Statistics and Computing
https://doi.org/10.1007/s11222-019-09872-2

Spherical regression models with general covariates and anisotropic errors

P. J. Paine[1] · S. P. Preston[2] · M. Tsagris[3] · Andrew T. A. Wood[2]

Received: 5 July 2018 / Accepted: 2 April 2019
© The Author(s) 2019

Abstract
Existing parametric regression models in the literature for response data on the unit sphere assume that the covariates have particularly simple structure, for example that they are either scalar or are themselves on the unit sphere, and/or that the error distribution is isotropic. In many practical situations, such models are too inflexible. Here, we develop richer parametric spherical regression models in which the covariates can have quite general structure (for example, they may be on the unit sphere, in Euclidean space, categorical or some combination of these) and in which the errors are anisotropic. We consider two anisotropic error distributions—the Kent distribution and the elliptically symmetric angular Gaussian distribution—and two parametrisations of each which enable distinct ways to model how the response depends on the covariates. Various hypotheses of interest, such as the significance of particular covariates, or anisotropy of the errors, are easy to test, for example by classical likelihood ratio tests. We also introduce new model-based residuals for evaluating the fitted models. In the examples we consider, the hypothesis tests indicate strong evidence to favour the novel models over simpler existing ones.

Keywords Angular Gaussian distribution · Kent distribution · Model selection · Residuals · Spherical data

1 Introduction

Spherical data are observations that lie on the unit sphere $\mathbb{S}^{p-1} = \{\mathbf{y} \in \mathbb{R}^p : \mathbf{y}^\top \mathbf{y} = 1\}$. They arise in many scientific disciplines, including shape analysis, geology and meteorology [e.g. Mardia and Jupp (2000)] and more recently areas as diverse as genome sequence representations and text analysis [e.g. Hamsici and Martinez (2007)]. In this paper, we consider the regression problem in which the data are pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$, $i = 1, \ldots, n$, involving a $q \times 1$ covariate vector, $\mathbf{x}_i$, and a spherical response variable, $\mathbf{y}_i \in \mathbb{S}^2$. The aim of regression modelling is to establish how the response variable $\mathbf{y}_i$ depends on $\mathbf{x}_i$.

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/KO22547/1].

✉ S. P. Preston
simon.preston@nottingham.ac.uk

[1] School of Mathematics and Statistics, University of Sheffield, Hicks Building, Hounsfield Road, Sheffield S3 7RH, UK

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

[3] Department of Economics, University of Crete, 74100 Rethymnon, Greece

Typical parametric regression models currently in use for spherical responses in dimension $p \geq 3$ are fairly restrictive in the sense that (i) the covariates are assumed to have special structure, e.g. that the covariate is a scalar (such as time) or is itself on the sphere (i.e. a direction); and/or (ii) the models assume isotropic error distributions. Examples of (i) and (ii) in the literature are Chang (1986), Rivest (1989) and Rosenthal et al. (2014), see also Di Marzio et al. (2014) in a nonparametric context. Recent work in regression modelling on general Riemannian manifolds, for which the unit sphere is a special case, includes the nonparametric approach of Lin et al. (2017), who develop local regression models assuming Euclidean covariates, and the semi-parametric approach of Cornea et al. (2017), who use parametric link functions mapping from a general covariate space to the manifold, with a nonparametric error distribution; though in neither is the possibility of anisotropic errors explicitly considered.

The principal contribution of this paper is to introduce parametric regression models for spherical response data that relax both (i) and (ii). The motivation for doing so is that in many applications the covariates do not have the simple structure described in (i), and that there is rarely any basis for assuming a priori that the error distribution is isotropic.

Published online: 13 April 2019

There are two main ingredients of the spherical regression models we develop: a distribution on the sphere, to play the role of an error distribution, and a structural model linking the parameters of this error distribution to the covariates. Our approach is similar in spirit to generalised linear models in the sense that we express parameters of the distribution of $\mathbf{y}_i$ in terms of $\mathbf{B}\mathbf{x}_i$, where $\mathbf{B}$ is a matrix of parameters. Two simple distributions on the sphere, each broadly analogous to the isotropic normal distribution in $\mathbb{R}^2$, are the von Mises–Fisher and isotropic angular Gaussian (IAG) distributions. Both are "isotropic" (or equivalently "rotationally symmetric") on the sphere at the mean direction, $\tilde{\boldsymbol{\mu}} \in \mathbb{S}^2$, meaning that their contours are small circles centred on $\tilde{\boldsymbol{\mu}}$. The von Mises–Fisher distribution arises from conditioning an isotropic multivariate normal random variable $\mathbf{z} \in \mathbb{R}^p$ to have unit norm. On $\mathbb{S}^2$, to which we specialise henceforth, it is often called the Fisher distribution. It has three free parameters: two to define the mean direction, $\tilde{\boldsymbol{\mu}} \in \mathbb{S}^2$, and another scalar parameter, $\kappa > 0$, that controls concentration. Its density function on $\mathbb{S}^2$ is

$$f_{\text{Fisher}}(\mathbf{y}|\kappa, \tilde{\boldsymbol{\mu}}) = \frac{\kappa}{4\pi \sinh(\kappa)} \exp\left(\kappa \mathbf{y}^\top \tilde{\boldsymbol{\mu}}\right). \tag{1}$$

The IAG distribution arises from projecting (as opposed to conditioning) $\mathbf{z}$ to lie on $\mathbb{S}^{p-1}$. On $\mathbb{S}^2$ its density function is

$$f_{\text{IAG}}(\mathbf{y}|\boldsymbol{\mu}) = \frac{1}{2\pi} \exp\left[\frac{1}{2}\left\{\left(\mathbf{y}^\top \boldsymbol{\mu}\right)^2 - \boldsymbol{\mu}^\top \boldsymbol{\mu}\right\}\right] M\left(\mathbf{y}^\top \boldsymbol{\mu}\right), \tag{2}$$

where $M(\alpha) = \alpha\phi(\alpha) + (1 + \alpha^2)\Phi(\alpha)$, and where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal probability density function and cumulative distribution function, respectively. It is parametrised by the vector $\boldsymbol{\mu} \in \mathbb{R}^3$. In terms of $\boldsymbol{\mu}$, the mean direction is $\boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and the concentration is determined by $\|\boldsymbol{\mu}\|$. Note that (2) could equally be re-parametrised in terms of $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and $\kappa = \|\boldsymbol{\mu}\|$, analogous to the parametrisation of (1), and likewise (1) could be re-parametrised in terms of a parameter $\kappa\tilde{\boldsymbol{\mu}} \in \mathbb{R}^3$; the distinction between parametrisations is a matter of modelling convenience and in the following we shall make use of both.

Because they are isotropic, the 3-parameter Fisher and IAG distributions are too restrictive for many applications. Each, however, has a 5-parameter anisotropic generalisation: the Kent (1982) distribution, and the elliptically symmetric angular Gaussian (ESAG) distribution (Paine et al. 2017), respectively. Both the Kent and ESAG distributions have *elliptical* symmetry about the mean direction, that is, they have ellipse-like contours centred on the mean direction. The two extra parameters over their isotropic counterparts control the orientation and eccentricity of the elliptical contours. We describe the Kent and ESAG distributions in more detail in Sect. 2, but here introduce two parametrisations we shall use for each. The first parametrisation we shall consider is in terms of $(\kappa, \beta, \boldsymbol{\Gamma})$, in which $\kappa > 0$ is a concentration parameter, $\beta \geq 0$ is an eccentricity parameter, and $\boldsymbol{\Gamma} = (\tilde{\boldsymbol{\mu}}\ \boldsymbol{\xi}_1\ \boldsymbol{\xi}_2) \in O(3)$ is an orthogonal matrix (i.e. $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix), in which $\tilde{\boldsymbol{\mu}}$ is the mean direction (having 2 degrees of freedom) and $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ are the major and minor axes that identify the orientation of the elliptical contours (together having 1 remaining degree of freedom). This parametrisation generalises that of (1).

The second parametrisation we consider, generalising (2), is in terms of a pair of vectors, $\boldsymbol{\mu} \in \mathbb{R}^3$ and $\boldsymbol{\gamma} \in \mathbb{R}^2$, in which, as in (2), $\boldsymbol{\mu}$ controls the mean direction and concentration; then $\boldsymbol{\gamma} \in \mathbb{R}^2$ controls eccentricity and orientation of the elliptical contours.

These two parametrisations lend themselves to different ways of modelling how the response variable depends on covariates. We consider models with the following structures

$$\text{Structure 1:}\quad \mathbf{Q}^\top\mathbf{y}_i \sim \text{H}(\kappa, \beta, \boldsymbol{\Gamma}(\mathbf{x}_i)); \tag{3}$$

$$\text{Structure 2:}\quad \mathbf{Q}^\top\mathbf{y}_i \sim \text{H}(\boldsymbol{\mu}(\mathbf{x}_i), \boldsymbol{\gamma}(\mathbf{x}_i)); \tag{4}$$

where $\text{H}(\cdot)$ is one of $\text{Kent}(\cdot)$ or $\text{ESAG}(\cdot)$ and $\mathbf{Q}$ is an orthogonal matrix, discussed later in Sect. 3 and the "Appendix", which is needed so that inference does not depend in undesirable ways on the particular, and possibly arbitrary, coordinate system in which the $\mathbf{y}_i$ are defined. A primary difference between (3) and (4) is that in Structure 1 we allow the mean direction and orientation of the dispersion to depend on the covariate vector, but the magnitude of dispersion and anisotropy is fixed. For Structure 2, all of these depend on the covariate. We specify in Sect. 3 particular functional forms for the $\boldsymbol{\Gamma}(\cdot)$, $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\gamma}(\cdot)$, but for now note that we will consider four models that result from the different combinations of these two structures and two error distributions. We will call these Kent1, ESAG1, Kent2, and ESAG2, where, for example, ESAG1 means using $\text{H}(\cdot) = \text{ESAG}(\cdot)$ as the error distribution and Structure 1 to model dependence on covariates. This modelling approach, in which we assume that the parameters of the error distribution depend on the covariates in particular functional ways, closely parallels generalised linear modelling, although rather than having a single linear predictor, here we have several.

Before giving more details about the parametrisations and models, we briefly discuss some earlier papers on spherical regression. Rivest (1989) considered the case with covariates themselves on the sphere, $\mathbf{x}_i \in \mathbb{S}^2$, and a Fisher error distribution with the mean direction modelled as $\tilde{\boldsymbol{\mu}}(\mathbf{x}_i) = \mathbf{R}\mathbf{x}_i$, where $\mathbf{R} \in \text{SO}(3)$ is a rotation. Rosenthal et al. (2014) replaced the rotation with the "projective linear transformation" (PLT), $\tilde{\boldsymbol{\mu}}(\mathbf{x}_i) = \mathbf{A}\mathbf{x}_i/\|\mathbf{A}\mathbf{x}_i\|$, with $\mathbf{A} \in \text{SL}(3)$ where $\text{SL}(3) = \{\mathbf{A} \in \mathbb{R}^{3\times 3} : \det(\mathbf{A}) = 1\}$ is the special linear group. This is a generalisation of Rivest's model since SL(3) contains SO(3). We consider the PLT later, using it to benchmark performance of the new models we introduce.

Besides regression models on the unit sphere, $\mathbb{S}^2$, there are several models for regression on the unit circle, $\mathbb{S}^1$. Presnell et al. (1998) considers regression on $\mathbb{S}^1$ for a general covariate $\mathbf{x}_i$, assuming IAG errors. We mention this model in particular because it is a close analogue on $\mathbb{S}^1$ of our ESAG2 model on $\mathbb{S}^2$ in the isotropic case (which corresponds to $\boldsymbol{\gamma} = 0$), as discussed later. Related work includes the $\mathbb{S}^1$ regression model of Fisher and Lee (1992), but this is less relevant to the present paper because it does not generalise conveniently to $\mathbb{S}^2$ or higher dimensional spheres; see Mardia and Jupp (2000) for a discussion of this and of the wider context of regression on $\mathbb{S}^1$. We also mention a regression model for data on the simplex introduced by Scealy and Welsh (2011). Their approach is to use a "square-root transformation" to map the data from the simplex to the positive orthant of the sphere, then to develop regression models for the transformed data using the Kent distribution. On the sphere, as opposed to the simplex, however, we believe it is especially important to allow what Scealy and Welsh (2011) refer to as $\mathbf{K}^*$ to depend on regression variables, something that they do not consider due to the fact they focus on transformed compositional data; see the discussion in the concluding section of their paper.

The main goals of this paper are: to explore and compare the modelling Structures 1 and 2; to investigate in the regression context the advantages and disadvantages of the Kent and ESAG distributions as error distributions; and to develop hypothesis tests for the significance of particular covariates, and of anisotropy.

In the following section, we introduce the Kent and ESAG distributions in each of the two parametrisations, then in Sect. 3 we develop the two modelling structures and hypothesis testing procedures. In Sect. 4, we introduce some novel residuals for model fitting diagnostics; then in Sect. 5, we implement the models and methods on various examples involving both synthetic and real data. Code for fitting the models in this paper is available on the second author's web page.

2 Elliptically symmetric distributions on $\mathbb{S}^2$

Here, we give details of the $(\boldsymbol{\mu}, \boldsymbol{\gamma})$ and $(\kappa, \beta, \boldsymbol{\Gamma})$ parametrisations of the Kent and ESAG distributions.

2.1 Kent distribution

Kent (1982) introduced this distribution using a $(\kappa, \beta, \boldsymbol{\Gamma})$ parametrisation, in terms of which the density is

$$f_{\text{Kent}}(\boldsymbol{y}|\kappa, \beta, \boldsymbol{\Gamma}) = C(\kappa, \beta)^{-1} \times \exp\left(\kappa \boldsymbol{y}^\top \tilde{\boldsymbol{\mu}} + \beta\left(\left(\boldsymbol{y}^\top \boldsymbol{\xi}_1\right)^2 - \left(\boldsymbol{y}^\top \boldsymbol{\xi}_2\right)^2\right)\right), \tag{5}$$

where $C(\kappa, \beta)$ is the normalising constant.

Lemma 1 *The Kent density in a $(\boldsymbol{\mu}, \boldsymbol{\gamma})$ parametrisation is*

$$f_{Kent}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\gamma}) = C(\kappa, \beta)^{-1} \times \exp\left(\boldsymbol{\mu}^\top \mathbf{y} + \mathbf{y}^\top\left(\gamma_1\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_1^\top - \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_2^\top\right) + \gamma_2\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_2^\top + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_1^\top\right)\right)\mathbf{y}\right), \tag{6}$$

where $\kappa = \|\boldsymbol{\mu}\|$, $\beta = \sqrt{\gamma_1^2 + \gamma_2^2}$ and $(\tilde{\boldsymbol{\xi}}_1\ \tilde{\boldsymbol{\xi}}_2) = (\boldsymbol{\xi}_1\ \boldsymbol{\xi}_2)\mathbf{R}(\psi)^\top$, with $\mathbf{R}(\psi)$ defined as in (14), and where $\psi \in (0, \pi]$ is the solution of $\gamma_1 = \beta\cos 2\psi$ and $\gamma_2 = \beta \sin 2\psi$.

The proof of Lemma 1 is in the "Appendix".

2.2 Elliptically symmetric angular Gaussian (ESAG) distribution

The general angular Gaussian distribution is the marginal distribution of the directional component of the multivariate normal distribution; that is, for a general mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\mathbf{V}$, let $\mathbf{z} \sim N(\boldsymbol{\mu}, \mathbf{V}) \in \mathbb{R}^p$, then $\mathbf{z}/\|\mathbf{z}\|$ has a general angular Gaussian distribution. The elliptically symmetric angular Gaussian (ESAG) distribution, developed in Paine et al. (2017), is a subfamily of the general angular Gaussian distribution. It is defined by the two conditions

$$\mathbf{V}\boldsymbol{\mu} = \boldsymbol{\mu}, \qquad \det(\mathbf{V}) = 1, \tag{7}$$

and on $\mathbb{S}^2$ has density

$$f_{\text{ESAG}}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V}) = \frac{1}{2\pi(\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y})^{3/2}} \times \exp\left[\frac{1}{2}\left\{\frac{(\mathbf{y}^\top\boldsymbol{\mu})^2}{\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y}} - \boldsymbol{\mu}^\top\boldsymbol{\mu}\right\}\right] \times M\left\{\frac{\mathbf{y}^\top\boldsymbol{\mu}}{(\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y})^{1/2}}\right\}, \tag{8}$$

where $M(\cdot)$ is defined as in (2). Distribution (8) has 5 free parameters, which can be seen by first fixing the 3 free parameters of $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^\top$ then observing that conditions (7) leave 2 degrees of freedom in $\mathbf{V}$. Let ρ_1, ρ_2, ρ_3 be the eigenvalues of $\mathbf{V}$, with corresponding orthonormal eigenvectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3$, respectively. Then, by the spectral decomposition theorem,

$$\begin{aligned}\mathbf{V}^{-1} &= \rho_1^{-1}\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top + \rho_2^{-1}\boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top + \rho_3^{-1}\boldsymbol{\xi}_3\boldsymbol{\xi}_3^\top \\ &= \rho_1^{-1}\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top + \rho_1\boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top + \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top,\end{aligned} \tag{9}$$

where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. The final term in (9) is a consequence of the constraint $\mathbf{V}\boldsymbol{\mu} = \boldsymbol{\mu}$, and $\rho_2^{-1} = \rho_1$ then follows from

$\det(\mathbf{V}) = 1$. Once $\boldsymbol{\mu}$ is fixed, then in $\mathbf{V}^{-1}$ there is one degree of freedom from ρ_1, and one degree of freedom from fixing the orientation of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$.

Lemma 2 *The ESAG density in a* $(\kappa, \beta, \boldsymbol{\Gamma})$ *parametrisation, where* $\boldsymbol{\Gamma} = (\tilde{\boldsymbol{\mu}}\ \boldsymbol{\xi}_1\ \boldsymbol{\xi}_2)$*, is given by* (8) *with* $\mathbf{V} = \mathbf{V}(\beta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ *defined by*

$$\mathbf{V}^{-1} = \mathbf{I} + \beta\left(\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top - \boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top\right) + \left(\sqrt{\beta^2+1} - 1\right)\left(\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top + \boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top\right), \tag{10}$$

with $\beta = 2^{-1}\left(\rho_1^{-1} - \rho_1\right)$

Lemma 2 follows directly from substituting (10) and $\boldsymbol{\mu} = \kappa\tilde{\boldsymbol{\mu}}$, with $\kappa \geq 0$ and $\tilde{\boldsymbol{\mu}} \in \mathbb{S}^2$, into (8). Note that $\beta = 0$ in (10) implies isotropy.

Paine et al. (2017) chose to parametrise $\mathbf{V}^{-1}$ via

$$\tilde{\boldsymbol{\xi}}_1 \equiv \tilde{\boldsymbol{\xi}}_1(\boldsymbol{\mu}) = \left(-\mu_0^2, \mu_1\mu_2, \mu_1\mu_3\right)^\top / (\mu_0\|\boldsymbol{\mu}\|) \tag{11}$$

and

$$\tilde{\boldsymbol{\xi}}_2 \equiv \tilde{\boldsymbol{\xi}}_2(\boldsymbol{\mu}) = (0, -\mu_3, \mu_2)^\top / \mu_0, \tag{12}$$

where $\mu_0 = (\mu_2^2 + \mu_3^2)^{1/2} > 0$. Hence, $\tilde{\boldsymbol{\xi}}_1$ and $\tilde{\boldsymbol{\xi}}_2$ are unit vectors which are orthogonal to each other and to the mean direction $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. Each is a function of $\boldsymbol{\mu}$ and related to $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ via a rotation

$$(\boldsymbol{\xi}_1\ \boldsymbol{\xi}_2) = (\tilde{\boldsymbol{\xi}}_1\ \tilde{\boldsymbol{\xi}}_2)\,\mathbf{R}(\psi), \tag{13}$$

where

$$\mathbf{R}(\psi) = \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix}. \tag{14}$$

Substituting (13) into (9), and using the fact that $\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top + \boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top + \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, leads to

$$\mathbf{V}^{-1} = \mathbf{I} + \gamma_1\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_1^\top - \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_2^\top\right) + \gamma_2\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_2^\top + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_1^\top\right) + \left\{(\gamma_1^2 + \gamma_2^2 + 1)^{1/2} - 1\right\}\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_1^\top + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_2^\top\right), \tag{15}$$

where

$$\begin{pmatrix}\gamma_1 \\ \gamma_2\end{pmatrix} = 2^{-1}\left(\rho_1^{-1} - \rho_1\right)\begin{pmatrix}\cos 2\psi \\ \sin 2\psi\end{pmatrix};$$

See Lemma 1 in Paine et al. (2017). The $(\boldsymbol{\mu}, \boldsymbol{\gamma})$ parametrisation of the density, $f_{\mathrm{ESAG}}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\gamma})$, is hence given by (8), with $\mathbf{V} = \mathbf{V}(\boldsymbol{\mu}, \boldsymbol{\gamma})$ defined by (15). An advantage of this parametrisation is that γ_1 and γ_2 are unconstrained, which is helpful for regression modelling. The isotropic subfamily, IAG, is the special case with $\boldsymbol{\gamma} = (0, 0)^\top$.

2.3 Practical differences between Kent and ESAG distributions

Both the Kent and ESAG distributions have similar characteristics from a modelling perspective: each typically has ellipse-like contours of constant probability density centred on the mean direction in the unimodal case and for different parameter values each has unimodal and bimodal cases. On practical grounds, the two distributions have different advantages and disadvantages. The Kent distribution belongs to the exponential family, and hence, its density, (5), has a simple mathematical form. In comparison, the ESAG density, (8), is rather cumbersome. On the other hand, the ESAG density and likelihood can be computed exactly, whereas the Kent density and likelihood involves a normalising constant, $C(\kappa, \beta)$ in (5), which is not known in closed form and hence needs to be approximated, by truncating an infinite series (Kent 1982), or else by saddlepoint or holonomic gradient methods (Kume and Sei 2017; Kume et al. 2013). In the present context, we maximise the likelihood for the regression models numerically, so the ESAG likelihood having a cumbersome form is no drawback, and the fact that it can be computed exactly is an advantage. For simulation, the Kent distribution requires a rejection algorithm (Kent et al. 2018), whereas ESAG can be simulated quickly and easily. Fast simulation is especially helpful in simulation-heavy inference procedures, e.g. the parametric bootstrap.

3 Regression model structures

In this section, we specify the two model structures in (3) and (4) and then discuss the advantages and disadvantages of each. It is assumed throughout the paper that the first element of $\mathbf{x}_i$ is 1, which is analogous to the inclusion in linear modelling of an "intercept term". For Structure 2 models, see (4), this means that the simpler model of $\{\mathbf{y}_i\}$ being IID, i.e. not depending on the covariates, is nested in the general regression model and this is helpful for testing the significance of regression. The motivation for including the intercept term is less clear-cut a priori for Structure 1 models, see (3), though empirical results, for example in Table 2 later, suggest there is sometimes a benefit from doing so.

Each model structure is defined in terms of a preliminary orthogonal transformation, $\mathbf{Q}$. For Structure 1 models, $\mathbf{Q}$ is assumed to be a population quantity, defined explicitly in the "Appendix", and estimated by a sample version $\hat{\mathbf{Q}}$. For Structure 2 models, $\mathbf{Q}$ is treated as a tuning parameter and optimised with respect to. These preliminary transformations are needed so that desirable invariance and equivariance

properties, discussed in the "Appendix", hold when an arbitrary orthogonal transformation is applied to the $\mathbf{y}_i$.

3.1 Structure 1: $\mathbf{Q}^\top \mathbf{y}_i \sim H(\kappa, \boldsymbol{\beta}, \boldsymbol{\Gamma}(\mathbf{x}_i))$

In this structure, $\mathbf{Q}$ is a population quantity, as defined in the "Appendix", and in all calculations involving data it is replaced by a sample version $\hat{\mathbf{Q}} = [\hat{\boldsymbol{\xi}}\ \hat{\boldsymbol{\xi}}_1\ \hat{\boldsymbol{\xi}}_2]$ with $\hat{\boldsymbol{\xi}} = \sum_{i=1}^n \mathbf{y}_i / \|\sum_{i=1}^n \mathbf{y}_i\|$ and $\hat{\boldsymbol{\xi}}_1$ and $\hat{\boldsymbol{\xi}}_2$ unit eigenvectors corresponding to the larger and smaller positive eigenvalues of

$$\left(\mathbf{I}_3 - \hat{\boldsymbol{\xi}}\hat{\boldsymbol{\xi}}^\top\right) \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top \left(\mathbf{I}_3 - \hat{\boldsymbol{\xi}}\hat{\boldsymbol{\xi}}^\top\right).$$

Here, $\hat{\mathbf{Q}}$ is the moment estimator defined in Kent (1982, p. 74) of $\boldsymbol{\Gamma}$ in defined in Kent (1982, p. 74) (5) under the assumption of IID $\mathbf{y}_i$.

We consider for $\boldsymbol{\Gamma}(\mathbf{x}_i)$ viewed as a function of $\mathbf{x}_i$:

$$\boldsymbol{\Gamma}(\mathbf{x}_i) = \mathbf{R}(\mathbf{x}_i)\text{diag}[1, \mathbf{S}(\mathbf{x}_i)], \qquad i = 1, \ldots, n, \tag{16}$$

where the $\mathbf{R}(\mathbf{x}_i)$ are orthogonal 3-by-3 matrices, the $\mathbf{S}(\mathbf{x}_i)$ are orthogonal 2-by-2 matrices, diag[., .] is a 3-by-3 block diagonal matrix with 1×1 and 2×2 blocks, and $\mathbf{x}_i$ is $q \times 1$.

The dependence of $\mathbf{R}(\mathbf{x}_i)$ and $\mathbf{S}(\mathbf{x}_i)$ on the covariate vector $\mathbf{x}_i$ needs to be prescribed. We choose to do so using the Cayley transform: for any skew-symmetric matrix $\mathbf{A}$, i.e. $\mathbf{A} = -\mathbf{A}^\top$, the matrix $(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A})^{-1}$ is a 3-by-3 rotation matrix (i.e. an orthogonal matrix with determinant $+1$). The Cayley transform maps the skew-symmetric matrices onto the set of rotations minus a set of lower dimension (see the "Appendix"). This is an injective mapping, which is the reason we favour it over, e.g., the exponential of $\mathbf{A}$. Define

$$\begin{aligned} \mathbf{R}(\mathbf{x}_i) &= (\mathbf{I} - \mathbf{A}_{\mathrm{R},i})(\mathbf{I} + \mathbf{A}_{\mathrm{R},i})^{-1}, \quad \text{and} \\ \mathbf{S}(\mathbf{x}_i) &= (\mathbf{I} - \mathbf{A}_{\mathrm{S},i})(\mathbf{I} + \mathbf{A}_{\mathrm{S},i})^{-1} \end{aligned} \tag{17}$$

where

$$\begin{aligned} \mathbf{A}_{\mathrm{R},i} &= \begin{pmatrix} 0 & \boldsymbol{\beta}_1^\top \mathbf{x}_i & \boldsymbol{\beta}_2^\top \mathbf{x}_i \\ -\boldsymbol{\beta}_1^\top \mathbf{x}_i & 0 & \boldsymbol{\beta}_3^\top \mathbf{x}_i \\ -\boldsymbol{\beta}_2^\top \mathbf{x}_i & -\boldsymbol{\beta}_3^\top \mathbf{x}_i & 0 \end{pmatrix}, \quad \text{and} \\ \mathbf{A}_{\mathrm{S},i} &= \begin{pmatrix} 0 & \boldsymbol{\beta}_4^\top \mathbf{x}_i \\ -\boldsymbol{\beta}_4^\top \mathbf{x}_i & 0 \end{pmatrix}, \end{aligned} \tag{18}$$

are skew-symmetric. Here, $\mathbf{R}(\cdot)$ and $\mathbf{S}(\cdot)$, and hence $\boldsymbol{\Gamma}(\cdot)$, are playing a role analogous to link functions in generalised linear models, linking linear predictors to the parameters of the distribution of the response variable. The nature of the link functions means that interpreting the influence of individual β_js is somewhat harder for this model than for Structure 2 models described below. This model is fitted by maximising the likelihood function of observed data with respect to the 4-by-q parameter matrix $\mathbf{B} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4\right)^\top$.

3.2 Structure 2: $\mathbf{Q}^\top \mathbf{y}_i \sim H(\boldsymbol{\mu}(\mathbf{x}_i), \boldsymbol{\gamma}(\mathbf{x}_i))$

In this parametrisation, $\boldsymbol{\mu} \in \mathbb{R}^3$ and $\boldsymbol{\gamma} \in \mathbb{R}^2$ are unrestricted, and $\boldsymbol{\mu}(\mathbf{x}_i)$ and $\boldsymbol{\gamma}(\mathbf{x}_i)$ are easy to specify as functions mapping from the q-dimensional domain of the $\{\mathbf{x}_i\}$ to $\mathbb{R}^3$ and $\mathbb{R}^2$, respectively. Here, we limit attention to linear functions,

$$\begin{aligned} \boldsymbol{\mu}(\mathbf{x}_i) &= \begin{pmatrix} \boldsymbol{\beta}_1^\top \mathbf{x}_i \\ \boldsymbol{\beta}_2^\top \mathbf{x}_i \\ \boldsymbol{\beta}_3^\top \mathbf{x}_i \end{pmatrix} = \mathbf{B}_1 \mathbf{x}_i, \quad \text{and} \\ \boldsymbol{\gamma}(\mathbf{x}_i) &= \begin{pmatrix} \boldsymbol{\beta}_4^\top \mathbf{x}_i \\ \boldsymbol{\beta}_5^\top \mathbf{x}_i \end{pmatrix} = \mathbf{B}_2 \mathbf{x}_i, \end{aligned} \tag{19}$$

where $\mathbf{B}_1 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)^\top$ and $\mathbf{B}_2 = (\boldsymbol{\beta}_4, \boldsymbol{\beta}_5)^\top$. In keeping with the notation of the preceding model, we collect these parameters together into a 5-by-q matrix $\mathbf{B} = \left(\mathbf{B}_1^\top, \mathbf{B}_2^\top\right)^\top$, where the influence of the subsets of parameters can be clearly distinguished: $\mathbf{B}_1$ controls the influence of the covariates, via $\boldsymbol{\mu}$, on the concentration and mean direction; and $\mathbf{B}_2$ controls influence, via $\boldsymbol{\gamma}$, on the degree and orientation of anisotropy. This leads to natural tests, e.g. for anisotropy, discussed below.

Unlike in Structure 1, in which model (16) is naturally tied to the particularly defined $\mathbf{Q}$, for Structure 2 and model (19) there is no a priori reason to select a particular $\mathbf{Q} \in O(3)$; hence, we treat $\mathbf{Q}$ as a tuning parameter, seeking to maximise the likelihood of the data $\left\{\mathbf{Q}^\top \mathbf{y}_i\right\}$ over $\{\mathbf{Q}, \mathbf{B}\}$. A practical way to do so at least approximately is via a brute-force search for $\mathbf{Q}$ over $O(3)$, for each value of $\mathbf{Q}$ on a grid over $O(3)$ computing the maximum likelihood estimator $\hat{\mathbf{B}}$ of $\mathbf{B}$, then selecting the pair $\left\{\mathbf{Q}, \hat{\mathbf{B}}\right\}$ corresponding to the largest maximised likelihood. In this paper, when comparing models for a particular data set, we compute $\mathbf{Q}$ for the most general ESAG2 model and keep this $\mathbf{Q}$ fixed for submodels and Kent2 models.

Model (19) with ESAG errors,

$$\mathbf{y}_i \sim \text{ESAG}(\mathbf{B}_1 \mathbf{x}_i, \mathbf{B}_2 \mathbf{x}_i), \tag{20}$$

is close in spirit to the circular $p = 2$ regression models of Presnell et al. (1998) and Wang and Gelfand (2013), particularly in the isotropic special case, with $\mathbf{B}_2 = \mathbf{0}$, in which this model is a direct analogue for $p = 3$ of Presnell et al.'s regression model on the circle.

A helpful property proved by Presnell et al. in the circular case is that the log-likelihood function is a concave function of the regression parameters—in our notation $\mathbf{B}_1$—that determine $\boldsymbol{\mu}$; this guarantees that the MLE of $\mathbf{B}_1$ is unique and easily determined by numerical optimisation. The corresponding result holds for ESAG2 (20) in the $p = 3$ case with isotropic errors, i.e. $\mathbf{B}_2 = \mathbf{0}$, as follows [in which vec

is the standard vectorisation operator; see e.g. Mardia et al. (1979)].

Proposition 1 *Consider model* (20)*, let* $\mathbf{B}_2 = \mathbf{0}$*, and let* $l(\mathbf{B}_1)$ *denote the log of the likelihood function for parameter* $\mathbf{B}_1$ *given observations* $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$*. Provided* $(\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ *has full rank, the negative Hessian*

$$-\frac{\partial^2 l(\mathbf{B}_1)}{\partial \operatorname{vec} \mathbf{B}_1 \partial \operatorname{vec} \mathbf{B}_1^\top}$$

is positive definite, hence $l(\mathbf{B}_1)$ *is a concave function of* $\mathbf{B}_1$*, and the MLE of* $\mathbf{B}_1$ *is unique.*

The proof of this Proposition is given in the "Appendix".

3.3 Tests for the significance of anisotropy and regression

In this section, we discuss procedures for performing hypothesis tests required for model selection and inference. To do so, we introduce the notation for the parameters $\mathbf{B} = \left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(q)}\right)$, i.e. such that $\boldsymbol{\beta}^{(j)}$ is the jth column of $\mathbf{B}$ and corresponds to the covariate appearing as the jth element of $\mathbf{x}_i$. A test of the significance of this particular covariate corresponds to a test with null and alternative hypotheses

$$\mathrm{H}_0{:}\boldsymbol{\beta}^{(j)} = \mathbf{0} \quad \text{versus} \quad \mathrm{H}_1{:}\boldsymbol{\beta}^{(j)} \text{ free.}$$

Since the null hypothesis is nested in the alternative then by Wilks' theorem, subject to the usual regularity conditions, under H_0,

$$T = -2 \log (L_0/L_1) \sim \chi^2_\nu, \tag{21}$$

asymptotically as $n \to \infty$, where L_0 and L_1 are the maximised likelihood functions under H_0 and H_1, respectively; and ν is the difference in the number of free parameters between H_0 and H_1, here equal to 4 or 5 for the Structure 1 and 2 models, respectively. The significance of the parameter can be assessed by referring the observed test statistic, T, to the χ^2_ν distribution. An alternative possibility, preferable when n is insufficiently large for the null asymptotic distribution (21) to be reasonable, is to approximate the null distribution using a bootstrap procedure.

Within Structure 2 models, it may be relevant to consider whether particular covariates are significant in $\boldsymbol{\mu}$ or $\boldsymbol{\gamma}$ distinctly. For example, for the covariate corresponding to the jth element of $\mathbf{x}_i$, a test that the covariate is significant in $\boldsymbol{\gamma}$ corresponds to the hypotheses

$$\begin{aligned} &\mathrm{H}_0{:}(\boldsymbol{\beta}^{(j)})_4 = (\boldsymbol{\beta}^{(j)})_5 = 0 \quad \text{versus} \\ &\mathrm{H}_1{:}(\boldsymbol{\beta}^{(j)})_4,\ (\boldsymbol{\beta}^{(j)})_5 \text{ free,} \end{aligned} \tag{22}$$

for which the degrees of freedom in (21) is $\nu = 2$. Having isotropic errors corresponds to $\boldsymbol{\gamma} = \mathbf{0}$, so for a test of the significance of anisotropy the hypotheses are $\mathrm{H}_0{:}\mathbf{B}_2 = \mathbf{0}$ versus $\mathrm{H}_1{:}\mathbf{B}_2$ free, where $\mathbf{B}_2$ is as defined in (19), and $\nu = 2q$.

4 Residuals for model diagnostics

For spherical regression models, there are many possible ways to define a residual. Here, we describe some general spherical residuals defined by Jupp (1988) before defining some particular model-based residuals for regression models with ESAG and Kent errors.

For observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ denote the fitted values by $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n$. Jupp defined "crude residuals" as

$$\mathbf{r}_i = \left(\mathbf{I} - \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i^\top\right) \mathbf{y}_i,$$

i.e. as projections of each observation, $\mathbf{y}_i$, into the tangent plane at its fitted value, $\hat{\mathbf{y}}_i$. Since the $\mathbf{r}_1, \ldots, \mathbf{r}_n$ lie in different tangent planes, Jupp defined the "rotated residuals"

$$\mathbf{s}_i = \mathbf{R}(\hat{\mathbf{y}}_i, \mathbf{y}_0)\mathbf{r}_i, \tag{23}$$

where $\mathbf{y}_0$ is an arbitrary point on the sphere which is not dependent on i, and $\mathbf{R}(\hat{\mathbf{y}}_i, \mathbf{y}_0)$ is a rotation from $\hat{\mathbf{y}}_i$ to $\mathbf{y}_0$, where $\mathbf{R}(\cdot, \cdot)$ does not depend on i. Then, the $\mathbf{s}_1, \ldots, \mathbf{s}_n$ lie in the plane tangent to the sphere at $\mathbf{y}_0$. Let $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2$ be an arbitrary pair of unit vectors orthogonal to each other and to $\mathbf{y}_0$, then a plot of the projected residuals

$$\mathbf{t}_i = \begin{pmatrix} \boldsymbol{\zeta}_1^\top \\ \boldsymbol{\zeta}_2^\top \end{pmatrix} \mathbf{s}_i, \tag{24}$$

can be inspected to identify structure amongst residuals that could indicate a shortcoming of the model.

For parametric regression models with ESAG or Kent errors, Jupp's residuals are potentially limited in that they are not model-based and hence do not take into account the dispersion of errors in the fitted model, i.e. (23) is a function of the fitted value $\hat{\mathbf{y}}_i$ but not of the parameters that determine dispersion.

We define model-based residuals for ESAG and Kent error models as follows, in each case motivated by high-concentration Gaussian limits of each distribution, although we expect the residuals to be useful for detecting model inadequacy even in non high-concentration settings. For a random variable $\mathbf{y} \sim \mathrm{ESAG}(\boldsymbol{\mu}, \boldsymbol{\gamma})$ consider the corresponding random variable

$$\boldsymbol{\eta}_{\mathrm{ESAG}}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\gamma}) = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \|\boldsymbol{\mu}\| \begin{pmatrix} \rho_1^{-1/2} \boldsymbol{\xi}_1^\top \\ \rho_1^{1/2} \boldsymbol{\xi}_2^\top \end{pmatrix} \mathbf{y},$$

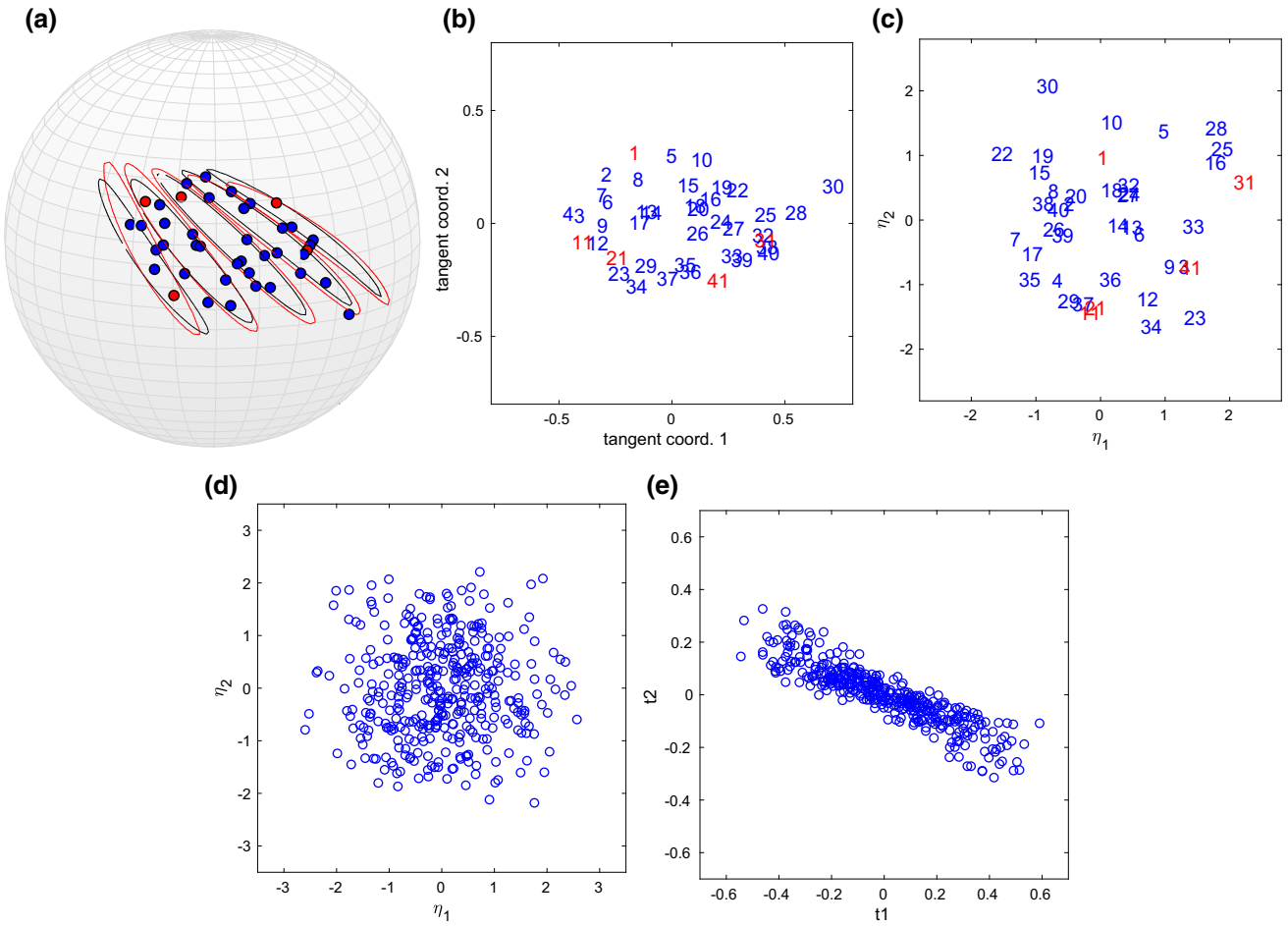


Fig. 1 Data and residuals for the model described in Sect. 5.1: **a** A data set with $n = 41$ plotted on the sphere; points with indices $i = 1, 11, 21, 31, 41$ are marked in red with corresponding 90%-coverage contours of ESAG($\boldsymbol{\mu}_i, \boldsymbol{\gamma}_i$), shown in red for the true parameters and black for the fitted. **b** the same data projected into the tangent plane at the sample mean, and **c** $\boldsymbol{\eta}$-residuals (25) for the fitted model. **d**, **e**, respectively, show $\boldsymbol{\eta}$-residuals and Jupp residuals (24) for a larger data set of $n = 401$ data points from the same model

where $\rho_1, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ are as defined in (9). From Proposition 2 in Paine et al. (2017), provided $\|\boldsymbol{\mu}\|$ is large then, approximately, $\boldsymbol{\eta}_{\text{ESAG}} \sim N_2(\mathbf{0}, \mathbf{I})$. Hence for regression models with ESAG errors, we define residuals

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_{\text{ESAG}}(\mathbf{y}_i; \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\gamma}}_i) \text{ for } i = 1, \ldots, n, \tag{25}$$

where $\hat{\boldsymbol{\mu}}_i = \hat{\mathbf{B}}_1 \mathbf{x}_i$ and $\hat{\boldsymbol{\gamma}}_i = \hat{\mathbf{B}}_2 \mathbf{x}_i$. Then, a scatterplot of $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_n$ can be compared with random $N_2(\mathbf{0}, \mathbf{I})$ scatter; see Fig. 1 for examples.

Similarly, for a random variable $\mathbf{y} \sim \text{Kent}(\kappa, \beta, \boldsymbol{\Gamma})$ then

$$\boldsymbol{\eta}_{\text{Kent}}(\mathbf{y}; \kappa, \beta, \boldsymbol{\Gamma}) = \sqrt{\kappa} \begin{pmatrix} \beta^{-1/2} \boldsymbol{\xi}_1^\top \\ \beta^{1/2} \boldsymbol{\xi}_2^\top \end{pmatrix} \mathbf{y} \sim N_2(\mathbf{0}, \mathbf{I}),$$

approximately, for large κ; see property (e) in Kent (1982). For models with Kent errors, writing $\hat{\boldsymbol{\Gamma}}_i = \hat{\boldsymbol{\Gamma}}(\mathbf{x}_i)$, we hence define the residuals

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_{\text{Kent}}(\mathbf{y}; \hat{\kappa}, \hat{\beta}, \hat{\boldsymbol{\Gamma}}_i).$$

5 Applications

Here, we consider three applications, in each investigating the spherical regression models towards different statistical goals.

The first involves a simulated data set with a scalar covariate, $t \in \mathbb{R}$. We exploit having a simple data-generating model to illustrate the flexibility within this regression framework for the mean direction and dispersion to depend on the covariate; to investigate the performance of hypothesis tests in detecting anisotropy and regression; and to compare Jupp and $\boldsymbol{\eta}$-residuals in the special setting where the model being fitted is the true one.

The second data set concerns the movement of clouds between two consecutive days. The cloud shapes are repre-

Table 1 Results from fitting various models to the synthetic data, which were generated from model M_1 with H taken to be ESAG, $n = 41$, and using parameters described in Sect. 5.1

Model	Model for $\mathbf{y}_i$		Log-lik. ESAG	Log-lik. Kent	(*d.f.*)
Structure 2					
M_1	$\mathrm{H}(\mathbf{B}_1\mathbf{x}_i, \mathbf{B}_2\mathbf{x}_i)$		91.7	81.5	(10)
M_2	$\mathrm{H}(\mathbf{B}_1\mathbf{x}_i, \boldsymbol{\gamma})$		63.5	62.3	(8)
M_3	$\mathrm{H}(\mathbf{B}_1\mathbf{x}_i, \mathbf{0})$	Isotropic errors	28.5	28.0	(6)
M_4	$\mathrm{H}(\boldsymbol{\mu}, \boldsymbol{\gamma})$	IID observations	9.6	10.2	(5)
Structure 1					
M_5	$\mathrm{H}(\kappa, \beta, \boldsymbol{\Gamma}(\mathbf{x}_i))$		89.7	88.9	(10)
M_6	$\mathrm{H}(\kappa, \beta, \boldsymbol{\Gamma}(\mathbf{x}_i)), \boldsymbol{\beta}_4 = \mathbf{0}$		88.6	87.6	(8)
M_7	$\mathrm{H}(\kappa, 0, \boldsymbol{\Gamma}(\mathbf{x}_i))$	Isotropic errors	29.2	28.9	(7)
M_8	$\mathrm{H}(\kappa, \beta, \boldsymbol{\Gamma})$	IID observations	2.5	2.8	(2)

sented by landmarks spaced around the cloud outline, and the position of these landmarks is regressed on their positions the previous day. This data set has been considered previously in the context of spherical–spherical regression models with isotropic errors (Rosenthal et al. 2014); hence, it makes for an interesting comparison with the more general framework developed in this paper.

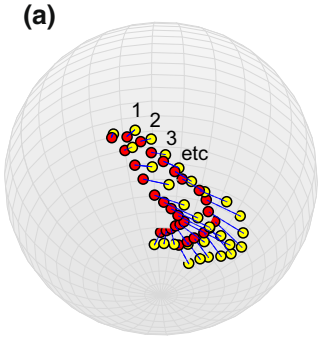
The third data set is derived from vectorcardiogram measurement of heart activity in children. These data too have been studied in the context of spherical–spherical isotropic regression (Chang 1986), but with the non-spherical covariates disregarded. The primary goal is inference, to understand which covariates are significantly related to the response. The framework of the present paper enables us to incorporate easily the additional non-spherical covariates, as well as anisotropic errors, and furthermore then to test formally whether such generalisations are warranted by the data.

5.1 Simulated data set (involving a scalar covariate)

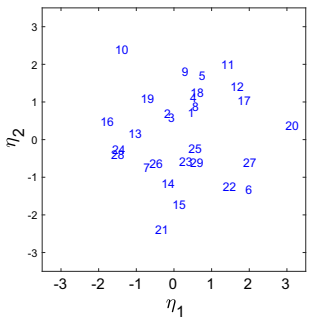
Denote by M_1 the ESAG2 regression model with $\boldsymbol{\mu}_i = (1 - t_i)\boldsymbol{\mu}^{(0)} + t_i\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\gamma}_i = (1 - t_i)\boldsymbol{\gamma}^{(0)} + t_i\boldsymbol{\gamma}^{(1)}$, and $t_i = (i - 1)/(n - 1)$ for $i = 1, \ldots, n$. In the notation of (20), $\mathbf{B}_1 = \left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(0)}\right)$, $\mathbf{B}_2 = \left(\boldsymbol{\gamma}^{(0)}, \boldsymbol{\gamma}^{(1)} - \boldsymbol{\gamma}^{(0)}\right)$, and $\mathbf{x}_i = (1, t_i)^\top$. Figure 1a–c is plot for a synthetic data set generated from M_1 using $\boldsymbol{\mu}^{(0)} = (5, 10, 2)^\top$, $\boldsymbol{\mu}^{(1)} = (-5, 10, 2)^\top$, $\boldsymbol{\gamma}^{(0)} = (2, 3)^\top$, $\boldsymbol{\gamma}^{(1)} = (-2, 5)^\top$, and $n = 41$. As a visual aid, plot markers corresponding to the subset of points with indices $i = 1, 11, 21, 31, 41$ are coloured red. Figure 1a shows the data, together with contours of constant probability density with 90% coverage for the true and fitted parameters, for the covariates corresponding to the red-marked points. The data-generating parameters are deliberately chosen here to produce highly anisotropic dispersion, as can be seen from the highly eccentric contours. These contours are well matched by corresponding contours of the fitted model, indicating that the parameters have been estimated well. Figure 1b shows the same data projected onto the tangent plane at the sample mean, with the index used as the point marker so that the ordering of the points can be seen. Figure 1c is a plot of $\boldsymbol{\eta}$-residuals, which seem consistent with IID bivariate normal scatter, indicating that the model is reasonable. This is expected since the data-generating model is a special case of the model being fitted. Exploring residuals further, Fig. 1d shows residuals analogous to those in c but this time for a larger sample size of $n = 401$, with the corresponding Jupp projecting residuals (24) shown in e. The Jupp residuals appear non-Gaussian and anisotropic, even though the fitted model is appropriate to the data, making these residuals harder in general to interpret for model diagnostics.

We can use the inference procedures described in Sect. 3.3 to test for significance of anisotropy and regression. Table 1 shows the maximised log-likelihood for the true model, M_1, and some different models involving various combinations of the two model structures and two error distributions. Using Wilks' statistic and the null asymptotic χ^2 approximation (21) to compare M_1 with each of models M_2, M_3, M_4 with errors assumed to be ESAG results in p values $< 10^{-5}$ in each case, indicating very strong evidence to favour the data-generating model over the simpler alternatives, which include the isotropic (M_3) and IID (M_4) models. When Kent errors are assumed for the fitted model, i.e. in contrast to the ESAG errors used in generating the data, the statistical conclusions (and even to some extent the numerical values of the maximised log-likelihoods) are very robust to this misspecification. This is probably a consequence of how similar the ESAG and Kent densities are in practise, especially if the concentration is reasonably high. The table also shows the results of fitting Structure 2 models M_5–M_8 to the Structure 1-generated data. Here, model M_5 is not favoured strongly over M_6, in contrast to how M_1 is strongly favoured over M_2. The explanation is that models M_2 and M_6 are only loosely analogous as submodels of M_1 and M_5, respectively. A major difference is that M_2 cannot capture the way the orientation of the anisotropy substantially

(a)



(b)



(c)

Model	Model for $\mathbf{y}_i$	log lik.	(d.f.)
ESAG1	ESAG$(\alpha, \beta, \boldsymbol{\Gamma}(\mathbf{x}_i))$	139.6	(18)
IAG1	ESAG$(\alpha, 0, \boldsymbol{\Gamma}(\mathbf{x}_i))$	105.2	(17)
Kent1	Kent$(\alpha, \beta, \boldsymbol{\Gamma}(\mathbf{x}_i))$	140.3	(18)
Fisher1	Kent$(\alpha, 0, \boldsymbol{\Gamma}(\mathbf{x}_i))$	105.2	(17)
ESAG2	ESAG$(\mathbf{B}_1\mathbf{x}_i, \mathbf{B}_2\mathbf{x}_i)$	161.4	(20)
IAG2	ESAG$(\mathbf{B}_1\mathbf{x}_i, \mathbf{0})$	108.4	(12)
Kent2	Kent$(\mathbf{B}_1\mathbf{x}_i, \mathbf{B}_2\mathbf{x}_i)$	143.5	(20)
Fisher2	Kent$(\mathbf{B}_1\mathbf{x}_i, \mathbf{0})$	107.4	(12)
PLT	Fisher$(\kappa, \tilde{\boldsymbol{\mu}} = \mathbf{A}\mathbf{u}_i / \lVert\mathbf{A}\mathbf{u}_i\rVert)$	84.4	(9)

Fig. 2 **a** The cloud formation data described in Sect. 5.2. The red points are landmarks on the outline of the cloud on a particular day, and the yellow points connected by blue lines are the corresponding landmarks on the following day. In the regression, we treat the red points as covariates and the yellow points as the response. **b** $\boldsymbol{\eta}$-residuals (25) for the fitted ESAG2 model; the residuals are numbered clockwise starting from the point indicated in **a**. The table in **c** results from fitting various models to the cloud data. For the PLT model, $\mathbf{A} \in \mathrm{SL}(3)$, and the covariate vector is $\mathbf{u}_i \in \mathbf{S}^2$, without an "intercept" element included. (Color figure online)

depends on the covariate, because $\boldsymbol{\gamma}$ does not depend on the covariate, whereas M_6 can still do so via $\mathbf{R}(\mathbf{x}_i)$ even when $\mathbf{S}(\mathbf{x}_i)$ is fixed to be the identity matrix. The conclusion to reject isotropy (M_7) and the assumption of IID data (M_8) in favour of M_5 are both robust to the model misspecification.

5.2 Cloud formation data (involving a spherical covariate)

These data involve 29 landmarks spaced around the outline of a cloud to represent its shape on each of two consecutive days, 4th and 5th Sept 2012. The data, see Fig. 2, are from NASA's Visible Earth project [with original cloud images from XPlanet (2018)] and were used as an application by Rosenthal et al. (2014) in assessing accuracy of their PLT model, albeit with a focus on prediction rather than inference. The goal is to regress the landmarks $\{\mathbf{y}_i\}_{i=1}^{29}$ for the second day on those $\{\mathbf{u}_i\}_{i=1}^{29}$ of the first. We hence define a covariate vector, including "intercept", as $\mathbf{x}_i = \left(1, \mathbf{u}_i^\top\right)^\top$.

The models we fitted to these data and the corresponding values of the maximised log-likelihood, are shown in Fig. 2c. The maximised log-likelihood values show that each of the models with anisotropic errors is very strongly favoured over its isotropic counterpart. The non-nestedness of the models otherwise makes them hard to select between formally. Model ESAG2 has substantially the largest log-likelihood value, although recall that the transformation $\boldsymbol{Q}$ used for the Structure 2 models is chosen specifically to maximise the ESAG2 log-likelihood.

The residuals of the fitted ESAG2 model, shown in Fig. 2b, show a small amount of serial correlation (points 21–27), but otherwise little to suggest the model is poorly fitting.

5.3 Vectorcardiogram data (involving a mixed-type covariate)

This data set was considered by Chang (1986) in the context of his spherical–spherical regression models. Here, our more general model enables incorporation of other covariates, and of anisotropic errors.

The data themselves are derived from vectorcardiogram measurements of the electrical activity of the heart of children of different ages and genders. The vectorcardiogram involves three leads being connected to the torso produce a time-dependent vector that traces approximately closed curves, each representing a heartbeat cycle, in $\mathbb{R}^3$. Sometimes used as a summary for clinical diagnosis is a unit vector defined as the directional component of the vector at a particular extremum across the cycles. The data comprise such unit vectors derived from data for two different lead placement systems, the Frank system ($\mathbf{y}_i \in \mathbb{S}^2$) and for the McFee system ($\mathbf{u}_i \in \mathbb{S}^2$), for each of 98 children of different ages and gender. Age is represented by a binary variable $A_i \in \{0, 1\}$ (0 meaning aged 2–10 years, and 1 meaning aged 11–18 years) and gender by a variable $G_i \in \{0, 1\}$ (0 for a boy, and 1 for a girl). We aim to regress $\mathbf{y}_i$ on the other variables, and hence take the covariate to be $\mathbf{x}_i = \left(1, \mathbf{u}_i^\top, A_i, G_i\right)^\top$, for $i = 1, \ldots, 98$.

To identify the meaning of the parameters in the parameter matrix $\mathbf{B}$, we write it in the block structure

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1^0 & \mathbf{B}_1^u & \boldsymbol{\beta}_1^A & \boldsymbol{\beta}_1^G \\ \boldsymbol{\beta}_2^0 & \mathbf{B}_2^u & \boldsymbol{\beta}_2^A & \boldsymbol{\beta}_2^G \end{pmatrix}, \tag{26}$$

where $\boldsymbol{\beta}_1^0$, $\boldsymbol{\beta}_1^A$ and $\boldsymbol{\beta}_1^G$ are 3×1 and $\mathbf{B}_1^u$ is 3×3; and $\boldsymbol{\beta}_2^0$, $\boldsymbol{\beta}_2^A$ and $\boldsymbol{\beta}_2^G$ are $s \times 1$ and $\mathbf{B}_2^u$ is $s \times 3$, where $s = 1$ for Structure 1 models and $s = 2$ for Structure 2 models. Setting any of

Table 2 Results for Structure 1 models and submodels fitted to the vectorcardiogram data shown in Fig. 3, and described in Sect. 5.3

Model	$\mathbf{A}_\mathrm{R}$ params set to zero	$\mathbf{A}_\mathrm{S}$ params set to zero	Log-lik. ESAG1	Log-lik Kent1	(*d.f.*)
M_1	–	–	32.12	27.46	(26)
M_2	–	β_2^G	31.79	27.45	(25)
M_3	–	β_2^A	31.22	26.70	(25)
M_4	–	β_2^G, β_2^A	28.78	25.34	(24)
M_5	$\boldsymbol{\beta}_1^G$	β_2^G	31.79	27.45	(22)
M_6	$\boldsymbol{\beta}_1^A$	β_2^A	31.20	26.70	(22)
M_7	$\boldsymbol{\beta}_1^G, \boldsymbol{\beta}_1^A$	β_2^G, β_2^A	28.78	25.34	(18)
M_8	–	$\mathbf{B}_2^u, \beta_2^G, \beta_2^A$	30.81	26.26	(21)
M_9	$\boldsymbol{\beta}_1^G$	”	30.80	26.24	(18)
M_{10}	$\boldsymbol{\beta}_1^A$	”	30.06	25.47	(18)
M_{11}	$\boldsymbol{\beta}_1^G, \boldsymbol{\beta}_1^A$	”	28.03	24.18	(15)
M_{12}	–	$\beta_2^0, \mathbf{B}_2^u, \beta_2^G, \beta_2^A$	30.79	26.12	(20)
M_{13}	$\boldsymbol{\beta}_1^G$	”	28.68	24.87	(17)
M_{14}	$\boldsymbol{\beta}_1^A$	”	29.95	25.15	(17)
M_{15}	$\boldsymbol{\beta}_1^G, \boldsymbol{\beta}_1^A$	”	28.00	23.94	(14)
M_{16}	–	$\beta = 0$ (isotropic errors)	10.26	4.11	(18)

Table 3 Results for Structure 2 models and submodels fitted to the vectorcardiogram data

Model	$\boldsymbol{\mu}$ params set to zero	$\boldsymbol{\gamma}$ params set to zero	Log-lik. ESAG2	Log-lik Kent2	(*d.f.*)
M_1	–	–	54.88	50.54	(30)
M_2	–	$\boldsymbol{\beta}_2^G$	50.04	47.65	(28)
M_3	–	$\boldsymbol{\beta}_2^A$	54.56	50.43	(28)
M_4	–	$\boldsymbol{\beta}_2^G, \boldsymbol{\beta}_2^A$	49.80	47.58	(26)
M_5	$\boldsymbol{\beta}_1^G$	$\boldsymbol{\beta}_2^G$	46.78	45.35	(25)
M_6	$\boldsymbol{\beta}_1^A$	$\boldsymbol{\beta}_2^A$	48.23	42.81	(25)
M_7	$\boldsymbol{\beta}_1^G, \boldsymbol{\beta}_1^A$	$\boldsymbol{\beta}_2^G, \boldsymbol{\beta}_2^A$	43.23	38.93	(20)
M_8	–	$\mathbf{B}_2^u, \boldsymbol{\beta}_2^G, \boldsymbol{\beta}_2^A$ ($\boldsymbol{\gamma} = \text{const}$)	33.64	28.77	(20)
M_9	$\boldsymbol{\beta}_1^G$	”	30.94	25.96	(17)
M_{10}	$\boldsymbol{\beta}_1^A$	”	29.38	22.99	(17)
M_{11}	$\boldsymbol{\beta}_1^G, \boldsymbol{\beta}_1^A$	”	26.16	19.87	(14)
M_{12}	–	$\boldsymbol{\beta}_2^0, \mathbf{B}_2^u, \boldsymbol{\beta}_2^G, \boldsymbol{\beta}_2^A$ (isotropic error)	25.18	15.38	(18)
M_{13}	$\boldsymbol{\beta}_1^G$	”	20.48	10.66	(15)
M_{14}	$\boldsymbol{\beta}_1^A$	”	21.06	12.56	(15)
M_{15}	$\boldsymbol{\beta}_1^G, \boldsymbol{\beta}_1^A$	”	16.60	7.85	(12)

these blocks equal to zero amounts to removing the influence of a particular covariate on particular parameters in the error model. For example, in Structure 2 models, setting $\boldsymbol{\beta}_1^A = \mathbf{0}$ means that the covariate A_i does not influence $\boldsymbol{\mu}$.

Tables 2 and 3, respectively, show the results of fitting Structure 1 and 2 models, and several submodels, to the vectorcardiogram data. Within each table, each of the submodels is nested within M_1, and some of the submodels are further nested within each other. Pairwise comparisons of relevant nested models using likelihood ratio tests described in Sect. 3.3 at 5% level suggest that for both ESAG1 and Kent1 the preferred model is M_{15}. This suggests that Structure 1 is poor for characterising how the response depends on the covariates for this application, to the extent that there is little benefit to retaining the covariates in the model. In contrast, for both ESAG2 and Kent2 the preferred model is M_3, which retains all of the covariates.

Figure 3b shows $\boldsymbol{\eta}$-residuals for ESAG2 models M_3 and M_{15}. For M_3, which is the preferred model, the residuals are reasonably consistent with $N_2(\mathbf{0}, \mathbf{I})$ scatter. For M_{15}, which

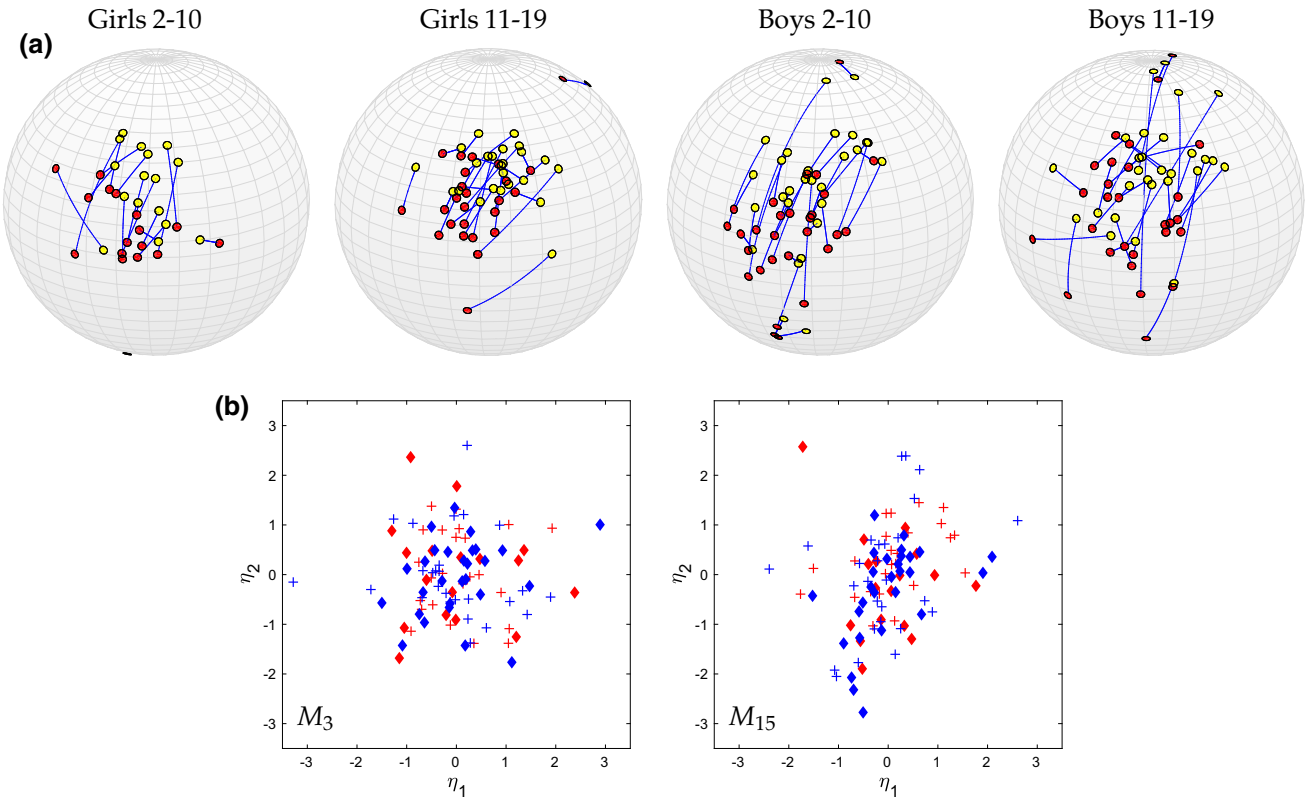


Fig. 3 **a** The vectorcardiogram data described in Sect. 5.3. Red and yellow markers, respectively, indicate covariates, $\{\mathbf{u}_i\}$, and responses, $\{\mathbf{y}_i\}$, and the blue lines indicate the pairings. **b** The $\boldsymbol{\eta}$-residuals for fitted ESAG2 models M_3 (the preferred model) and M_{15} (the model in which age and gender covariates are ignored and errors are assumed isotropic). Red point markers denote girls and blue denotes boys; diamonds denote the 2–10 age group and crosses denote the 11–19 age group. (Color figure online)

assumes isotropic errors and neglects the age and gender covariates, the scatter appears less isotropic, and there are slight differences in the scatter according to age and gender, consistent with there being residual variation due to these factors not being incorporated in the model.

6 Conclusions

The regression models we have introduced are rather more general than existing regression models in the literature, allowing covariates with general structure, and errors that are nonisotropic. We have also introduced novel model-based residuals that enable simple visual diagnostics to check fitted models, to identify for example any residual structure dependent on a covariate, any serial correlation or any outliers, and to explore adequacy of the error models.

For the anisotropic error model, there is little to choose on statistical grounds between using Kent or ESAG, though we have found occasions for models based on the Kent that the likelihood function is harder to maximise numerically (perhaps owing to roughness in the likelihood approximation arising from approximating the normalising constant). Models based on ESAG are free from this issue, and the computation of the ESAG likelihood is much faster. Of the two model structures we considered, models with Structure 2 tended to perform better; such models are also simpler and enable the influence of particular covariates to be related more directly to the response variable. On the foregoing grounds, ESAG2 models are our preferred ones.

The likelihood framework in which we have developed the models makes it very easy to use classical methods to compare nested models of different complexity, in particular to test hypotheses about significance of regression or the anisotropy of errors. Indeed, applying such tests for the examples considered provides strong support that the regression modelling generalisations we have developed are warranted.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

Preliminary transformation in Sect. 3. In this section, we define the orthogonal matrix $\mathbf{Q} = [\boldsymbol{\xi}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2]$ in (3). In the "correlation" approach to regression, where we view the pairs $(\mathbf{y}_i, \mathbf{x}_i)$ as being independent and identically distributed, we define $F_{\mathbf{X}}(.)$ to be the marginal distribution of the $\mathbf{x}_i$; and in the "conditional" approach to regression, where only the conditional distributions of $\mathbf{y}_i$ given $\mathbf{x}_i$ are specified, define $F_{\mathbf{X}}(.) = n^{-1}\sum_{i=1}^{n}\delta_{\mathbf{x}_i}(.)$, where $\delta_{\mathbf{x}}(A) = 1$ if $\mathbf{x} \in A$ and $\delta_{\mathbf{x}}(A) = 0$ otherwise, with $A \subseteq \mathbb{R}^p$. Then, we define

$$\boldsymbol{\xi} = \int_{\mathbf{x}\in\mathbb{R}^p} \mathbf{g}(\mathbf{x})\mathrm{d}F_{\mathbf{X}}(\mathbf{x}) \Big/ \left|\left| \int_{\mathbf{x}\in\mathbb{R}^p} \mathbf{g}(\mathbf{x})\mathrm{d}F_{\mathbf{X}}(\mathbf{x}) \right|\right|, \tag{27}$$

where $\mathbf{g}(\mathbf{x}) = E[\mathbf{y}|\mathbf{x}]$ denotes conditional expectation under the population model, and $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are unit eigenvectors corresponding to the larger and smaller positive eigenvalues of

$$\left(\mathbf{I}_3 - \boldsymbol{\xi}\boldsymbol{\xi}^\top\right) \int_{\mathbf{x}\in\mathbb{R}^p} \mathbf{g}(\mathbf{x})\mathbf{g}(\mathbf{x})^\top \mathrm{d}F_{\mathbf{X}}(\mathbf{x}) \left(\mathbf{I}_3 - \boldsymbol{\xi}\boldsymbol{\xi}^\top\right). \tag{28}$$

A key point to note is that, when we apply an orthogonal transformation $\mathbf{y} \mapsto \mathbf{U}\mathbf{y}$ to $\mathbf{y}$, then $\mathbf{Q} \mapsto \mathbf{U}\mathbf{Q}$ and

$$\mathbf{Q}^\top\mathbf{y} \mapsto (\mathbf{U}\mathbf{Q})^\top \mathbf{U}\mathbf{y} = \mathbf{Q}^\top\mathbf{U}^\top\mathbf{U}\mathbf{y} = \mathbf{Q}^\top\mathbf{y},$$

so that structures (3) and (4) are invariant with respect to orthogonal transformations of the $\mathbf{y}_i$.

When fitting models with structures (3) and (4) we estimate $\mathbf{Q}$ using its sample analogue defined in Sect. 3.1. Providing the $\mathbf{y}_i$ have absolutely continuous distributions and the sample size n is at least 3, $\hat{\mathbf{Q}}$ is well defined with probability 1 even if the population version $\mathbf{Q}$ is not well defined, i.e. even if the denominator of (27) is 0 and/or the positive eigenvalues of (28) are equal. If $\mathbf{Q}$ is well defined then $\hat{\mathbf{Q}}$ estimates $\mathbf{Q}$ consistently.

Finally, we consider what happens if we start with the original coordinate system for the $\mathbf{y}_i$, in which case the models concerned are equivariant rather than invariant with respect to orthogonal transformations of the $\mathbf{y}_i$. In the case of Structure 1 in (3), the orthogonal matrices $\boldsymbol{\Gamma}(\mathbf{x}_i)$ transform according to $\boldsymbol{\Gamma}(\mathbf{x}_i) \mapsto \mathbf{Q}\boldsymbol{\Gamma}(\mathbf{x}_i)$, $i = 1, \ldots, n$, with κ and β unchanged.

For Structure 2, in which $\mathbf{Q} \in O(3)$ is considered a tuning parameter and optimised with respect to, the $\boldsymbol{\mu}(\mathbf{x}_i)$ and $\boldsymbol{\gamma}(\mathbf{x}_i)$ are invariant to orthogonal transformations of the $\mathbf{y}_i$.

The range of the Cayley transform. Write

$$\mathbf{A} = \begin{pmatrix} 0 & a & b \\ -a & 0 & c \\ -b & -c & 0 \end{pmatrix}.$$

Then, elementary calculations show that

$$(\mathbf{I}_3 - \mathbf{A})(\mathbf{I}_3 + \mathbf{A})^{-1} = \frac{1}{1 + a^2 + b^2 + c^2}\mathbf{B} \tag{29}$$

where $\mathbf{B} = (B_{jk})_{j,k=1}^3$, with $B_{11} = 1 + c^2 - a^2 - b^2$,

$$\begin{aligned} &B_{22} = 1 + b^2 - a^2 - c^2, \quad B_{33} = 1 + a^2 - b^2 - c^2, \\ &B_{12} = 2(a - bc), \quad B_{13} = 2(ac + b), \quad B_{21} = -2(a + bc), \\ &B_{23} = 2(c - ab), \quad B_{31} = 2(ac - b), \quad B_{32} = -2(ab + c). \end{aligned}$$

To identify the set of rotations not in the range of the Cayley map, we need to consider all limits as subsets of a, b and c go to $\pm\infty$. In the most general case, if $a \mapsto \lambda a$, $b \mapsto \lambda b$ and $c \mapsto \lambda c$, then as $\lambda \to \infty$ and assuming not all of a, b and c are 0, the matrix in (29) converges to

$$\frac{1}{a^2 + b^2 + c^2} \begin{pmatrix} c^2 - a^2 - b^2 & -2bc & 2ac \\ -2bc & b^2 - a^2 - c^2 & -2ab \\ 2ac & -2ab & a^2 - b^2 - c^2 \end{pmatrix}.$$

The trace of this matrix is equal to -1, and in a certain sense these matrices are as different from the identity matrix (which has trace $+3$) as it is possible for a 3-by-3 rotation matrix to be. Moreover, this family can be parametrised by $\tau_1 = a/R$, $\tau_2 = b/R$ and $\tau_3 = c/R$ where $R = \sqrt{a^2 + b^2 + c^2}$ and, since $\tau_1^2 + \tau_2^2 + \tau_3^2 = 1$, it follows that this set is 2-dimensional rather than 3-dimensional, and so has measure 0 with respect to Haar (or geometric) measure on the space of rotations. A similar result holds for the Cayley transform in higher dimensions. That the Cayley transform is not a surjection seems to be of no practical or computational significance in the present setting, however.

Proof of Lemma 1 Start with the Kent density function as given in (5). In this parametrisation, we already have the mean vector $\boldsymbol{\mu}$ as an unconstrained vector of parameters. Now, to allow the axes of symmetry, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, to be an arbitrary rotation of $\tilde{\boldsymbol{\xi}}_1$ and $\tilde{\boldsymbol{\xi}}_2$ as defined in (12) write the matrix $\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top - \boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top$ as follows:

$$\begin{aligned} &\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top - \boldsymbol{\xi}_2\boldsymbol{\xi}_2^\top \\ &= \left(\cos\psi\tilde{\boldsymbol{\xi}}_1 + \sin\psi\tilde{\boldsymbol{\xi}}_2\right)\left(\cos\psi\tilde{\boldsymbol{\xi}}_1 + \sin\psi\tilde{\boldsymbol{\xi}}_2\right)^\top \\ &\quad - \left(-\sin\psi\tilde{\boldsymbol{\xi}}_1 + \cos\psi\tilde{\boldsymbol{\xi}}_2\right)\left(-\sin\psi\tilde{\boldsymbol{\xi}}_1 + \cos\psi\tilde{\boldsymbol{\xi}}_2\right)^\top \\ &= \tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_1^\top\left(\cos^2\psi - \sin^2\psi\right) + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_2^\top\left(\sin^2\psi - \cos^2\psi\right) \\ &\quad + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_1^\top(\sin\psi\cos\psi + \cos\psi\sin\psi) \\ &\quad + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_1^\top(\cos\psi\sin\psi + \sin\psi\cos\psi) \\ &= \left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_1^\top - \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_2^\top\right)\left(\cos^2\psi - \sin^2\psi\right) \\ &\quad + \left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_2^\top + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_1^\top\right)2\cos\psi\sin\psi \end{aligned}$$

$$= \cos(2\psi)\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_1^\top - \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_2^\top\right) + \sin(2\psi)\left(\tilde{\boldsymbol{\xi}}_1\tilde{\boldsymbol{\xi}}_2^\top + \tilde{\boldsymbol{\xi}}_2\tilde{\boldsymbol{\xi}}_1^\top\right).$$

As in the parametrisation of the ESAG distribution, we want the unknown parameters to be unconstrained, therefore define parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^\top$

$$\gamma_1 = \beta\cos(2\psi), \qquad \gamma_2 = \beta\sin(2\psi).$$

Proof of Proposition 1 This proof follows a similar course to the proof in Presnell et al. (1998). The log-likelihood is

$$l(\mathbf{B}_1) = \sum_{i=1}^{n} -\frac{3}{2}\log(2\pi) - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\mu}_i + \psi(\boldsymbol{\mu}_i^\top\mathbf{y}_i),$$

where $\boldsymbol{\mu}_i = \mathbf{B}_1\mathbf{x}_i = (\mathbf{x}_i \otimes \mathbf{I}_3)^\top \operatorname{vec}\mathbf{B}_1$ and $\otimes$ denotes Kronecker product, and

$$\psi(\alpha) = \log\left(\alpha + \frac{(1+\alpha^2)\Phi(\alpha)}{\phi(\alpha)}\right). \tag{30}$$

Then,

$$-\frac{\partial^2 l(\mathbf{B}_1)}{\partial \operatorname{vec}\mathbf{B}_1\, \partial \operatorname{vec}\mathbf{B}_1^\top} = \sum_{i=1}^{n}(\mathbf{x}_i \otimes \mathbf{I}_3)\left(\mathbf{I}_3 - \psi''(\boldsymbol{\mu}_i^\top\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^\top\right)(\mathbf{x}_i \otimes \mathbf{I}_3)^\top, \tag{31}$$

which is positive definite if $\left(\mathbf{I}_3 - \psi''(\boldsymbol{\mu}_i^\top\mathbf{y}_i)\mathbf{y}_i\mathbf{y}_i^\top\right)$ is positive definite for all i. Since the eigenvalues of the latter matrix are $1,\ 1,\ 1-\psi''(\boldsymbol{\mu}_i^\top\mathbf{y}_i)$, it is sufficient to show that $1-\psi''(\alpha) > 0$ for all α. From (30), following some re-arrangement,

$$1 - \psi''(\alpha) = \frac{1}{(1+\alpha^2)^2\Phi(\alpha)^2}\left(1 - \alpha\exp\left(-\psi(\alpha)\right)\right)^2 h(\alpha),$$

where

$$h(\alpha) = -2\Phi(\alpha)(\alpha\phi(\alpha) + (1+\alpha^2)\Phi(\alpha)) + 4(\phi(\alpha) + \alpha\Phi(\alpha))^2, \tag{32}$$

and hence it is sufficient now to show that $h(\alpha) > 0$ for all α. For $\alpha < 0$, using Mills' ratio $\Phi(\alpha) > -\alpha\phi(\alpha)(1+\alpha^2)^{-1}$, and substituting for the second instance of $\Phi(\alpha)$ in (32) gives

$$h(\alpha) > 4\left(\phi(\alpha) + \alpha\Phi(\alpha)\right)^2 > 0.$$

For $0 \le \alpha < 1$, $h'(\alpha) \ge 0$ since

$$h'(\alpha) = \left(2 - 2\alpha^2\right)\phi(\alpha)\Phi(\alpha) - 2\alpha\phi(\alpha)^2 + 4\alpha\Phi(\alpha)^2 > 2\alpha(2\Phi(\alpha)^2 - \phi(\alpha)^2)$$

and $2\Phi(\alpha)^2 - \phi(\alpha)^2 > 0$; since also $h(0) > 0$ therefore $h(\alpha) > 0$. Finally, for $\alpha \ge 1$,

$$h(\alpha) = 6\alpha\phi(\alpha)\Phi(\alpha) + 2\left(2\phi(\alpha)^2 + (\alpha^2-1)\Phi(\alpha)^2\right) > 0.$$

□

References

Chang, T.: Spherical regression. Ann. Stat. **14**, 907–924 (1986)

Cornea, E., Zhu, H., Kim, P., Ibrahim, J.G.: Regression models on Riemannian symmetric spaces. J. R. Stat. Soc. Ser. B **79**, 463–482 (2017)

Di Marzio, M., Panzera, A., Taylor, C.: Nonparametric regression for spherical data. J. Am. Stat. Assoc. **109**, 748–763 (2014)

Fisher, N.I., Lee, A.J.: Regression models for angular response. Biometrics **48**, 665–677 (1992)

Hamsici, O.C., Martinez, A.M.: Spherical-homoscedastic distributions: the equivalency of spherical distributions and normal distributions in classification. J. Mach. Learn. Res. **8**, 1583–1623 (2007)

Jupp, P.E.: Residuals for directional data. J. Appl. Stat. **15**, 137–147 (1988)

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

Kent, J.T., Ganeiber, A.M., Mardia, K.V.: A new unified approach for the simulation of a wide class of directional distributions. J. Comput. Graph. Stat. **27**, 291–301 (2018)

Kume, A., Sei, T.: On the exact maximum likelihood inference of Fisher–Bingham distributions using an adjusted holonomic gradient method. Stat. Comput. **28**, 835–847 (2018)

Kume, A., Preston, S.P., Wood, A.T.A.: Saddlepoint approximations for the normalising constant of Fisher–Bingham distributions on products of spheres and Stiefel manifolds. Biometrika **100**, 971–984 (2013)

Lin, L., St Thomas, B., Zhu, H., Dunson, D.B.: Extrinsic local regression on manifold-valued data. J. Am. Stat. Assoc. **112**, 1261–1273 (2017)

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

Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, London (1979)

Paine, P.J., Preston, S.P., Tsagris, M., Wood, A.T.A.: The elliptically symmetric angular Gaussian distribution. Stat. Comput. **28**, 689–697 (2017)

Presnell, B., Morrison, S.P., Littel, R.C.: Projected multivariate linear models for directional data. J. Am. Stat. Assoc. **93**, 1068–1077 (1998)

Rivest, L.-P.: Spherical regression for concentrated Fisher–von Mises distributions. Ann. Stat. **17**, 307–317 (1989)

Rosenthal, M., Wei, W., Klassen, E., Srivastava, A.: Spherical regression models using projective linear transformations. J. Am. Stat. Assoc. **109**, 1615–1624 (2014)

Scealy, J.L., Welsh, A.H.: Regression for compositional data by using distributions defined on the hyper-sphere. J. R. Stat. Soc. Ser. B **73**, 351–375 (2011)

Wang, F., Gelfand, A.E.: Directional data analysis under the general projected normal distribution. Stat. Methodol. **10**, 113–127 (2013)

XPlanet: Real time cloud map. http://xplanet.sourceforge.net/clouds.php. Accessed 28 June 2018

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.