_3+\hat{\delta}_4)}, x_i^{-(\hat{\delta}_3-\hat{\delta}_4)})\boldsymbol{H}^*(\hat{\boldsymbol{\mu}}(\boldsymbol{x}_i))^\top \boldsymbol{y}_i.$$

The middle right, bottom left and bottom right plots in Figures 3 and 4 contain plots of $r_{2,i}$ vs x_i, $r_{3,i}$ vs x_i and $r_{3,i}$ vs $r_{2,i}$, respectively (we also included a cubic smoothing spline for comparison in the first two plots). These plots show no obvious patterns and the residuals appear randomly dispersed about zero with constant variance.

6 Simulation

We simulated 1000 samples of size $n = 50$ from the following models fitted to the data of Case 1 in Section 5: (i) Distribution defined in Section 3 with $a_1 = 1$ ($\mathcal{P}_1$), and (ii) distribution defined in Section 3 with $a_1 = 6$ ($\mathcal{P}_6$). For each sample we fitted both models and calculated five different estimates of $\boldsymbol{\mu}$ using (a) moment estimator, (b) normalised spatial median, (c) maximum likelihood estimator obtained using the Nelder-Mead algorithm under the $\mathcal{P}_6$ model, (d) maximum likelihood estimator obtained using the Nelder-Mead algorithm under the $\mathcal{P}_1$ model, and (e) maximum likelihood estimator for the Kent model with shape parameters estimated via the asymptotic Gaussian approximation. We also calculated the maximum likelihood estimates of κ and $\boldsymbol{V}$ for each sample under the true model. Table 3 contains the estimated standard errors and true values of $\boldsymbol{\mu}$ conditioned on in the simulations (the estimated standard errors are parametric bootstrap estimates for the Case 1 data in Section 5). The estimated biases in the mean direction estimators were all negligible and the standard errors and root mean squared errors were all very similar. Table

4 contains the true values of κ and $\boldsymbol{V}$ conditioned on in the simulations as well as standard error and bias estimates, where V_{ij} denotes the (i,j)th element in $\boldsymbol{V}$.

In Table 3 there are no results for μ_1 because we have the identity $\mu_1^2+\mu_2^2+\mu_3^2=1$ and μ_1 is determined up to sign from μ_2 and μ_3. It is best to give results for μ_2 and μ_3 only because the marginal distributions of y_2 and y_3 are centred close to zero and are approximately symmetric, whereas the marginal distributions of y_1 are highly left skewed due to being distributed close to the upper boundary 1. Biases and standard errors are less informative in this asymmetric case. The standard errors suggest that μ_2 and μ_3 could both be zero. This is due to the fact that we applied the prior orthogonal transformation before the analysis to guarantee equivariance (this transformation recentred the data so that the sample mean direction is at the north pole).

Table 3: Estimated standard errors for the $\boldsymbol{\mu}=(\mu_1,\mu_2,\mu_3)^\top$ estimators

			$\mathcal{P}_1$			
parameter	true value	(a)	(b)	(c)	(d)	(e)
μ_2	-0.0006	0.018	0.020	0.020	0.018	0.018
μ_3	0.0002	0.013	0.015	0.015	0.013	0.013
			$\mathcal{P}_6$			
parameter	true value	(a)	(b)	(c)	(d)	(e)
μ_2	-0.0039	0.018	0.012	0.011	0.021	0.023
μ_3	0.0096	0.020	0.014	0.013	0.025	0.028

From Table 3 we see that the moment estimator has similar efficiency to the Kent and $\mathcal{P}_1$ maximum likelihood estimator when simulating under $\mathcal{P}_1$. The normalised spatial median and the $\mathcal{P}_6$ maximum likelihood estimator are slightly less efficient. When simulating under $\mathcal{P}_6$, the normalised spatial median and $\mathcal{P}_6$ maximum likelihood estimator of the mean direction were more efficient than the $\mathcal{P}_1$ maximum likelihood estimator, the moment estimator and the Kent maximum likelihood estimator. The normalised spatial median is only slightly less efficient than the $\mathcal{P}_6$

Table 4: Bias and standard error estimates for the κ and $\boldsymbol{V}$ estimators

$\mathcal{P}_1$				
parameter	κ	V_{11}	V_{12}	V_{22}
true value	84.31	1.39	0.0029	0.7210
bias	4.64	0.02	0.0017	0.018
standard error	12.93	0.20	0.15	0.11
$\mathcal{P}_6$				
parameter	κ	V_{11}	V_{12}	V_{22}
true value	5.09	0.91	0.088	1.10
bias	0.11	0.029	0.000	0.029
standard error	0.79	0.16	0.17	0.19

maximum likelihood estimator when simulating under $\mathcal{P}_6$. Table 4 shows that $\boldsymbol{V}$ is not significantly different from the identity matrix when simulating under $\mathcal{P}_6$ and there is evidence that the true model could be rotationally symmetric. This is not the case for $\mathcal{P}_1$ as both V_{11} and V_{22} appear marginally significantly different from 1 based on the size of the standard errors in Table 4.

7 Conclusion

We introduced a flexible heteroscedastic regression model for paleomagnetic directional data. The error distribution, which is obtained via a novel transformation of the von Mises-Fisher distribution, has some desirable properties. Specifically, the error density has elliptical symmetry; and its normalising constant is tractable, so that the shape parameters can be estimated directly using maximum likelihood estimation. The new model was successfully applied to the analysis of paleomagnetic data in the GEOMAGIA50.v3 database. It is evident from our analysis that there is significant heteroscedasticity in the data and that the new regression model provides a useful framework which captures non-linear features in the data. Moreover, the model has a tuning parameter that enables the accommodation of both light-tailed and heavy-tailed directional data.

Acknowledgments

Both authors are grateful to EPSRC for supporting this research through grant EP/K022547/1 and the first author was additionally supported by an Australian Research Council Discovery Early Career Researcher Award. We thank Andrew P. Roberts, David Heslop and Sanja Panovska for their assistance with the paleomagnetic data and useful conversations. We also thank three referees and an Associate Editor for their detailed reviews which have led to an improved paper.

A Proofs

A.1 Proof of Lemma 2.1

Suppose initially that $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$, where $\boldsymbol{v} = r\boldsymbol{t}$, $\boldsymbol{u} = \rho\boldsymbol{\tau}$, $r = ||\boldsymbol{v}||$, $\boldsymbol{t} = \boldsymbol{v}/r$, $\rho = ||\boldsymbol{u}||$ and $\boldsymbol{\tau} = \boldsymbol{u}/\rho$. Suppose also that $v_i = u_i/a_i$ for $i = 1, \ldots, p$. Then consider the sequence of transformations $(r, \boldsymbol{t}) \to \boldsymbol{v} \to \boldsymbol{u} \to (\rho, \boldsymbol{\tau})$. Ignoring sets of Lebesgue measure zero, these transformations are all bijections. Then, using $d\boldsymbol{t}$ and $d\boldsymbol{\tau}$ to denote unnormalised geometric measure on the unit sphere $\mathcal{S}^{p-1}$, and using the standard facts that $drd\boldsymbol{t} = \frac{1}{||\boldsymbol{v}||^{p-1}} d\boldsymbol{v}$ and $d\boldsymbol{u} = \rho^{p-1} d\rho d\boldsymbol{\tau}$, we obtain

$$
\begin{aligned}
drd\boldsymbol{t} &= \frac{1}{||\boldsymbol{v}||^{p-1}} d\boldsymbol{v} \\
&= \left\{ \sum_{i=1}^{p} \left(\frac{u_i}{a_i} \right)^2 \right\}^{-(p-1)/2} \left(\prod a_i^{-1} \right) d\boldsymbol{u} \\
&= \left\{ \sum_{i=1}^{p} \left(\frac{\rho\tau_i}{a_i} \right)^2 \right\}^{-(p-1)/2} \left(\prod a_i \right)^{-1} \rho^{p-1} d\rho d\boldsymbol{\tau} \\
&= J_a(\boldsymbol{\tau}) d\rho d\boldsymbol{\tau}.
\end{aligned}
$$

Now restrict attention to $\rho = 1$. Under this restriction, $\boldsymbol{u} = \boldsymbol{\tau}$ and Lemma 2.1 follows.

A.2 Proof of Proposition 1

Write $\boldsymbol{y}^\top \boldsymbol{\mu} = \cos(\theta)$. Then

$$\inf_{\boldsymbol{y}:\, \boldsymbol{y}^\top \boldsymbol{\mu} = \cos(\theta)} \sum_{j=1}^{p} \left(\frac{y_j}{a_j}\right)^2 = \frac{1}{a_1^2}\cos^2(\theta) + \frac{1}{a_2^2}\sin^2(\theta).$$

Consequently, any local mode will occur at a θ which maximises $f_Y(\boldsymbol{y})$ on the great circle $\boldsymbol{y} = \cos(\theta)\boldsymbol{\mu} + \sin(\theta)\boldsymbol{\gamma}_2$. On this great circle, $\log f_Y(\boldsymbol{y})$, written as a function of θ and with the normalising constant excluded, is given by

$$\log f_Y(\boldsymbol{y}) = -\frac{(p-1)}{2}\log\left(\cos^2(\theta) + \frac{a_1^2\sin^2(\theta)}{a_2^2}\right) + \frac{\kappa\cos(\theta)}{\left(\cos^2(\theta) + \frac{a_1^2\sin^2(\theta)}{a_2^2}\right)^{1/2}}. \quad (19)$$

Differentiating (19) with respect to θ and rearranging, we obtain

$$\frac{a_1^2\sin(\theta)}{a_2^2\left(\cos^2(\theta) + \frac{a_1^2\sin^2(\theta)}{a_2^2}\right)^{3/2}}\left[a_1(p-1)\left(\frac{a_2^2}{a_1^2} - 1\right)\cos(\theta)\left(\frac{\cos^2(\theta)}{a_1^2} + \frac{\sin^2(\theta)}{a_2^2}\right)^{1/2} - \kappa\right].$$

The expression inside the square bracket is negative for all $\theta \in [0, \pi]$ when $\kappa > a_1(p-1)\left((a_2/a_1)^2 - 1\right)$ and $a_1 \leq a_2$, in which case the unique mode of $f_Y(\boldsymbol{y})$ is at $\theta = 0$, i.e. $\boldsymbol{y} = \boldsymbol{\mu}$. In the case when $a_1 = 1$ and $\kappa = a_1(p-1)\left((a_2/a_1)^2 - 1\right)$, the expression inside the square bracket is zero at $\theta = 0$ but negative for all $\theta \in (0, \pi]$, and as a consequence there is still a unique mode at $\boldsymbol{y} = \boldsymbol{\mu}$. In both of these cases the density is unimodal.

When $\kappa > 0$ and $a_1 > a_2$ the expression inside the square bracket is negative for all $\theta \in [0, \frac{\pi}{2}]$, in which case there is a unique mode of $f_Y(\boldsymbol{y})$ at $\theta = 0$ on the interval

$\theta \in [0, \frac{\pi}{2}]$, i.e. $\boldsymbol{y} = \boldsymbol{\mu}$. When $\kappa > 0$ and $a_1 > a_2$ the expression inside the square bracket may be negative, zero or positive for $\theta \in [\frac{\pi}{2}, \pi]$ and it is possible for there to be a maximum on the interior of this interval close to $\theta = \pi$ (resulting in a bimodal distribution over the entire interval $[0, \pi]$). Now consider point $\theta_1 \in [0, \frac{\pi}{2}]$ and its matching point $\theta_2 = \pi - \theta_1 \in [\frac{\pi}{2}, \pi]$. When $\kappa > 0$ the function (19) is always larger for $\theta = \theta_1$ than it is for $\theta = \theta_2$ and hence the global maximum of $f_Y(\boldsymbol{y})$ on the entire interval $\theta \in [0, \pi]$ is at $\theta = 0$. Therefore when $\kappa > 0$ and $a_1 > a_2$ there is still a global maximum at $\boldsymbol{y} = \boldsymbol{\mu}$, but the distribution is not unimodal in general.

A.3 Proof of Proposition 2

Since the transformation (1) is scale invariant an equivalent form for density (6) is obtained by replacing a_1 by 1 and each a_j by a_j/a_1 for $j = 2, 3, \ldots, p$ (this replacement also needs to occur in the Jacobian term). Then, similar to Scealy and Welsh (2011, p357), let $y_1^* = 1 - \parallel \boldsymbol{z}^* \parallel^2 /2$ and $y_j^* = (1 - \parallel \boldsymbol{z}^* \parallel^2 /4)^{1/2} z_j^*$ for $j = 2, 3, \ldots, p$, where $\boldsymbol{z}^* = (z_2^*, z_3^*, \ldots, z_p^*)^\top$ and it follows that $\parallel \boldsymbol{z}^* \parallel \leq 2$. For this transformation $d\boldsymbol{y}^* = (1 - \parallel \boldsymbol{z}^* \parallel^2 /4)^{(p-3)/2} d\boldsymbol{z}^*$. Hence the density of $\boldsymbol{z}^*$ is

$$\{c_p(\kappa)\}^{-1} a_1^{p-1} \left\{ \left(1 - \parallel \boldsymbol{z}^* \parallel^2 /2\right)^2 + \sum_{j=2}^{p} \left(\left(1 - \parallel \boldsymbol{z}^* \parallel^2 /4\right)^{1/2} z_j^* a_1/a_j \right)^2 \right\}^{-(p-1)/2}$$
$$\times \exp \left\{ \frac{\kappa \left(1 - \parallel \boldsymbol{z}^* \parallel^2 /2\right)}{\left\{ \left(1 - \parallel \boldsymbol{z}^* \parallel^2 /2\right)^2 + \sum_{j=2}^{p} \left(\left(1 - \parallel \boldsymbol{z}^* \parallel^2 /4\right)^{1/2} z_j^* a_1/a_j \right)^2 \right\}^{1/2}} \right\}$$
$$\times \left(1 - \parallel \boldsymbol{z}^* \parallel^2 /4\right)^{(p-3)/2}.$$

Now let $\boldsymbol{v} = (v_2, v_3, \ldots, v_p)^\top$, where $v_j = \kappa^{1/2} z_j^*$, $j = 2, 3, \ldots, p$. When κ is large, it follows that the density of $\boldsymbol{v}$ is

$$\begin{aligned} &\left\{c_p(\kappa)\right\}^{-1} a_1^{p-1} \kappa^{-(p-1)/2} \left\{\left(1 + O_p\left(\kappa^{-1}\right)\right)\right\}^{-(p-1)/2} \\ &\qquad \times \exp\left\{ \frac{\left(\kappa - \parallel \boldsymbol{v}^* \parallel^2 /2\right)}{\left\{1 - \kappa^{-1} \parallel \boldsymbol{v}^* \parallel^2 + \kappa^{-1} \sum_{j=2}^{p} \left(v_j^* a_1 / a_j\right)^2 + O_p(\kappa^{-2})\right\}^{1/2}} \right\} \qquad (20) \\ &\qquad\qquad \times \left(1 - O_p\left(\kappa^{-1}\right)\right)^{(p-3)/2}. \end{aligned}$$

By a Taylor series expansion, the exponential term in (20) simplifies to

$$\exp\left\{ \left(\kappa - \parallel \boldsymbol{v}^* \parallel^2 /2\right) \left(1 + (2\kappa)^{-1} \parallel \boldsymbol{v}^* \parallel^2 - (2\kappa)^{-1} \sum_{j=2}^{p} \left(v_j^* a_1 / a_j\right)^2 + O_p(\kappa^{-2}) \right) \right\},$$

which is equivalent to

$$\exp\left\{ \kappa - (2)^{-1} \sum_{j=2}^{p} \left(v_j^* a_1 / a_j\right)^2 + O_p(\kappa^{-1}) \right\}.$$

The term $\left\{c_p(\kappa)\right\}^{-1} \kappa^{-(p-1)/2} \exp\{\kappa\}$ simplifies to

$$(2\pi)^{-p/2} \left(I_{(p/2)-1}(\kappa)\right)^{-1} \kappa^{-1/2} \exp\{\kappa\} = (2\pi)^{-(p-1)/2} \left(1 + O(\kappa^{-1})\right),$$

since $I_\nu(\kappa) = (2\pi)^{-1/2} \kappa^{-1/2} \exp\{\kappa\} (1 + O(\kappa^{-1}))$ (e.g. Mardia and Jupp 2000, p. 349). The density then converges to

$$(2\pi)^{-(p-1)/2} a_1^{p-1} \exp\left\{ -(2)^{-1} \sum_{j=2}^{p} \left(v_j^* a_1 / a_j\right)^2 \right\},$$

the $(p-1)$ dimensional Gaussian density with mean 0 and diagonal covariance matrix. The variables v_j for $j = 2, 3, \ldots, p$ are each $O_p(1)$ and therefore z_j^* for $j = 2, 3, \ldots, p$ is

$O_p(\kappa^{-1/2})$. By definition, $\kappa^{1/2}\boldsymbol{y}_L^* = \kappa^{1/2}\left(1 - \parallel \boldsymbol{z}^* \parallel^2 /4\right)^{1/2}\boldsymbol{z}^* = \kappa^{1/2}\boldsymbol{z}^* + \boldsymbol{O}_p(\kappa^{-1})$ and therefore $\mid \kappa^{1/2}\boldsymbol{z}^* - \kappa^{1/2}\boldsymbol{y}_L^* \mid \to 0$ in probability and $\boldsymbol{y}_L^*$ also has the same asymptotic Gaussian distribution as $\boldsymbol{z}^*$.

A.4 Proof of Proposition 3

By definition $\boldsymbol{y} = \boldsymbol{\Gamma}\boldsymbol{y}^*$, implying $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\Gamma}\,\mathrm{E}(\boldsymbol{y}^*)$ and $\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^\top) = \boldsymbol{\Gamma}\,\mathrm{E}(\boldsymbol{y}^*\boldsymbol{y}^{*\top})\boldsymbol{\Gamma}^\top$. Define $\boldsymbol{y}_L^* = (y_2^*, y_3^*, \ldots, y_p^*)^\top$. From symmetry arguments we observe that $\mathrm{E}(y_1^*) > 0$, $\mathrm{E}(\boldsymbol{y}_L^*) = \boldsymbol{0}_{p-1}$, $\mathrm{E}(y_1^*\boldsymbol{y}_L^*) = \boldsymbol{0}_{p-1}$ and $\mathrm{E}(y_m^* y_r^*) = 0$ for $m \neq r$, $m = 2, 3, \ldots, p$ and $r = 2, 3, \ldots, p$ and the result then follows.

A.5 Proof of Proposition 4

First note that

$$\tilde{\boldsymbol{\mu}}_w = \frac{\sum_{i=1}^n \boldsymbol{w}_i}{\parallel \sum_{i=1}^n \boldsymbol{w}_i \parallel} = \boldsymbol{A}\tilde{\boldsymbol{\mu}}_y.$$

Moreover, since $\boldsymbol{A}$ is an orthogonal matrix,

$$\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_w\tilde{\boldsymbol{\mu}}_w^\top = \boldsymbol{A}\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_y\tilde{\boldsymbol{\mu}}_y^\top\right)\boldsymbol{A}^\top,$$

and therefore

$$\begin{aligned}
&\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_w\tilde{\boldsymbol{\mu}}_w^\top\right)\left(\sum_{i=1}^n \boldsymbol{w}_i\boldsymbol{w}_i^\top\right)\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_w\tilde{\boldsymbol{\mu}}_w^\top\right) \\
&\qquad = \boldsymbol{A}\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_y\tilde{\boldsymbol{\mu}}_y^\top\right)\boldsymbol{A}^\top\boldsymbol{A}\left(\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^\top\right)\boldsymbol{A}^\top\boldsymbol{A}\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_y\tilde{\boldsymbol{\mu}}_y^\top\right)\boldsymbol{A}^\top \\
&\qquad = \boldsymbol{A}\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_y\tilde{\boldsymbol{\mu}}_y^\top\right)\left(\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^\top\right)\left(\boldsymbol{I}_p - \tilde{\boldsymbol{\mu}}_y\tilde{\boldsymbol{\mu}}_y^\top\right)\boldsymbol{A}^\top
\end{aligned}$$

Consequently, the first part of the proposition holds and, in particular, we may choose $\tilde{\boldsymbol{\Gamma}}_w = \boldsymbol{A}\tilde{\boldsymbol{\Gamma}}_y$. It then follows directly from (17) that $\tilde{\boldsymbol{Q}}_w = \tilde{\boldsymbol{Q}}_y\boldsymbol{A}^\top$;

$$\tilde{\boldsymbol{w}}_i = \tilde{\boldsymbol{Q}}_w\boldsymbol{w}_i = \tilde{\boldsymbol{Q}}_y\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{y}_i = \tilde{\boldsymbol{Q}}_y\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i, \quad i = 1, \ldots, n,$$

so that, in particular, the $\hat{\boldsymbol{\mu}}(\boldsymbol{x}_i)$ are invariant (as opposed to equivariant); and, finally,

$$\hat{\boldsymbol{w}}_i = \tilde{\boldsymbol{Q}}_w^\top\hat{\boldsymbol{\mu}}(\boldsymbol{x}_i) = \boldsymbol{A}\tilde{\boldsymbol{Q}}_y^\top\hat{\boldsymbol{\mu}}(\boldsymbol{x}_i) = \boldsymbol{A}\hat{\boldsymbol{y}}_i, \quad i = 1, \ldots, n,$$

as required.

References

Acton, G. D., Galbrun, B. and King, J. W. (2000). Paleolatitude of the Caribbean plate since the late Cretaceous. *Proceedings of the Ocean Drilling Program, Scientific Results* **165** 149–173.

Borradaile, G. (2003). *Statistics of Earth Science Data*. Springer-Verlag, Berlin.

Brown, M. C., Donadini, F., Korte, M., Nilsson, A., Korhonen, K., Lodge, A., Lengyel, S. N., Constable, C. G. (2015). GEOMAGIA50.v3: 1. general structure and modifications to the archeological and volcanic database. *Earth, Planets and Space* **67:83**

Di Marzio, M., Panzera, A. and Taylor, C. C. (2014). Nonparametric regression for spherical data. *Journal of the American Statistical Association* **109** 748-763.

Downs, T. D. and Mardia, K. V. (2002). Circular regression. *Biometrika* **89** 683–697.

Fisher, N. I., Lewis, T. and Embleton, B. J. J. (1987). *Statistical Analysis of*

Spherical Data. Cambridge University Press, Cambridge.

Jones, M. C. and Pewsey, A. (2005). A family of symmetric distributions on the circle. *Journal of the American Statistical Association* **100** 1422–1428.

Jupp, P. E. and Kent, J. T. (1987). Fitting smooth paths to spherical data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **36** 34–46.

Kato, S. and Jones, M. C. (2010). A family of distributions on the circle with links to, and applications arising from, Möbius transformation. *Journal of the American Statistical Association* **105** 249-262.

Kato, S. and McCullagh, P. (2015). Conformal mapping for multivariate Cauchy families. *arXiv.1510.07679v.*

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society, Series B* **44** 71–80.

Kent, J. T., Mardia, K. V. and McDonnell, P. (2006). The complex Bingham quartic distribution and shape analysis. *Journal of the Royal Statistical Society, Series B* **68**, 747-765.

Ko, D. and Chang, T. (1993). Robust M-estimators on spheres. *Journal of Multivariate Analysis* **45** 104–136.

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

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84** 881–896.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley, Chichester.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics* **4** 51–67.

Noel, M. and Batt, C. M. (1990). A method for correcting geographically sepa-

rated remanence directions for the purpose of archaeomagnetic dating. *Geophysical Journal International* **102** 753–756.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimisation. *Computer Journal* **7** 308–313.

Paine, P. J., Preston, S. P., Tsagris, M. and Wood, A. T. A. (2018). An elliptically symmetric angular Gaussian distribution. *Statistics and Computing* **28** 689–697.

Panovska, S., Korte. M., Finlay, C. and Constable, C. G. (2015). Limitations in paleomagnetic data and modelling techniques and their impact on Holocene geomagnetic field models. *Geophysical Journal International* **202** 402–418.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Rivest, L. (1984). On the information matrix for symmetric distributions on the hypersphere. *Annals of Statistics* **12** 1085–1089.

Rivest, L., Duchesne, T., Nicosia, A. and Fortin, D. (2016). A general angular regression model for the analysis of data on animal movement in ecology. *Journal of the Royal Statistical Society Series C* **65** 445-463.

Scealy, J. L. and Welsh, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society, Series B* **73** 351–375.

Scealy, J. L. and Welsh, A. H. (2014). Fitting Kent models to compositional data with small concentration. *Statistics and Computing* **24** 165–179.

Scealy, J. L. and Welsh, A. H. (2017). A directional mixed effects model for compositional expenditure data. *Journal of the American Statistical Association* **112** 24–36.

Schuenemeyer, J. H. and Drew, L. J. (2011). *Statistics for Earth and Environ-*

mental Scientists. John Wiley & Sons, Hoboken, New Jersey.

Tauxe, L. (2010). *Essentials of Paleomagnetism*. University of California Press, Oakland, California.

Taylor, J. M. G. (1992). Properties of modelling the error distribution with an extra shape parameter. *Computational Statistics and Data Analysis* **13** 33–46.

Walker, M. R. and Jackson, A. (2000). Robust modelling of the Earth's magnetic field. *Geophysical Journal International* **143** 799–808.

Watson, G. S. (1983). *Statistics on Spheres*. Wiley, New York.