Electronic Journal of Statistics
Vol. 18 (2024) 301–334
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2210

A new parameterization for elliptically symmetric angular Gaussian distributions of arbitrary dimension

Zehao Yu[1] **and Xianzheng Huang**[2]

[1]*Department of Statistics, University of South Carolina*
e-mail: zehaoy@email.sc.edu

[2]*Department of Statistics, University of South Carolina*
e-mail: huang@stat.sc.edu

Abstract: We consider a class of angular Gaussian distributions that allows different degrees of isotropy for directional random variables of arbitrary dimension. To incorporate constraints imposed on the original model parameters, we propose a new parameterization of the distribution so that all new model parameters are free of constraints. Via the new parameterization, we translate the original problem of maximum likelihood estimation subject to complex constraints to a routine optimization problem free of constraints, which in turn leads to theoretically sound and numerically stable procedures for drawing likelihood-based inference. Byproducts from the likelihood-based inference are used to develop graphical and numerical diagnostic tools for assessing goodness of fit of this distribution in a data application. Simulation study and application to data from a hydrogeology study are used to demonstrate implementation and performance of the inference procedures and diagnostics methods.

Keywords and phrases: Angular Gaussian distribution, compositional data, maximum likelihood, residual.

Received July 2022.

1. Introduction

Directional data are ubiquitous in many scientific fields. For example, wave directions are directional data studied in oceanography [30], wind directions are of interest in meteorology [1], and protein backbone structures are directional data researchers study in biology [16]. These exemplify directional data of dimension no higher than three. Other examples of low dimensional direction data include migratory movements of animals, and measurements on a periodic scale, such as weekdays and hours. Directional data of higher dimensions arise in bioinformatics and hydrogeology, among many other fields of research. For example, gene expression data associated with a large number of genes for each experimental unit are often standardized to preserve directional characteristics when studying the fluctuation of gene expressions over cell cycles [8]. By transforming the original gene expression data on a high dimensional Euclidean space to a unit hypersphere, one ignores absolute expression levels and can obtain better

clustering of genes that are functionally related [4]. Although not on a spherical space, compositional data on a simplex [22, 2] can be easily transformed to become directional data. For instance, microbiome data are often summarized as the composition of bacterial taxa so that one can focus on the microbial relative abundances as opposed to absolute abundances in microbiome analysis [28]. A compositional data point is a vector with non-negative components that sum to one, hence a component-wise square-root transformation of this vector yields a vector on a unit hypersphere [26].

Each of the above examples of directional data can be viewed as realizations of a random variable supported on a unit-radius d-dimensional spherical space defined by $\mathbb{S}^{d-1} = \{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\| = 1\}$, for $d \geq 2$, where $\|\mathbf{y}\|$ is the L_2-norm of $\mathbf{y}$. [14] provided a brief survey of statistical methods for analyzing circular data, i.e., directional data with $d = 2$. Two general strategies for constructing a circular distribution are highlighted in this review paper: one uses a "wrapped" circular version of a random variable supported on $\mathbb{R}$ to formulate a circular distribution; the other deduces a circular distribution via projecting a univariate random variable on $\mathbb{R}$ or a bivariate random variable on $\mathbb{R}^2$ onto the circle. Both strategies have been generalized and used to formulate directional distributions on $\mathbb{S}^{d-1}$ for $d > 2$. With the Gaussian distribution playing an important role in statistics, it is not surprising that directional distributions originating from a Gaussian distribution have been most studied and adopted in practice, including the so-called wrapped normal distribution and projected normal distribution, with more attention on the latter in recent literature. In particular, [23] used a projected multivariate normal distribution to construct a regression model for a circular response and linear predictors, and employed the maximum likelihood method to infer unknown parameters. [32] incorporated projected normal distributions to develop Bayesian hierarchical models for analyzing circular data. [11] proposed Bayesian inferential method for directional data of arbitrary dimension, again modelled by projected normal distributions.

Projected normal distributions are also referred to as angular Gaussian distributions. Different angular Gaussian distributions are created by imposing different constraints on the parameter space associated with a multivariate Gaussian distribution in order to resolve the non-identifiability issue that arises when the support of a random variable changes from a Euclidean space to a spherical space. [20] imposed constraints on the mean vector and variance-covariance matrix of a Gaussian distribution so that the resultant angular Gaussian distribution is identifiable and, more interestingly, elliptically symmetric. The authors thus coined their proposed distribution as the elliptically symmetric angular Gaussian distribution, ESAG for short. [21] further developed regression models for directional data assuming an ESAG distribution for the response given covariates. Both works on ESAG focus on directional data with $d \leq 3$. More recently, [27] proposed a new directional distribution, called scaled von Mises-Fisher distribution, using grouped transformations of the von Mises-Fisher distribution to achieve elliptical symmetry. The authors used this new distribution to model archeomagnetic data that can be converted to directional data with $d = 3$. The feature of elliptical symmetry of a distribution makes capturing cer-

tain anisotropic pattern of directional data possible. An added benefit of ESAG is that the normalization constant in its probability density function is much easier to compute compared to many existing directional distributions, such as the Kent distribution [13]. This makes maximum likelihood estimation under the ESAG model for directional data more straightforward.

To incorporate the constraints imposed on the mean vector and variance-covariance matrix of a Gaussian distribution when formulating ESAG, [20] designed a parameterization of ESAG when $d = 3$, which allows one to bypass the complicated problem of optimization with constraints when finding the maximum likelihood estimators of the induced parameters. But their parameterization cannot be easily generalized to cases with $d > 3$. This limits the use of ESAG in applications where directional data of higher dimension are observed. The first contribution of our study presented in this paper is a novel parameterization of ESAG of arbitrary dimension that allows one to bypass optimization subject to complicated constraints on model parameters in maximum likelihood estimation. This new parameterization of ESAG for $d \geq 3$ is presented in Section 2. Under the new parameterization, maximum likelihood estimation translates to a routine numerical problem of optimization without constraints, as we describe in Section 3. A legitimate concern in any parametric modelling is potential violations of certain model assumptions in a given application. To address this concern, we propose model diagnostics methods that exploit directional residuals in Section 4, which constitutes a second major contribution of our study. Operating characteristics of the proposed model diagnostics methods are demonstrated in simulation study in Section 5. In Section 6, we entertain data from hydrogeological research, where we fit ESAG to transformed compositional data from different geographic locations. Section 7 summarizes the contributions of the study and outlines the follow-up research agenda.

2. The ESAG distribution

2.1. Constraints on parameters

Let $\mathbf{X}$ be a d-dimensional Gaussian variable with mean $\boldsymbol{\mu}$ and variance-covariance $\mathbf{V}$, i.e., $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \mathbf{V})$. Then the normalized variable, $\mathbf{Y} = \mathbf{X}/\|\mathbf{X}\|$, follows an angular Gaussian distribution, $\mathrm{AG}(\boldsymbol{\mu}, \mathbf{V})$, supported on $\mathbb{S}^{d-1}$. Parameters in $\boldsymbol{\mu}$ and $\mathbf{V}$ associated with $\mathrm{AG}(\boldsymbol{\mu}, \mathbf{V})$ are not identifiable because $\mathbf{X}/\|\mathbf{X}\|$ and $c\mathbf{X}/\|c\mathbf{X}\|$ are equal for $c > 0$, and thus they follow the same angular distribution, even though $\mathbf{X}$ and $c\mathbf{X}$ have different mean or/and variance-covariace when $c \neq 1$. To construct an identifiable angular Gaussian distribution, [20] impose the following two sets of constraints on $\boldsymbol{\mu}$ and $\mathbf{V}$, where $\det(\cdot)$ refers to the determinant of a matrix,

$$\mathbf{V}\boldsymbol{\mu} = \boldsymbol{\mu}, \tag{2.1}$$

$$\det(\mathbf{V}) = 1, \tag{2.2}$$

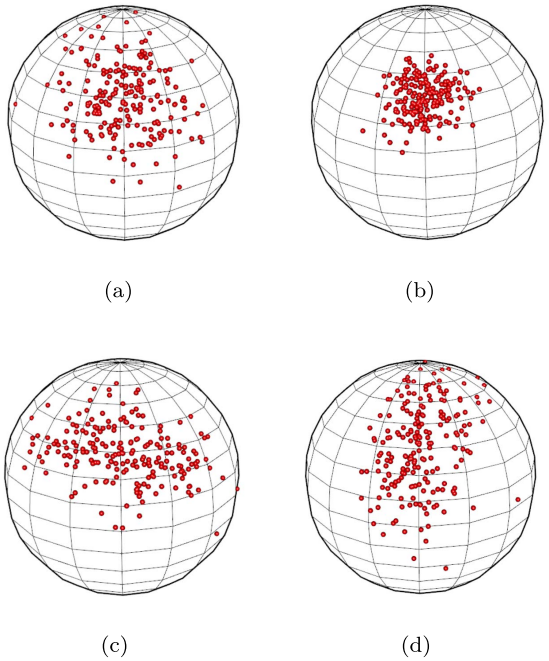

FIG 1. *Four random samples from ESAG$_2(\boldsymbol{\mu}, \mathbf{V})$ with $\boldsymbol{\mu}$ and $\mathbf{V}$ specified by (a)–(d) in Section 2.1.*

leading to the ESAG distribution, with the probability density function given by

$$f(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V}) = \frac{(2\pi)^{-(d-1)/2}}{(\mathbf{y}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y})^{d/2}} \exp\left[\frac{1}{2}\left\{\frac{(\mathbf{y}^{\mathrm{T}}\boldsymbol{\mu})^2}{\mathbf{y}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}} - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}\right\}\right] M_{d-1}\left\{\frac{\mathbf{y}^{\mathrm{T}}\boldsymbol{\mu}}{(\mathbf{y}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y})^{1/2}}\right\}, \tag{2.3}$$

where $M_{d-1}(t) = (2\pi)^{-1/2}\int_0^\infty x^{d-1}\exp\{-(x-t)^2/2\}dx$. Henceforth, we say that $\mathbf{Y}$ follows a $(d-1)$-dimensional ESAG, or $\mathbf{Y} \sim \mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, if $\mathbf{Y}$ follows a distribution specified by the density in (2.3) with constraints in (2.1) and (2.2).

Figure 1 presents four random samples scattering on 3-dimensional spheres, generated from ESAG$_2(\boldsymbol{\mu}, \mathbf{V})$ with the following parameters specifications, where $\mathbf{1}_d$ is a vector of d ones and $\mathbf{I}_d$ is d-dimensional identity matrix:

(a) $\boldsymbol{\mu} = 2 \times \mathbf{1}_3$, $\mathbf{V} = \mathbf{I}_3$;
(b) $\boldsymbol{\mu} = 4 \times \mathbf{1}_3$, $\mathbf{V} = \mathbf{I}_3$;
(c) $\boldsymbol{\mu} = 2 \times \mathbf{1}_3$,
$$\mathbf{V} = \begin{bmatrix} 1.57 & -0.08 & -0.50 \\ -0.08 & 0.74 & 0.34 \\ -0.50 & 0.34 & 1.16 \end{bmatrix};$$
(d) $\boldsymbol{\mu} = 2 \times \mathbf{1}_3$,
$$\mathbf{V} = \begin{bmatrix} 0.74 & -0.08 & 0.34 \\ -0.08 & 1.57 & -0.50 \\ 0.34 & -0.50 & 1.16 \end{bmatrix}.$$

Comparing the four data clouds depicted in Figure 1, one can see that a larger $\|\boldsymbol{\mu}\|$ leads to less variability in a random sample (e.g., contrasting (a) with (b)); and $\mathbf{V}$ also influences the orientation of the data cloud (e.g., comparing (a), (c), and (d)). In other words, $\|\boldsymbol{\mu}\|$ controls the overall concentration, with a higher concentration (i.e., a larger $\|\boldsymbol{\mu}\|$) indicating a lower overall variability, whereas $\mathbf{V}$ dictates orientation of dispersion in different subspaces on the hypersphere.

Because the dimension of the parameter space associated with $N_d(\boldsymbol{\mu}, \mathbf{V})$ is $d(d+3)/2$, and there are $d+1$ constraints imposed by (2.1) and (2.2), there are at most $p = (d-1)(d+2)/2$ identifiable parameters for $\mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$. Let $\boldsymbol{\Omega}$ be the $p \times 1$ parameter vector that specifies $\mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$. To facilitate likelihood-based inference, it is desirable to formulate $\boldsymbol{\Omega}$ so that the parameter space is $\mathbb{R}^p$. For this purpose, we define $\boldsymbol{\Omega} = (\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, where, clearly, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}} \in \mathbb{R}^d$, and thus $\boldsymbol{\gamma} \in \mathbb{R}^{(d-2)(d+1)/2}$ includes parameters needed to specify $\mathbf{V}$ that satisfies (2.1) and (2.2) after $\boldsymbol{\mu}$ is given.

The parameterization leading to $\boldsymbol{\gamma}$ starts from the spectral decomposition of $\mathbf{V}$,

$$\mathbf{V} = \sum_{j=1}^{d} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^{\mathrm{T}}, \tag{2.4}$$

where $\lambda_1, \ldots, \lambda_d \in (0, +\infty) \triangleq \mathbb{R}_+$ are eigenvalues of $\mathbf{V}$, and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d$ are the corresponding orthonormal eigenvectors. According to (2.1), one of the eigenvalues of $\mathbf{V}$ is equal to 1, with $\boldsymbol{\mu}$ being the corresponding (non-zero) eigenvector. Without loss of generality, we set $\lambda_d = 1$ and $\boldsymbol{\xi}_d = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. To this end, once $\boldsymbol{\mu}$ is given, one needs to formulate $\boldsymbol{\gamma}$ so that it can be mapped to $\lambda_1, \ldots, \lambda_{d-1}$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{d-1}$, through which $\mathbf{V}$ is determined via (2.4). In what follows, we present the derivations leading to $\boldsymbol{\gamma}$ in three steps: (i) parameterizing $\lambda_1, \ldots, \lambda_{d-1}$; (ii) parameterizing $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{d-1}$; (iii) grouping new parameters from Steps (i) and (ii), then relating each group of new parameters to entries in $\boldsymbol{\gamma}$.

2.2. Step (i): parameterization for eigenvalues of $\mathbf{V}$

Now that we set $\lambda_d = 1$, and by the constraint $\prod_{j=1}^{d-1} \lambda_j = 1$ implied by (2.2), we only need $d-2$ parameters to specify the first $d-1$ eigenvalues. Without loss of generality, we let $\lambda_1 \leq \cdots \leq \lambda_{d-1}$, then write $\lambda_j = (r_{j-1}+1)\lambda_{j-1}$, where $r_{j-1} \geq 0$, for $j = 2, \ldots, d-1$. Using the constraint $\prod_{j=1}^{d-1} \lambda_j = 1$, one can show that

$$\lambda_1 = \left\{ \prod_{j=1}^{d-2} (r_j + 1)^{d-(j+1)} \right\}^{-1/(d-1)} \quad \text{and} \quad \lambda_j = \lambda_1 \prod_{k=1}^{j-1} (r_k + 1), \ \text{for } j = 2, \ldots, d-1. \tag{2.5}$$

In effect, we parameterize the first $d-1$ eigenvalues of $\mathbf{V}$ using $d-2$ non-negative new parameters, $r_1, \ldots, r_{d-2}$. We call these new parameters radial parameters for a reason to become clear in Step (iii).

2.3. Step (ii): parameterization for eigenvectors of $\mathbf{V}$

With $\boldsymbol{\xi}_d = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ as the eigenvector corresponding to the eigenvalue $\lambda_d = 1$, we now parameterize the remaining $d-1$ eigenvectors $\{\boldsymbol{\xi}_j\}_{j=1}^{d-1}$. We first define an orthonormal basis of $\mathbb{R}^d$, $(\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_d)$, with $\tilde{\boldsymbol{\xi}}_j = \mathbf{u}_j/\|\mathbf{u}_j\|$, for $j = 1, \ldots, d$, and

$$\mathbf{u}_j = \begin{cases} (-\mu_2,\, \mu_1,\, 0, \ldots, 0)^{\mathrm{T}}, & \text{for } j = 1, \\ (\mu_1\mu_{j+1}, \ldots, \mu_j\mu_{j+1},\, -\sum_{k=1}^{j} \mu_k^2,\, 0, \ldots, 0)^{\mathrm{T}}, & \text{for } j = 2, \ldots, d-1, \\ \boldsymbol{\mu} & \text{for } j = d. \end{cases} \tag{2.6}$$

If (2.6) yields $\mathbf{u}_j = \mathbf{0}_d$, for $j \in \{1, \ldots, d-1\}$, then we set $\mathbf{u}_j = \mathbf{e}_j$, i.e., the unit vector with 1 at the j-th entry. This yields an orthonormal basis $(\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_d)$ uniquely determined by $\boldsymbol{\mu}$, and thus no new parameters are introduced in formulating this basis.

By (2.6), $\boldsymbol{\xi}_d = \tilde{\boldsymbol{\xi}}_d$. Then we let $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{d-1}) = (\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_{d-1})\mathcal{R}_{d-1}$, where $\mathcal{R}_{d-1}$ is a $(d-1)$-dimensional rotation matrix that depends on $(d-2)(d-1)/2$ new parameters introduced next. According to [18], $\mathcal{R}_{d-1}$ can be expressed as a product of $(d-2)(d-1)/2$ plane rotation matrices, for $d > 3$, with each plane rotation matrix depending on a longitude angle in $[-\pi, \pi)$ or a latitude angle in $[0, \pi]$. More specifically,

$$\mathcal{R}_{d-1} = \left[\prod_{m=1}^{d-3} \left\{ R_{12}^*(\theta_{d-m-1}) \prod_{j=1}^{d-m-2} R_{j+1,j+2}^*(\phi_{1-j+(d-m-1)(d-m-2)/2}) \right\} \right] \times R_{12}^*(\theta_1), \tag{2.7}$$

where $\theta_1, \ldots, \theta_{d-2} \in [-\pi, \pi)$ are longitude angles, $\phi_1, \ldots, \phi_{(d-2)(d-3)/2} \in [0, \pi]$ are latitude angles, and $R_{jk}^*(\cdot)$ is a $(d-1)$-dimensional plane rotation matrix resulting from replacing the (j,j), (j,k), (k,j), and (k,k) entries of $\mathbf{I}_{d-1}$ by $\cos(\cdot)$, $-\sin(\cdot)$, $\sin(\cdot)$, and $\cos(\cdot)$, respectively. [26] exploited the same formulation of a rotation matrix to parameterize the Kent distribution. Since rotating a set of orthonormal vectors yields another set of orthonormal vectors, we have $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{d-1})$ as $d-1$ orthonormal eigenvectors of $\mathbf{V}$ that are orthogonal to $\boldsymbol{\xi}_d$.

To recap, we introduce $(d-2)(d-1)/2$ angles in (2.7) as new parameters in Step (ii) to parameterize the first $d-1$ eigenvectors of $\mathbf{V}$ after $\boldsymbol{\mu}$ is given. We call these angles orientation parameters in the sequel.

TABLE 1
Transformations linking radial and orientation parameters in $\widetilde{\boldsymbol{\Omega}}$ to $\boldsymbol{\gamma} \in \mathbb{R}^9$ when $d = 5$.

3 groups of parameters in $\widetilde{\boldsymbol{\Omega}}$	Spherical coordinates $\Downarrow$ Cartesian coordinates	Cartesian coordinates $\Downarrow$ Spherical coordinates
r_1, θ_1	$\begin{cases}\gamma_{1,1} = r_1 \cos\theta_1, \\ \gamma_{1,2} = r_1 \sin\theta_1\end{cases}$ $\rightarrow \tilde{\boldsymbol{\gamma}}_1 = (\gamma_{1,1}, \gamma_{1,2})^{\mathrm{T}}$	$\begin{cases}r_1 = \lVert\tilde{\boldsymbol{\gamma}}_1\rVert, \\ \theta_1 = \operatorname{atan2}(\gamma_{1,2}, \gamma_{1,1})\end{cases}$ $\rightarrow (r_1, \theta_1)$
r_2, θ_2, ϕ_1	$\begin{cases}\gamma_{2,1} = r_2 \cos\phi_1, \\ \gamma_{2,2} = r_2 \sin\phi_1 \cos\theta_2, \\ \gamma_{2,3} = r_2 \sin\phi_1 \sin\theta_2\end{cases}$ $\rightarrow \tilde{\boldsymbol{\gamma}}_2 = (\gamma_{2,1}, \boldsymbol{\gamma}_{2,2}, \gamma_{2,3})^{\mathrm{T}}$	$\begin{cases}r_2 = \lVert\tilde{\boldsymbol{\gamma}}_2\rVert, \\ \theta_2 = \operatorname{sign}(\gamma_{2,3}) \arccos \dfrac{\gamma_{2,2}}{\sqrt{\gamma_{2,2}^2 + \gamma_{2,3}^2}}, \\ \phi_1 = \arccos \dfrac{\gamma_{2,1}}{\lVert\tilde{\boldsymbol{\gamma}}_2\rVert} \triangleq \tilde{\phi}_{2,1}\end{cases}$ $\rightarrow (r_2, \theta_2, \phi_1) = (r_2, \theta_2, \tilde{\phi}_{2,1})$
$r_3, \theta_3, \phi_2, \phi_3$	$\begin{cases}\gamma_{3,1} = r_3 \cos\phi_2, \\ \gamma_{3,2} = r_3 \sin\phi_2 \cos\phi_3, \\ \gamma_{3,3} = r_3 \sin\phi_2 \sin\phi_3 \cos\theta_3, \\ \gamma_{3,4} = r_3 \sin\phi_2 \sin\phi_3 \sin\theta_3\end{cases}$ $\rightarrow \tilde{\boldsymbol{\gamma}}_3 = (\gamma_{3,1}, \gamma_{3,2}, \gamma_{3,3}, \gamma_{3,4})^{\mathrm{T}}$	$\begin{cases}r_3 = \lVert\tilde{\boldsymbol{\gamma}}_3\rVert, \\ \theta_3 = \operatorname{sign}(\gamma_{3,3}) \arccos \dfrac{\gamma_{3,3}}{\sqrt{\gamma_{3,3}^2 + \gamma_{3,4}^2}}, \\ \phi_2 = \arccos \dfrac{\gamma_{3,1}}{\lVert\tilde{\boldsymbol{\gamma}}_3\rVert} \triangleq \tilde{\phi}_{3,1}, \\ \phi_3 = \arccos \dfrac{\gamma_{3,2}}{\sqrt{\gamma_{3,2}^2 + \gamma_{3,3}^2 + \gamma_{3,4}^2}} \triangleq \tilde{\phi}_{3,2}\end{cases}$ $\rightarrow (r_3, \theta_3, \phi_2, \phi_3) = (r_3, \theta_3, \tilde{\boldsymbol{\phi}}_3^{\mathrm{T}})$, where $\tilde{\boldsymbol{\phi}}_3 = (\tilde{\phi}_{3,1}, \tilde{\phi}_{3,2})^{\mathrm{T}}$
	$\boldsymbol{\gamma} = (\tilde{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \tilde{\boldsymbol{\gamma}}_2^{\mathrm{T}}, \tilde{\boldsymbol{\gamma}}_3^{\mathrm{T}})^{\mathrm{T}}$	$\widetilde{\boldsymbol{\Omega}} = (r_1, r_2, r_3, \theta_1, \theta_2, \theta_3, \tilde{\phi}_{2,1}, \tilde{\boldsymbol{\phi}}_3^{\mathrm{T}})^{\mathrm{T}}$

2.4. Step (iii): relating spherical coordinates to Cartesian coordinates

Gathering the new parameters introduced above, including the radial parameters from Step (i) and the orientation parameters from Step (ii), we define $\widetilde{\boldsymbol{\Omega}} = (r_1, \ldots, r_{d-2}, \theta_1, \ldots, \theta_{d-2}, \phi_1, \ldots, \phi_{(d-2)(d-3)/2})^{\mathrm{T}}$. We now partition $\widetilde{\boldsymbol{\Omega}}$ into $d-2$ groups such that each group of parameters can be viewed as coordinates under a spherical coordinate system of certain dimension, consisting of one radial parameter, one longitude angle ranging over $[-\pi, \pi)$, and, for a spherical coordinate system of dimension higher than one, latitude angle(s) ranging over $[0, \pi]$. Following this partition of $\widetilde{\boldsymbol{\Omega}}$, we exploit the connection between a spherical coordinate system and the corresponding Cartesian coordinate system [5] to transform each group of radial and orientation parameters to a group of parameters in $\boldsymbol{\gamma}$.

For illustration purposes, we use a four dimensional ESAG (thus $d = 5$) as an example to demonstration the grouping and transformations linking $\widetilde{\boldsymbol{\Omega}}$ to $\boldsymbol{\gamma}$. Now with $d = 5$, we need radial and orientation parameters in $\widetilde{\boldsymbol{\Omega}} = (r_1, r_2, r_3, \theta_1, \theta_2, \theta_3, \phi_1, \phi_2, \phi_3)^{\mathrm{T}}$ to specify $\mathbf{V}$ given $\boldsymbol{\mu}$. Table 1 shows in the first column $3 (= d-2)$ groups of parameters forming a partition of $\widetilde{\boldsymbol{\Omega}}$. Recollecting the composition of a set of spherical coordinates, one can see that the three groups of parameters correspond to coordinates under a circular coor-

dinate system (i.e., one-dimensional spherical coordinate system), coordinates under a two-dimensional spherical coordinate system, and those under a three-dimensional spherical coordinate system, respectively. The transformations from each set of spherical coordinates to the corresponding set of Cartesian coordinates are shown in the second column of Table 1. Each set of Cartesian coordinates is viewed as a group of parameters in $\boldsymbol{\gamma}$. Lastly, the inverse transformations that map each set of Cartesian coordinates (as entries of $\boldsymbol{\gamma}$) back to the corresponding spherical coordinates (as entries of $\widetilde{\boldsymbol{\Omega}}$) are given in the third column of Table 1, where all denominators appearing in the transformations are assumed nonzero for simplicity. To signify the grouping of the latitude angles in $\widetilde{\boldsymbol{\Omega}}$, we re-define (ϕ_1, ϕ_2, ϕ_3) as $(\tilde{\phi}_{2,1}, \tilde{\phi}_{3,1}, \tilde{\phi}_{3,2})$ in Table 1, with the first subscript in $\tilde{\phi}_{j,k}$ being the group index. These new notations with a double subscript for the latitude angles will replace the original notations with a single subscript henceforth.

Focusing on the first group of parameters, (r_1, θ_1), in $\widetilde{\boldsymbol{\Omega}}$ in Table 1 allows a closer comparison between our parameterization and that in [20, Section 2.3] for ESAG_2 (thus $d = 3$). Like our new unrestricted parameters in $\tilde{\boldsymbol{\gamma}}_1 = (\gamma_{1,1}, \gamma_{1,2})^{\mathrm{T}}$, they also defined two unrestricted parameters, γ_1 and γ_2 in their notations, as new model parameters of ESAG_2. But their new model parameters are formulated as functions of λ_1 and an orientation parameter falling in $(0, \pi]$. Even though their strategy leads to a simple and interesting expression for $\mathbf{V}^{-1}$ [see Lemma 1 in 20], the formulation of their γ_1 and γ_2 cannot be easily generalized to higher dimentions. In contrast, writing our new parameters as functions of (r_1, θ_1) amounts to transforming circular coordinates to Cartesian coordinates in $\mathbb{R}^2$, the kind of transformation that can be easily generalized to higher dimensions as demonstrated in Table 1.

In general, for $d \geq 3$, we divide $(d-2)(d+1)/2$ parameters in $\widetilde{\boldsymbol{\Omega}}$ into $d-2$ groups, with the first group being (r_1, θ_1), and (if $d > 3$), for $j = 2, \ldots, d-2$, the j-th group being $(r_j, \theta_j, \tilde{\boldsymbol{\phi}}_j)$, where $\tilde{\boldsymbol{\phi}}_j = (\tilde{\phi}_{j,1}, \ldots, \tilde{\phi}_{j,j-1})^{\mathrm{T}}$. To adapt to the grouping for $\widetilde{\boldsymbol{\Omega}}$, we also define $\boldsymbol{\gamma}$ as $d-2$ groups of parameters, $\boldsymbol{\gamma} = (\tilde{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\gamma}}_{d-2}^{\mathrm{T}})^{\mathrm{T}}$, where $\tilde{\boldsymbol{\gamma}}_j = (\gamma_{j,1}, \ldots, \gamma_{j,j+1})^{\mathrm{T}} \in \mathbb{R}^{j+1}$, for $j = 1, \ldots, d-2$. The first group $\tilde{\boldsymbol{\gamma}}_1$ consists of $\gamma_{1,1} = r_1 \cos\theta_1$ and $\gamma_{1,2} = r_1 \sin\theta_1$; for $j = 2, \ldots, d-2$, the j-th group $\tilde{\boldsymbol{\gamma}}_j$ consists of entries given by

$$\begin{cases} \gamma_{j,1} = r_j \cos\tilde{\phi}_{j,1}, \\ \gamma_{j,2} = r_j \sin\tilde{\phi}_{j,1} \cos\tilde{\phi}_{j,2}, \\ \vdots \\ \gamma_{j,j} = r_j \sin\tilde{\phi}_{j,1} \sin\tilde{\phi}_{j,2} \cdots \sin\tilde{\phi}_{j,j-1} \cos\theta_j, \\ \gamma_{j,j+1} = r_j \sin\tilde{\phi}_{j,1} \sin\tilde{\phi}_{j,2} \cdots \sin\tilde{\phi}_{j,j-1} \sin\theta_j. \end{cases} \tag{2.8}$$

This completes the derivations leading to $\boldsymbol{\gamma}$ for specifying $\mathbf{V}$ in $\mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$ after $\boldsymbol{\mu}$ is given.

Looking back, one can see that parameters introduced in Steps (i) and (ii) collected in $\widetilde{\boldsymbol{\Omega}}$ are transitional parameters that connect $\mathbf{V}$ subject to ESAG

constraints and the unrestricted $\boldsymbol{\gamma}$. Figure 2 gives a recap of the proposed parameterization and highlights the transitional nature of $\widetilde{\boldsymbol{\Omega}}$. Viewing (2.8) as a set of Cartesian coordinates $\tilde{\boldsymbol{\gamma}}_j$ in the $(j+1)$-dimensional Euclidean space when $d > 3$, for $j = 2, \ldots, d-2$, one can transform $\tilde{\boldsymbol{\gamma}}_j$ to the corresponding spherical coordinates in the j-dimensional spherical space given by

$$
\begin{aligned}
r_j &= \|\tilde{\boldsymbol{\gamma}}_j\|, \\
\theta_j &= \begin{cases} 0, & \text{if } \gamma_{j,j}^2 + \gamma_{j,j+1}^2 = 0, \\ \arccos \dfrac{\gamma_{j,j}}{\sqrt{\gamma_{j,j}^2 + \gamma_{j,j+1}^2}}, & \text{if } \gamma_{j,j+1} \geq 0 \text{ and } \gamma_{j,j}^2 + \gamma_{j,j+1}^2 \neq 0, \\ -\arccos \dfrac{\gamma_{j,j}}{\sqrt{\gamma_{j,j}^2 + \gamma_{j,j+1}^2}}, & \text{if } \gamma_{i,i+1} < 0, \end{cases} \\
\tilde{\phi}_{j,k} &= \begin{cases} 0, & \text{if } \sum_{\ell=k}^{j+1} \gamma_{j,\ell}^2 = 0, \text{ for } k = 1, \ldots, j-1, \\ \arccos \dfrac{\gamma_{j,k}}{\sqrt{\sum_{\ell=k}^{j+1} \gamma_{j,\ell}^2}}, & \text{otherwise, for } k = 1, \ldots, j-1, \end{cases}
\end{aligned} \tag{2.9}
$$

producing the j-th group of parameters in $\widetilde{\boldsymbol{\Omega}}$; and the first group contains $r_1 = \|\tilde{\boldsymbol{\gamma}}_1\|$ and $\theta_1 = \operatorname{atan2}(\gamma_{1,2}, \gamma_{1,1})$. In (2.9), we do not assume the denominators appearing in the transformations are always nonzero as we do in Table 1.

This completes the reparameterization of $\mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$ using $\boldsymbol{\Omega} = (\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ for any $d \geq 3$. Having the parameter space being $\mathbb{R}^p$ without any constraints greatly simplifies the implementation of maximum likelihood estimation for $\boldsymbol{\Omega}$.

3. Maximum likelihood estimation

Using the parameterization of ESAG developed in Section 2, one can easily derive the likelihood function of a sample from ESAG, following which one can maximize the logarithm of it with respect to $\boldsymbol{\Omega}$ over $\mathbb{R}^p$ to obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\Omega}$. Because all new parameters we bring in for the proposed parameterization are in $\boldsymbol{\gamma}$, we zoom in on $\boldsymbol{\gamma}$ next for its interpretations and implications on inferences for $\mathbf{V}$.

3.1. Interpretations of parameters

Because $r_j = \|\tilde{\boldsymbol{\gamma}}_j\|$, for $j = 1, \ldots, d-2$, and by (2.5), $\boldsymbol{\gamma} = \mathbf{0}$ implies $\lambda_j = 1$, for $j = 1, \ldots, d$, and thus $\mathbf{V} = \mathbf{I}_d$, leading to an isotropic hyperspherical distribution [15]. If $\mathbf{Y} \sim \mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{I}_d)$, then, for any orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}\boldsymbol{\mu} = \boldsymbol{\mu}$, we have $\mathbf{PY} \sim \mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{I}_d)$, i.e., $\mathbf{PY} = \mathbf{Y}$ in distribution, or, $\mathbf{PY} \overset{\mathcal{L}}{=} \mathbf{Y}$ in short. In addition, if $\tilde{\boldsymbol{\gamma}}_j = \mathbf{0}$, then $r_j = 0$, and thus $\lambda_{j+1} = (r_j + 1)\lambda_j = \lambda_j$, in which case we say that the distribution is isotropic in the subspace spanned by $\{\boldsymbol{\xi}_j, \boldsymbol{\xi}_{j+1}\}$, or partially isotropic. That is, given any orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}\boldsymbol{\mu} = \boldsymbol{\mu}$ and $\mathbf{P}\boldsymbol{\xi}_k = \boldsymbol{\xi}_k$, for $k \neq j, j+1$, we have

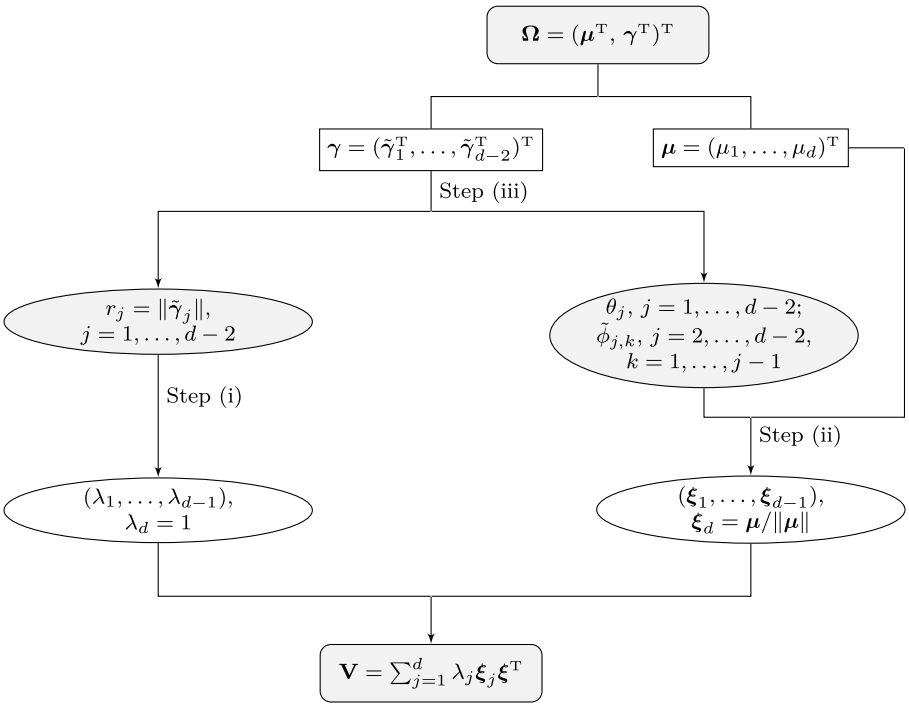

FIG 2. *A pictorial illustration of the proposed parameterization of* $ESAG_{d-1}(\boldsymbol{\mu}, \mathbf{V})$ *via the unconstrained* $\boldsymbol{\Omega}$. *The radial and orientation parameters in* $\widetilde{\boldsymbol{\Omega}}$ *are in the gray ellipses. The eigenvalues and eigenvectors of* $\mathbf{V}$ *are in the clear ellipses.*

$\mathbf{PY} \overset{\mathcal{L}}{=} \mathbf{Y}$. Practically speaking, this means that rotating data from an isotropic (a partially isotropic) ESAG via certain orthogonal matrix that rotates the mean direction to itself (and rotates certain eigenvectors of $\mathbf{V}$ to themselves) does not change the distribution of the data. From the modelling point of view, any level of isotropy of ESAG implies a reduced model. Hence, testing whether or not a data set can be modelled by a reduced, thus more parsimonious, ESAG amounts to testing hypotheses regarding parameters in $\boldsymbol{\gamma}$. For example, testing $\mathbf{V} = \mathbf{I}_d$ is equivalent to testing $\boldsymbol{\gamma} = \mathbf{0}$, which is the same for the unrestricted parameters (also called $\boldsymbol{\gamma}$) introduced in [20] when $d = 3$. Certainly, when $d = 3$, the concept of partially isotropic is irrelevant because $\boldsymbol{\gamma}$ only contains one group of parameters, $\tilde{\boldsymbol{\gamma}}_1$, in this case.

As one can see in (2.4), if $\boldsymbol{\xi}_j$ is an eigenvector of $\mathbf{V}$ corresponding to the eigenvalue λ_j, then so is $-\boldsymbol{\xi}_j$. This suggests that there exist $\boldsymbol{\gamma} \neq \boldsymbol{\gamma}'$ (corresponding to eigenvectors different by a sign) yet both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ lead to the same $\mathbf{V}$ given $\boldsymbol{\mu}$. When this happens, we say that $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are equivalent. We show in Appendix A that, if $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are equivalent, then $\|\tilde{\boldsymbol{\gamma}}_j\| = \|\tilde{\boldsymbol{\gamma}}'_j\|$, for $j = 1, \ldots, d-2$, which in turn suggests that the interpretations of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ relevant to isotropy of ESAG are the same. A theoretical implication of the

existence of equivalent $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ is that, although one cannot claim consistency of the MLE of $\boldsymbol{\gamma}$ (since the MLE may consistently estimate $\boldsymbol{\gamma}$ or $\boldsymbol{\gamma}'$), the consistency of the MLE of $\mathbf{V}$ is guaranteed by the invariance property of MLE [Theorem 7.2.10, 6]. A numerical implication of the existence of equivalent $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ is that maximum likelihood estimation of $\boldsymbol{\Omega}$ tends to be very forgiving in terms of the starting value for $\boldsymbol{\Omega}$, especially when the focal point of inference lies in $\boldsymbol{\mu}$ and $\mathbf{V}$. In other words, even though optimizing the log-likelihood under the new parameterization may lead to different members of an equivalent class due to different choices of starting values, these members all lead to the same estimation for $\boldsymbol{\mu}$ and $\mathbf{V}$. We provide empirical evidence of these implications in a simulation experiment in Section 3.2.

With our focal point of inference resting on $\boldsymbol{\mu}$ and $\mathbf{V}$ now fully specified by $\boldsymbol{\Omega}$, parameters in $\widetilde{\boldsymbol{\Omega}}$ are not of direct interest. Nevertheless, a noteworthy phenomenon similar to that discussed in [26, see remarks following Theorem 4] is that some orientation parameters in $\widetilde{\boldsymbol{\Omega}}$ are not identifiable when the truth of $\widetilde{\boldsymbol{\Omega}}$ falls in a subspace of the boundary of the parameter space. A detailed discussion of this issue is given in Appendix B, where we also provide empirical evidence indicating that finite-sample inferences for $\boldsymbol{\mu}$ and $\mathbf{V}$ are practically not affected whether or not some orientation parameters in $\widetilde{\boldsymbol{\Omega}}$ are identifiable. Asymptotic properties of MLEs of some model parameters however are expected to be affected (e.g., slower convergence or non-Gaussian limiting distribution), which we plan to address systematically in our follow-up research on ESAG regression models.

3.2. Empirical evidence

Using the proposed parameterization, we generate a random sample of size $n \in \{20, 50, 100\}$ from $\text{ESAG}_3(\boldsymbol{\mu}, \mathbf{V})$, where $\boldsymbol{\mu} = (2, -5, 3, 5)^{\mathrm{T}}$, and $\mathbf{V}$ is determined via $\boldsymbol{\mu}$ and $\boldsymbol{\gamma} = (\gamma_{1,1}, \gamma_{1,2}, \gamma_{2,1}, \gamma_{2,2}, \gamma_{2,3})^{\mathrm{T}} = (3, 5, -3, -4, 2)^{\mathrm{T}}$. We then maximize the log-likelihood function of this random sample to find the MLE of $\boldsymbol{\Omega}$, denoted by $\hat{\boldsymbol{\Omega}}$, using two different starting values of $\boldsymbol{\Omega}$: one coincides with the truth, the other is given by $\boldsymbol{\mu}_0 = \mathbf{1}_4$ and $\boldsymbol{\gamma}_0 = \mathbf{0}$. This produces two estimates of $\boldsymbol{\Omega}$. We repeat this experiment 1000 times. In all 1000 Monte Carlo replicates, we employ the Broyden-Fletcher-Goldfarb-Shanno algorithm [9] to find a maximizer of the log-likelihood function. In fact, we find that most commonly used optimization algorithms work well in maximizing the objective function despite the choice of starting values, partly thanks to the fact that transformations involved in the parameterization derivations in Section 2 are mostly smooth and simple enough.

Figure 3 presents graphical summaries of 1000 realizations of a subset of $\hat{\boldsymbol{\Omega}} = (\hat{\boldsymbol{\mu}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, $(\hat{\mu}_1, \hat{\gamma}_{1,1}, \hat{\gamma}_{2,1})$, corresponding to each choice of starting value at each level of the sample size n. In particular, for each parameter, a kernel density estimate based on 1000 realizations of its MLE is depicted in Figure 3. The top panels of Figure 3, which present results from using the truth of $\boldsymbol{\Omega}$ to start the optimization algorithm, provide empirical evidence suggesting that the

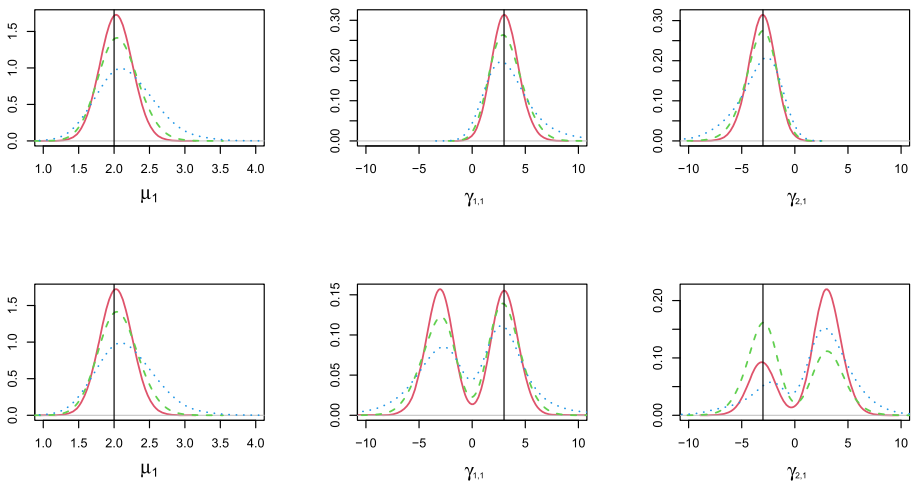

FIG 3. *Estimated distributions of estimators for three selected parameters in* $\boldsymbol{\Omega}$ *based on 1000 realizations of each parameter estimator when the true parameter values are used as the starting value (upper panels) and when* $\boldsymbol{\mu}_0$ *and* $\boldsymbol{\gamma}_0$ *not equal to the truth are used as starting values (lower panels) in search for a maximizer of the log-likelihood as* n *varies:* $n = 20$ *(blue dotted lines),* $n = 50$ *(green dashed lines),* $n = 100$ *(red solid lines). Vertical lines mark the true values of the corresponding parameters.*

usual asymptotic properties of an MLE, including consistency and asymptotic normality, are expected to hold for $\hat{\boldsymbol{\Omega}}$ when one uses a starting value in a neighborhood of the truth. The bottom panels of Figure 3, which show results from using a starting value that has little resemblance with the truth, indicate that $\hat{\boldsymbol{\mu}}$ still behaves like a regular MLE that is consistent and asymptotically normally distributed, but $\hat{\boldsymbol{\gamma}}$ appears to follow a bimodal distribution. The two modes of the distribution of $\hat{\boldsymbol{\gamma}}$ are expected to be the true value of $\boldsymbol{\gamma}$ and another value $\boldsymbol{\gamma}'$ that is equivalent to $\boldsymbol{\gamma}$.

Despite the potential bimodality of $\hat{\boldsymbol{\gamma}}$ when a less carefully chosen starting value of $\boldsymbol{\Omega}$ is used to find $\hat{\boldsymbol{\Omega}}$, the resultant estimate of $\mathbf{V}$, $\hat{\mathbf{V}}$, is similar, if not identical, to the estimate one obtains when using the truth as the starting value. Figure 4 shows boxplots of the Frobenius norm of $\mathbf{V} - \hat{\mathbf{V}}$ corresponding to 1000 realizations of $\hat{\mathbf{V}}$ resulting from each choice of the starting value at each level of n. From there one can see that $\hat{\mathbf{V}}$ is virtually unaffected by the choice of starting values.

Although the robustness of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$ to the choice of starting value is reassuring, one should not treat $\hat{\boldsymbol{\gamma}}$ as a conventional MLE due to its behavior observed in Figure 3. Consequently, the usual Fisher information matrix or the sandwich variance may not serve well for estimating the variance of $\hat{\boldsymbol{\Omega}}$. We thus recommend use of bootstrap for the uncertainty assessment of $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$, after mapping $\hat{\boldsymbol{\gamma}}$ to $\hat{\mathbf{V}}$ given $\hat{\boldsymbol{\mu}}$.

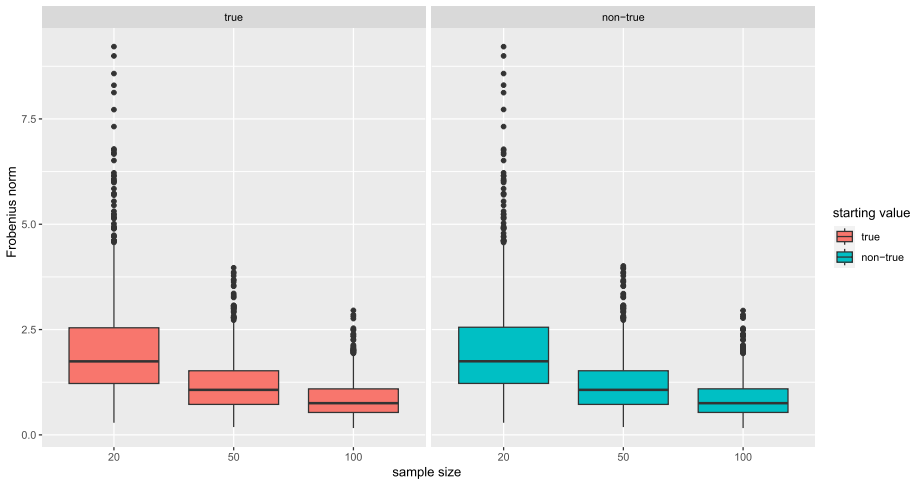

FIG 4. *Boxplots of the Frobenius norm of* $\mathbf{V} - \hat{\mathbf{V}}$ *as sample size* n *varies when the true parameter values are used as the starting value (in the left panel) and when* $\boldsymbol{\mu}_0$ *and* $\boldsymbol{\gamma}_0$ *not equal to the truth are used as starting values (in the right panel) in search for a maximizer of the log-likelihood.*

3.3. Composition estimation

When the original data are compositional data, a follow-up task after model parameters in $\text{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$ for the transformed data are estimated is the estimation of the mean composition of each component. This amounts to estimating $\text{E}(\mathbf{Y}^2)$, where $\mathbf{Y}^2$ is the element-wise quantity squared of the directional vector $\mathbf{Y}$. We show in Appendix C that, if $\mathbf{Y} \sim \text{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, then $E(\mathbf{Y}^2) = \boldsymbol{\Xi}^2\text{E}(\mathbf{K}^2)$, where $\boldsymbol{\Xi} = [\boldsymbol{\xi}_d \mid \boldsymbol{\xi}_{d-1} \mid \ldots \mid \boldsymbol{\xi}_1]$, $\mathbf{K} = \boldsymbol{\Xi}^{\text{T}}\mathbf{Y}$, and both $\boldsymbol{\Xi}^2$ and $\mathbf{K}^2$ refer to the element-wise quantity squared of the matrix/vector. This motivates an estimator of $E(\mathbf{Y}^2)$ based on a random sample $\{\mathbf{Y}_i\}_{i=1}^n$ given by $\hat{\boldsymbol{\Xi}}^2 \sum_{i=1}^n \hat{\mathbf{K}}_i^2/n$, where $\hat{\mathbf{K}}_i = \hat{\boldsymbol{\Xi}}^{\text{T}}\mathbf{Y}_i$, for $i = 1, \ldots, n$, and $\hat{\boldsymbol{\Xi}}$ is the MLE of $\boldsymbol{\Xi}$.

4. Model diagnostics

Even though the ESAG family accommodates certain anisotropic feature of a distribution and thus offers some flexibility in modelling, it remains fully parametric and thus is subject to model misspecification in a given application. In this section, we develop residual-based model diagnostics tools that data analysts can use to assess whether or not an ESAG distribution provides adequate fit for their directional data, either as a marginal distribution, or a conditional distribution of the directional response given covariates $\mathbf{W}$ as in a regression setting.

4.1. Residuals

Denote by $\{\mathbf{Y}_i\}_{i=1}^n$ the observed directional data of size n, where $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are independent with $\mathbf{Y}_i \sim \text{ESAG}_{d-1}(\boldsymbol{\mu}_i, \mathbf{V}_i)$, for $i = 1, \ldots, n$. The subscript i attached to $\boldsymbol{\mu}$ and $\mathbf{V}$ can be dropped if one aims to assess the goodness of fit (GOF) for the observed data using an ESAG as the marginal distribution. Otherwise the subscript implies covariate-dependent model parameters in ESAG as in a regression model for $\mathbf{Y}$.

In a non-regression or regression setting, after one obtains the MLE of all unknown parameters in the model, one has the MLEs $\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{V}}_i$, following which a prediction can be made by $\hat{\mathbf{Y}}_i = \hat{\boldsymbol{\mu}}_i/\|\hat{\boldsymbol{\mu}}_i\|$, for $i = 1, \ldots, n$. Similar to a directional residual defined in [12], we define residuals as

$$\hat{\mathbf{r}}_i = \left(\mathbf{I}_d - \hat{\mathbf{Y}}_i\hat{\mathbf{Y}}_i^{\mathrm{T}}\right)\mathbf{Y}_i, \text{ for } i = 1, \ldots, n. \tag{4.1}$$

In (4.1), $\hat{\mathbf{Y}}_i\hat{\mathbf{Y}}_i^{\mathrm{T}}$ can be viewed as the projection onto the space spanned by $\hat{\boldsymbol{\mu}}_i$, and thus $\mathbf{I}_d - \hat{\mathbf{Y}}_i\hat{\mathbf{Y}}_i^{\mathrm{T}}$ is the projection onto the space orthogonal to the space spanned by $\hat{\boldsymbol{\mu}}_i$. Equivalently, by the orthogonality of eigenvectors of $\hat{\mathbf{V}}_i$, $\mathbf{I}_d - \hat{\mathbf{Y}}_i\hat{\mathbf{Y}}_i^{\mathrm{T}}$ is the projection onto the space spanned by the $d-1$ eigenvectors of $\hat{\mathbf{V}}_i$ that are orthogonal to $\hat{\boldsymbol{\mu}}_i$, denote by $\{\hat{\boldsymbol{\xi}}_{i,j}\}_{j=1}^{d-1}$. Hence (4.1) can be re-expressed as $\hat{\mathbf{r}}_i = \hat{\mathbf{P}}_{-d}\hat{\mathbf{P}}_{-d}^{\mathrm{T}}\mathbf{Y}_i$, where $\hat{\mathbf{P}}_{-d} = [\hat{\boldsymbol{\xi}}_{i,1} \mid \ldots \mid \hat{\boldsymbol{\xi}}_{i,d-1}]$, that is, $\hat{\mathbf{P}}_{-d}$ is the $d \times (d-1)$ matrix with the j-th column being $\hat{\boldsymbol{\xi}}_{i,j}$, for $j = 1, \ldots, d-1$. The potential dependence $\hat{\mathbf{P}}_{-d}$ on covariates via the subscript i is suppressed for simplicity.

For model diagnostic purposes, we use the following quadratic form of residuals,

$$\hat{Q}_i = \hat{\mathbf{r}}_i^{\mathrm{T}}\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{r}}_i, \text{ for } i = 1, \ldots, n. \tag{4.2}$$

With consistent estimation of $\boldsymbol{\mu}_i$, along with consistent estimation of $\boldsymbol{\gamma}_i$ or an equivalent $\boldsymbol{\gamma}_i'$, we have $\hat{\mathbf{r}}_i = \hat{\mathbf{P}}_{-d}\hat{\mathbf{P}}_{-d}^{\mathrm{T}}\mathbf{Y}_i$ converge to $\mathbf{r}_i = \mathbf{P}_{-d}\mathbf{P}_{-d}^{\mathrm{T}}\mathbf{Y}_i$ in distribution, where $\mathbf{P}_{-d}$ results from excluding the d-th column of the $d \times d$ matrix $\mathbf{P} = [\boldsymbol{\xi}_1 \mid \ldots \mid \boldsymbol{\xi}_{d-1} \mid \boldsymbol{\xi}_d]$, and $\mathbf{P}_{-d}\mathbf{P}_{-d}^{\mathrm{T}} = \mathbf{I}_d - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathrm{T}}/\|\boldsymbol{\mu}_i\|^2$. Additionally, $\hat{\mathbf{V}}_i$ converges to $\mathbf{V}_i$ in probability as $n \to \infty$. Thus, (4.2) converges to $Q_i = \mathbf{r}_i^{\mathrm{T}}\mathbf{V}_i^{-1}\mathbf{r}_i$ in distribution as $n \to \infty$. In what follows, we investigate the distribution of Q_i to gain insight on the asymptotic distribution of (4.2). The subscript i as the data point index is suppressed in this investigation.

For $\mathbf{Y} \sim \text{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, the random variable can be expressed as $\mathbf{Y} = \mathbf{X}/\|\mathbf{X}\| = (\mathbf{V}^{1/2}\mathbf{Z} + \boldsymbol{\mu})/\|\mathbf{X}\|$, where $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$. Hence, $\mathbf{r} = \mathbf{P}_{-d}\mathbf{P}_{-d}^{\mathrm{T}}\mathbf{Y} = \mathbf{P}_{-d}\mathbf{P}_{-d}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{Z}/\|\mathbf{X}\|$, following which we show in Appendix D that

$$Q = \mathbf{r}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{r} = \frac{\|\mathbf{U}_{-d}\|^2}{\|\mathbf{X}\|^2}, \tag{4.3}$$

where $\mathbf{U}_{-d}$ results from replacing the d-th entry of $\mathbf{U} = \mathbf{P}^{\mathrm{T}}\mathbf{Z}$ with zero. Since $\mathbf{P}$ is an orthogonal matrix, $\mathbf{U} = \mathbf{P}^{\mathrm{T}}\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$, and thus $\|\mathbf{U}_{-d}\|^2 \sim \chi^2_{d-1}$.

Now we see that Q relates to the quotient of norms of Gaussian vectors, the distribution of which was studied in [17], following which one can derive the distribution of Q analytically. One then can see that Q is not a pivotal quantity and its distribution is not of a form familiar or easy enough for direct use for model diagnosis. We next construct a transformation of Q aiming at attaining an approximate pivotal quantity for the purpose of model diagnostics.

4.2. Graphical model diagnostic

[20] showed that, if $\mathbf{Y} = (Y_1, \ldots, Y_d)^{\mathrm{T}} \sim \mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, then $\|\boldsymbol{\mu}\|(Y_1, \ldots, Y_{d-1})^{\mathrm{T}}$ converges in distribution to $N_{d-1}(\mathbf{0}, \sum_{j=1}^{d-1} \lambda_j^{-1} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^{\mathrm{T}})$ as $\|\boldsymbol{\mu}\| \to \infty$. [21] defined model-based residuals motivated by this result, and proposed to inspect a scatter plot of such residuals to detect model inadequacy when fitting a regression model that assumes ESAG errors. Following this finding, one also has that $T_0 = \|\boldsymbol{\mu}\|^2 Q = (\|\boldsymbol{\mu}\|^2/\|\mathbf{X}\|^2)\|\mathbf{U}_{-d}\|^2$ converges in distribution to χ^2_{d-1} for ESAG, and thus is a pivot in limit as $\|\boldsymbol{\mu}\| \to \infty$ (instead of $n \to \infty$). One may thus assess adequacy of a posited ESAG model for a data set by checking if $\{\hat{T}_{0,i}\}_{i=1}^n = \{\|\hat{\boldsymbol{\mu}}_i\|^2 \hat{Q}_i\}_{i=1}^n$ approximately come from χ^2_{d-1}. As seen in Figure 1, a larger $\|\boldsymbol{\mu}\|$ implies that the distribution has a higher concentration and thus less variability in data. This diagnostic strategy based on T_0 is thus intuitively well motivated since, with $\|\boldsymbol{\mu}\|$ large, $\|\boldsymbol{\mu}\|^2/\|\mathbf{X}\|^2$ is expected to be close to one, making T_0 close to $\|\mathbf{U}_{-d}\|^2 \sim \chi^2_{d-1}$. However, empirical evidence from our extensive simulation study suggest that a practically unreasonably large $\|\boldsymbol{\mu}\|$ is needed to make χ^2_{d-1} a reasonably good approximation of the distribution of T_0. Consequently, this strategy based on T_0 is of little practical value since data observed in most applications can rarely have low enough variability to make this approximation satisfactory.

Motivated by the fact that $E(\|\mathbf{X}\|^2) = \|\boldsymbol{\mu}\|^2 + \sum_{j=1}^d \lambda_j$ [Theorem 5.2.1, 24], we propose the following random quantity for diagnostics purposes,

$$T_1 = \left(\|\boldsymbol{\mu}\|^2 + \sum_{j=1}^{d} \lambda_j \right) Q, \tag{4.4}$$

which follows χ^2_{d-1} approximately when $\|\boldsymbol{\mu}\|$ is large, with the approximation improves much faster than that for T_0 as $\|\boldsymbol{\mu}\|$ increases, and thus is more like a pivot than T_0 is. Figure 5 presents kernel density estimates of the distributions of T_0 and T_1 based on random samples of these random quantities, each of size 500, generated based on Monte Carlo replicates from $\mathrm{ESAG}_3(\boldsymbol{\mu}, \mathbf{V})$. More specifically, we set $\|\boldsymbol{\mu}\| = 4.24$, which is not large enough to make the χ^2-approximation for T_0 satisfactory, and $\sum_{j=1}^d \lambda_j = 11.1$. As one can see in this figure, the variability of T_0 is way too low to make χ^2_{d-1} approximate its distribution well, and T_1 greatly improves over T_0 in its proximity to χ^2_{d-1}. In general, T_1 only requires a moderate $\|\boldsymbol{\mu}\|$ to make the χ^2-approximation practically useful.

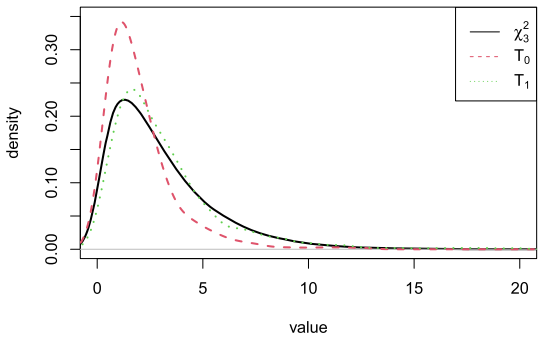

FIG 5. *Kernel density estimates of T_0 (dashed line) and T_1 (dotted line) comparing with the density of χ^2_3 (solid line).*

Following maximum likelihood estimation of all unknown parameters, one can exploit an empirical version of T_1, $\{\hat{T}_{1,i}\}_{i=1}^n$, where $\hat{T}_{1,i} = (\|\hat{\boldsymbol{\mu}}_i\|^2 + \sum_{j=1}^d \hat{\lambda}_{i,j})\hat{Q}_i$, for $i = 1, \ldots, n$, and check if $\{\hat{T}_{1,i}\}_{i=1}^n$ can be reasonably well modeled by χ^2_{d-1}. It can be a graphical check via a quantile-quantile (QQ) plot, for example, to see if there exists any clear signal of this sample deviating from χ^2_{d-1}. Such graphical check is easy to implement following parameter estimation, and can provide visual warning signs when ESAG is a grossly inadequate model for the observed data $\{\mathbf{Y}_i\}_{i=1}^n$. Certainly, in a given application, the quality of χ^2-approximation for T_1 is unknown with its true distribution yet to be estimated. We next propose a bootstrap procedure to facilitate a quantitative test for model misspecification, which leads to another graphical diagnostic tool as a byproduct that does not rely on a χ^2-approximation for T_1.

4.3. Goodness of fit test

Consider testing the null hypothesis that $\mathbf{Y}$ follows an ESAG. Although T_1 defined in (4.4) approximately follows χ^2_{d-1} under the null hypothesis, a testing procedure based on T_1 that does not acknowledge its exact null distribution can lead to misleading conclusion, e.g., an inflated Type I error for the test. Instead of estimating the exact null distribution of T_1, we use a random sample of T_1 induced from an ESAG as a reference sample, and quantify the dissimilarity between this reference sample and the observed empirical version of T_1, $\{\hat{T}_{1,i}\}_{i=1}^n$. One may use a nonparametric test for testing if two data sets come from the same distribution, such as the Kolmogorov–Smirnov (KS) test [7] and the Cramér-von Mises test [3], to compare $\{\hat{T}_{1,i}\}_{i=1}^n$ and the reference sample induced from an ESAG. We employ the KS test in all presented simulation study in this article. A smaller p-value from the test indicates a larger distance between the underlying distribution of $\{\hat{T}_{1,i}\}_{i=1}^n$ and that of the reference sample, with the latter approximately representing what one expects for T_1 under the null hypothesis. Here, the ultimate test statistic for testing the null hypothesis

Algorithm 1 Goodness-of-Fit Test Procedure

1: **procedure** COMPARE OBSERVED EMPIRICAL VERSION OF T_1 WITH A REFERENCE SAMPLE
2: Given data $\{\mathbf{Y}_i\}_{i=1}^n$ for a non-regression setting or $\{(\mathbf{Y}_i, \mathbf{W}_i)\}_{i=1}^n$ for a regression setting, find the MLE $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\gamma}}_i$, for $i = 1, \ldots, n$, assuming an ESAG model for $\mathbf{Y}_i$ or $\mathbf{Y}_i$ conditioning on $\mathbf{W}_i$.
3: Compute $\hat{\mathbf{V}}_i$ and $\{\hat{\lambda}_{i,j}\}_{j=1}^{d-1}$ based on $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\gamma}}_i$, for $i = 1, \ldots, n$.
4: Compute $\hat{T}_{1,i} = (\|\hat{\boldsymbol{\mu}}_i\|^2 + \sum_{j=1}^{d-1} \hat{\lambda}_{i,j})\hat{Q}_i$, for $i = 1, \ldots, n$.
5: Generate $\{\tilde{\mathbf{Y}}_i\}_{i=1}^n$, where $\tilde{\mathbf{Y}}_i \sim \text{ESAG}(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i)$, for $i = 1, \ldots, n$.
6: Compute $\tilde{T}_{1,i} = (\|\hat{\boldsymbol{\mu}}_i\|^2 + \sum_{j=1}^{d-1} \hat{\lambda}_{i,j})\tilde{Q}_i$, where $\tilde{Q}_i = \tilde{r}_i^{\mathrm{T}} \hat{\mathbf{V}}_i^{-1} \tilde{r}_i$ and $\tilde{r}_i = \hat{\mathbf{P}}_{-d} \hat{\mathbf{P}}_{-d}^{\mathrm{T}} \tilde{\mathbf{Y}}_i$, for $i = 1, \ldots, n$.
7: Use the KS test to test if $\{\hat{T}_{1,i}\}_{i=1}^n$ and $\{\tilde{T}_{1,i}\}_{i=1}^n$ arise from the same distribution. Denote by KS_p the resultant p-value of the KS test.
8: **end procedure**
9: **procedure** BOOTSTRAP PROCEDURE TO ESTIMATE THE NULL DISTRIBUTION OF KS_p
10: Set B = number of bootstraps
11: Initiate $s = 0$
12: **for** b in $1, \ldots, B$ **do**
13: Generate the b-th bootstrap sample $\{\mathbf{Y}_i^{(b)}\}_{i=1}^n$, where $\mathbf{Y}_i^{(b)} \sim \text{ESAG}(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i)$ for $i = 1, \ldots, n$.
14: Repeat steps 2–7 using data $\{\mathbf{Y}_i^{(b)}\}_{i=1}^n$ for a non-regression setting or $\{(\mathbf{Y}_i^{(b)}, \mathbf{W}_i)\}_{i=1}^n$ for a regression setting. Denote the p-value of the KS test as $\text{KS}_p^{(b)}$.
15: **if** $\text{KS}_p^{(b)} < \text{KS}_p$ **then** $s = s + 1$
16: **end for**
17: Define an estimated p-value for this GOF test as s/B.
18: **end procedure**

is a p-value from the KS test. Denote this test statistic as KS_p. Alternatively, one may use the Kolmogorov-Smirnov statistic (as the largest distance between two estimated distribution functions) as a test statistic. We adopt KS_p instead of the distance statistic mainly due to the bounded support of the former. Even when data are from an ESAG, it is analytically unclear what KS_p should be because the ESAG from which the reference sample is induced is not exactly the true ESAG (as to be seen next). We thus use parametric bootstrap to estimate the null distribution of KS_p to obtain an approximate p-value to compare with a preset nominal level, such as 0.05, according to which we conclude to reject or fail to reject the null at the chosen nominal level. Algorithm 1 above presents a detailed algorithm for this hypothesis testing procedure.

Several remarks are in order for this algorithm. First, in Step 5, $\text{ESAG}(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i)$, from which we induce a data point $\tilde{T}_{1,i}$ in the reference sample $\{\tilde{T}_{1,i}\}_{i=1}^n$, can be viewed as the member of the ESAG family that is closest to the distribution that characterizes the true data generating process producing $\mathbf{Y}_i$, where the closeness between two distributions is quantified by the Kullback-Leibler divergence [33]. Hence, $\tilde{\mathbf{Y}}_i$ generated from $\text{ESAG}(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i)$ at this step is expected to resemble $\mathbf{Y}_i$ if the null hypothesis is true, with $\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{V}}_i$ consistently estimating $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$, respectively. Second, in Step 6, $\tilde{T}_{1,i}$ is constructed in a way that closely mimics T_1 instead of $\hat{T}_{1,i}$. In particular, just like T_1 where all population parameters are used in its construction, such as $\boldsymbol{\mu}$, $\{\lambda_j\}_{j=1}^d$, as well as $\mathbf{V}$ and $\mathbf{P}_{-d}$ that Q depends on, computing $\tilde{T}_{1,i}$ (upon completing Steps

2–5) requires no parameter estimation although it depends on $\hat{\boldsymbol{\mu}}_i$, $\{\hat{\lambda}_{i,j}\}_{j=1}^d$, $\hat{\mathbf{V}}_i$ and $\hat{\mathbf{P}}_{-d}$, which are viewed as population parameters associated with $\tilde{\mathbf{Y}}_i$. One may certainly construct in Step 6 a random quantity closely mimicking $\hat{T}_{1,i}$ instead, but that would involve another round of parameters estimation based on $\{\tilde{\mathbf{Y}}_i\}_{i=1}^n$ and thus is computationally unattractive. Third, we acknowledge that, even under the null hypothesis, $\mathrm{ESAG}(\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{V}}_i)$ is not the true distribution of $\mathbf{Y}_i$, with MLEs in place of the true model parameters. Hence, even when the null hypothesis is true, $\{\hat{T}_{1,i}\}_{i=1}^n$ do not come from the same distribution as that of the reference sample $\{\tilde{T}_{1,i}\}_{i=1}^n$, but the two distributions are expected to be closer than when the null hypothesis is severely violated. The bootstrap procedure is designed to estimate the null distribution of the distance between these two distributions that is quantified by KS_p, with a smaller value of KS_p indicating a larger distance and thus stronger evidence against the null. As to be seen in the upcoming simulation study, this bootstrap procedure is capable of approximating the null distribution of KS_p well enough to yield an empirical size of the test matching closely with any given nominal level.

In the absence of model misspecification, the distribution of $\{\tilde{T}_{1,i}\}_{i=1}^n$ approximates the distribution of T_1, with the accuracy of the approximation depends less on $\|\boldsymbol{\mu}\|$ than the χ^2-approximation does. Therefore, a more reliable graphical diagnostic device than the aforementioned QQ plot using χ^2_{d-1} as a reference distribution is a QQ plot based on $\{\hat{T}_{1,i}\}_{i=1}^n$ and $\{\tilde{T}_{1,i}\}_{i=1}^n$, as we demonstrate in the upcoming empirical study.

5. Simulation study

5.1. Design of simulation

To demonstrate operating characteristics of the diagnostics methods proposed in Section 4, we apply them to data $\{\mathbf{Y}_i\}_{i=1}^n$ generated according to four data generating processes specified as follows:

(M1) An ESAG model, $\mathrm{ESAG}_3(\boldsymbol{\mu}, \mathbf{V})$, with $\boldsymbol{\mu} = (2, -2, 3, -3)^{\mathrm{T}}$ and $\mathbf{V}$ defined via $\boldsymbol{\mu}$ and $\boldsymbol{\gamma} = (2, 3, 5, 8, 2)^{\mathrm{T}}$.

(M2) A mixture of ESAG and angular Cauchy, with a mixing proportion of $1-\alpha$ on $\mathrm{ESAG}_3(\boldsymbol{\mu}, \mathbf{V})$ specified in (M1), where a random vector from an angular Cauchy is generated by normalizing a random vector from a multivariate Cauchy with mean $\boldsymbol{\mu}$. This creates a scenario where $(1-\alpha)\times 100\%$ of the data arise from ESAG but the rest of the data deviate from ESAG, where $\alpha \in \{0.05, 0.1, 0.2\}$.

(M3) An angular Gaussian distribution, $\mathrm{AG}(\boldsymbol{\mu}, \tilde{\mathbf{V}})$, where $\det(\tilde{\mathbf{V}}) = \alpha \neq 1$, which creates a scenario where the constraint in (2.2) is violated. More specifically, when formulating (M1), one has the eigenvalues $\{\lambda_j\}_{j=1}^{d-1}$ and the corresponding eigenvectors $\{\boldsymbol{\xi}_j\}_{j=1}^{d-1}$ of $\mathbf{V}$, besides $\lambda_d = 1$ and $\boldsymbol{\xi}_d = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. Using these quantities from (M1), we define $\tilde{\mathbf{V}} = \sum_{j=1}^d \tilde{\lambda}_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^{\mathrm{T}}$, where $\tilde{\lambda}_j = \alpha^{1/(d-1)}\lambda_j$, for $j = 1, \ldots, d-1$, and $\tilde{\lambda}_d = 1$, with $\alpha \in$

$\{0.05, 0.1, 5, 10\}$. Because $\tilde{\mathbf{V}}\boldsymbol{\mu} = \boldsymbol{\mu}$, the constraint in (2.1) for ESAG is satisfied for this angular Gaussian distribution.

(M4) Similar to (M3) but $\tilde{\lambda}_j = \alpha^{-1/(d-1)}\lambda_j$, for $j = 1, \ldots, d-1$, and $\tilde{\lambda}_d = \alpha \in \{0.1, 0.5, 2.5, 5\}$. This leads to $\tilde{\mathbf{V}}\boldsymbol{\mu} = \alpha\boldsymbol{\mu}$ and thus violates constraint (2.1). Because now $\det(\tilde{\mathbf{V}}) = 1$, the constraint in (2.2) for ESAG is satisfied for this angular Gaussian distribution.

We generate random samples of size $n \in \{250, 500, 1000\}$ following each data generating process. The proportions of data sets across 300 Monte Carlo replicates for which the GOF test rejects the null hypothesis at various significance levels are recorded for each simulation setting. This rejection rate estimates the size of the test under (M1), and sheds light on how sensitive the proposed diagnostic methods are to various forms and severity of deviations from ESAG exhibited in (M2)–(M4). We set $B = 200$ in the bootstrap algorithm.

5.2. Simulation results

Under (M1), Figure 6 shows the rejection rate versus the nominal level when the null hypothesis stating that $\mathbf{Y} \sim$ ESAG is true. This figure suggests that the null distribution of the test statistic KS_p is approximated well enough over a wide range of nominal levels based on merely $B = 200$ bootstrap samples, especially at the lower tail so that the size of the test is close to a low nominal level such as 0.05.

Table 2 presents rejection rates of the GOF test at nominal level 0.05 under the remaining three data generating processes (M2)–(M4). Under (M2), when $\alpha \times 100\%$ of the observed data are not from ESAG, the power of the test steadily increases as α increases. A larger sample size also boosts the power of detecting violation of the null. Under (M3), when data are from $\mathrm{AG}(\boldsymbol{\mu}, \tilde{\mathbf{V}})$ that does not satisfy constraint (2.2) due to $\det(\tilde{\mathbf{V}}) = \alpha(\neq 1)$, one can see from Table 2 that, depending on the severity of the violation of (2.2) that is controlled by the deviation of α from 1, the proposed test has a moderate power to detect this particular violation of ESAG, with a higher power at a larger sample size. Under (M4), when data are from $\mathrm{AG}(\boldsymbol{\mu}, \tilde{\mathbf{V}})$ with constraint (2.1) violated due to $\tilde{\mathbf{V}}\boldsymbol{\mu} = \alpha\boldsymbol{\mu}$, one can see from Table 2 that, as α deviates from 1 from either direction, the proposed test possesses moderate to high power to detect violation of the null hypothesis, with the power increasing quickly as n grows larger.

Besides the quantitative GOF test that performs satisfactorily according to the above empirical evidence, one can also inspect the QQ plot based on $\{\hat{T}_{1,i}\}_{i=1}^n$ and the bootstrap sample $\{\tilde{T}_{1,i}\}_{i=1}^n$ to graphically check ESAG assumptions. Figure 7 shows a collection of such plots based on a randomly chosen Monte Carlo replicate from each of the four considered data generating processes. As evidenced in Figure 7, violation of the ESAG assumptions as designed in (M2)–(M4) causes a QQ plot deviating from a straight-line pattern, a pattern more or less observed in the absence of model misspecification as in (M1). The similarity between the three QQ plots under (M2)–(M4) suggests that $\{\hat{T}_{1,i}\}_{i=1}^n$ are not informative in distinguishing different forms of ESAG

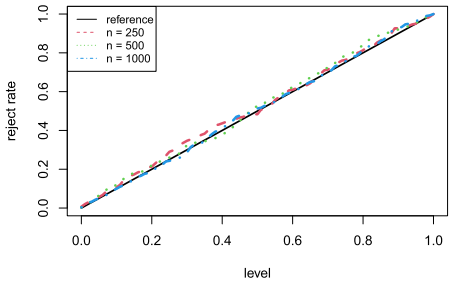

FIG 6. *Rejection rates of the GOF test versus nominal levels under (M1) when* $n = 250$ *(dashed line), 500 (dotted line), and 1000 (dash-dotted line). The solid line is the* 45° *reference line.*

TABLE 2
Rejection rates of the GOF test under (M2)–(M4) at nominal level 0.05

n	(M2)			(M3)				(M4)			
$\{\alpha\}$	0.05	0.1	0.2	0.05	0.1	5	10	0.1	0.5	2.5	5
250	0.10	0.27	0.65	0.27	0.17	0.14	0.17	0.47	0.16	0.75	1.00
500	0.17	0.42	0.89	0.38	0.30	0.22	0.33	0.73	0.26	0.98	1.00
1000	0.26	0.69	0.99	0.60	0.46	0.30	0.52	0.96	0.42	1.00	1.00

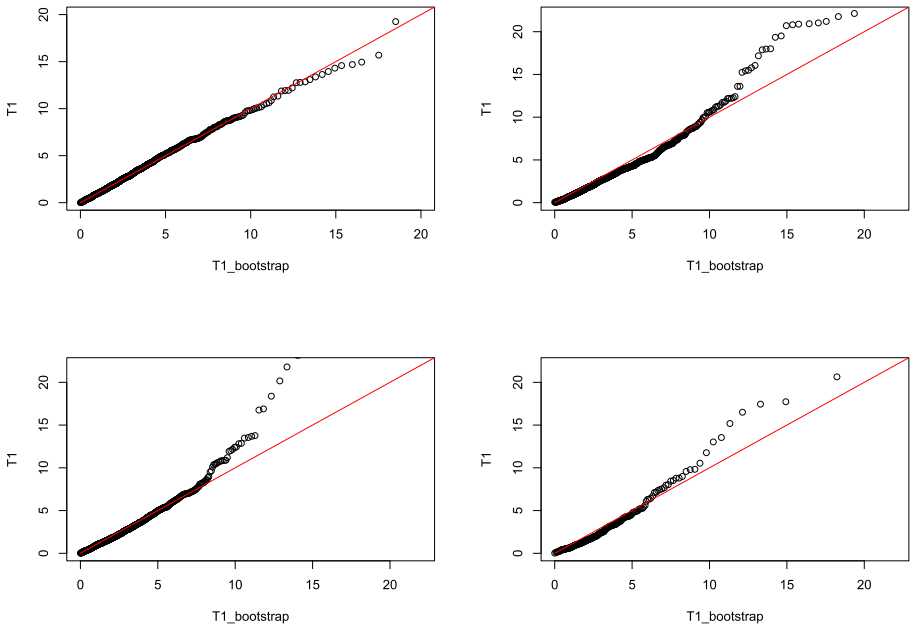

FIG 7. *QQ plots based on* $\{\hat{T}_{1,i}\}_{i=1}^n$ *and the bootstrap sample* $\{\tilde{T}_{1,i}\}_{i=1}^n$ *under (M1) (top-left panel), (M2) with* $\alpha = 0.2$ *(top-right panel), (M3) with* $\alpha = 0.05$ *(bottom-left panel), and (M4) with* $\alpha = 2.5$ *(bottom-right panel), respectively. Solid lines are* 45° *reference lines.*

violations. Regardless, such QQ plot provides a graphical check on the overall goodness of fit that is convenient to create because the full B-round bootstrap procedure in the above algorithm is not needed to make the plot.

6. Application to hydrochemical data

In this section, we analyze the hydrochemical data containing 14 molarities measured monthly at different stations along the Llobregat River and its tributaries in northeastern Spain between the summer of 1997 and the spring of 1999 [19]. The complete data are available in the `R` package, `compositions` [31]. For illustration purposes, we focus on the compositional data recording relative abundance of two major ions, K^+ and Na^+, and two minor ions, Ca^{2+} and Mg^{2+}. Taking the square-root transformation of the compostional data gives directional data with $d = 4$. The four considered ions are mostly from potash mine tailing, which is one of the major sources of anthropogenic pollution in the Llobregat Basin [29].

We first assume that the transformed composition of $(K^+, Na^+, Ca^{2+}, Mg^{2+})$ in tributaries of Anoia, one of the two main tributaries of the Llobregat River, follows an ESAG distribution. Fitting 67 records collected from stations placed along tributaries of Anoia to $\mathrm{ESAG}_3(\boldsymbol{\mu}, \mathbf{V})$, we obtain estimates of $\boldsymbol{\mu}$ and $\mathbf{V}$ as

$$\hat{\boldsymbol{\mu}}_{\mathrm{A}} = \begin{bmatrix} 1.99 \\ 5.74 \\ 7.95 \\ 4.59 \end{bmatrix}, \quad \hat{\mathbf{V}}_{\mathrm{A}} = \begin{bmatrix} 0.93 & 1.15 & -0.76 & -0.09 \\ 1.15 & 2.77 & -1.41 & -0.27 \\ -0.76 & -1.41 & 1.99 & 0.38 \\ -0.09 & -0.27 & 0.38 & 0.73 \end{bmatrix}.$$

The GOF test yields an estimated p-value of 0.66, suggesting that the estimated ESAG distribution may provide an adequate fit for the data. The QQ plot in Figure 8 (see the left panel) may indicate some disagreement in the upper tail when it comes to the distribution of $\hat{T}_1$ and its bootstrap counterpart induced from an ESAG distribution, but otherwise mostly resemble each other in distribution. Using the estimated model parameters and applying the method in Section 3.3, we obtain an estimate of the mean composition of $(K^+, Na^+, Ca^{2+}, Mg^{2+})$ to be (0.04, 0.28, 0.51, 0.17).

We repeat the above exercise for another compositional data of size 43 collected from stations placed along tributaries of the lower Llobregat course, and obtain estimates for $\boldsymbol{\mu}$ and $\mathbf{V}$ given by

$$\hat{\boldsymbol{\mu}}_{\mathrm{L}} = \begin{bmatrix} 3.27 \\ 8.56 \\ 9.01 \\ 5.78 \end{bmatrix}, \quad \hat{\mathbf{V}}_{\mathrm{L}} = \begin{bmatrix} 0.63 & 1.50 & -0.71 & -0.90 \\ 1.50 & 5.36 & -2.66 & -3.17 \\ -0.71 & -2.66 & 2.43 & 2.10 \\ -0.90 & -3.17 & 2.10 & 2.91 \end{bmatrix}.$$

The estimated p-value from the GOF test is 0.55 in this case. This, along with the QQ plot in Figure 8 (see the middle panel), also implies that the inferred ESAG distribution fits the data reasonably well. The estimated mean composition of

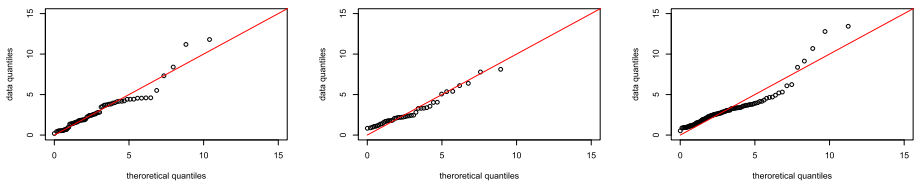

FIG 8. *QQ plots from the GOF test applied to compositional data from tributaries of Anoia (left panel), those from tributaries of the lower Llobregat course (middle panel), and the data that combine the previous two data sets (right panel).*

(K^+, Na^+, Ca^{2+}, Mg^{2+}) is $(0.06, 0.37, 0.40, 0.17)$, which shares some similarity with the estimated mean composition associated with Anoia tributaries in terms of Mg^{2+}, and also in that Ca^{2+} and Na^+ are the two dominating components among the four, and K^+ is the minority.

The two estimates for $\mathbf{V}$, $\hat{\mathbf{V}}_{\text{A}}$ and $\hat{\mathbf{V}}_{\text{L}}$, also share some implications in common: the two major ions, K^+ and Na^+, are positively correlated, so are the two minor ions, Ca^{2+} and Mg^{2+}; but a major ion is negatively correlated with a minor ion in composition. Diagonal entries of $\hat{\mathbf{V}}_{\text{A}}$ and $\hat{\mathbf{V}}_{\text{L}}$ should not be interpreted or compared here in the same way as if data were not directional because the variability of ESAG($\boldsymbol{\mu}$, $\mathbf{V}$) depends on both $\boldsymbol{\mu}$ and $\mathbf{V}$. For the compositional vector as a whole, with $\|\hat{\boldsymbol{\mu}}_{\text{A}}\| \approx 11.00 < \|\hat{\boldsymbol{\mu}}_{\text{L}}\| \approx 14.10$, we have data evidence suggesting that the transformed compositional data (as directional data) from Anoia tributaries are less concentrated around its mean direction, and thus more variable, than those from tributaries of the lower Llobregat course. When zooming in on one component at a time in the compositional vector, one can compare variability between two ESAG distributions based on $\mathbf{V}/\|\boldsymbol{\mu}\|^2$. For instance, even though $\hat{\mathbf{V}}_{\text{A}}[3,3] = 1.99 < \hat{\mathbf{V}}_{\text{L}}[3,3] = 2.43$, we would not jump to the conclusion that the composition of Ca^{2+} is less variable in Anoia tributaries than that in the other set of locations. Instead, because $\hat{\mathbf{V}}_{\text{A}}[3,3]/\|\hat{\boldsymbol{\mu}}_{\text{A}}^2\| = 0.18 > \hat{\mathbf{V}}_{\text{L}}[3,3]/\|\hat{\boldsymbol{\mu}}_{\text{L}}^2\| = 0.17$, we conclude that the composition of Ca^{2+} is similar in variability between the two sets of locations, but tributaries of Anoia may be subject to slightly higher variability in this regard. This conclusion is also consistent with the comparison of the sample standard deviation of the composition of Ca^{2+} between the two data sets.

Moreover, estimates for the other set of parameters of ESAG arising in the new parameterization, $\boldsymbol{\gamma}$, also provide statistically interesting insights on the underlying distributions. Denote by $\hat{\boldsymbol{\gamma}}_{\text{A}}$ the estimate based on data from Anoia tributaries, and by $\hat{\boldsymbol{\gamma}}_{\text{L}}$ the estimate based on data from tributaries of the lower Llobregat course. We find that $\|\hat{\boldsymbol{\gamma}}_{\text{A}}\| = 6.24 < \|\hat{\boldsymbol{\gamma}}_{\text{L}}\| = 17.03$, indicating that neither of the two ESAG distributions is isotropic, with the second ESAG deviating from isotropy further. To check partial isotropy, we look into the estimated eigenvalues associated with $\hat{\mathbf{V}}_{\text{A}}$ and $\hat{\mathbf{V}}_{\text{L}}$. With one eigenvalue fixed at 1, the three estimated eigenvalues associated with $\hat{\mathbf{V}}_{\text{A}}$ are 0.37 (0.05), 0.62 (0.10), and 4.44 (0.64), with the estimated standard errors in parentheses obtained based

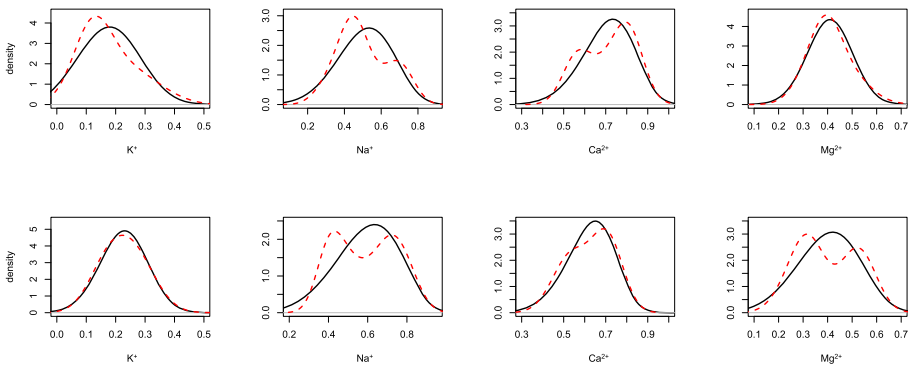

FIG 9. *Kernel density estimates associated with each of the four considered components, K^+, Na^+, Ca^{2+}, and Mg^{2+}, based on the square-root transformed compositional data (red dashed lines) from tributaries of Anoia (upper panels) and those from tributaries of the lower Llobregat course (lower panels), contrasting with the counterpart kernel density estimates based on data generated from the estimated ESAG distribution (black solid lines).*

on 300 bootstrap data sets, each of the same size as the raw data sampled from the raw data with replacement. Similarly, we have the three estimated eigenvalues associated with $\hat{\mathbf{V}}_{\mathrm{L}}$ given by 0.19 (0.04), 0.54 (0.29), and 9.61 (1.84). Taking the estimated standard errors into consideration, with the large discrepancy between the estimated (and fixed) eigenvalues, neither of the two data sets provides sufficient evidence indicating partial isotropy.

For each component in (K^+, Na^+, Ca^{2+}, Mg^{2+}) at each of the two considered collections of locations, Figure 9 shows two estimated probability density functions: a kernel density estimate based on the transformed compositions, and a kernel density estimate based on data generated from the estimated ESAG distribution for that collection of locations. Marginally, the distribution of some components deduced from the estimated ESAG distribution, such as K^+ along tributaries of the lower Llobregat course and Mg^{2+} along Anoia tributaries, matches closely with the counterpart estimated distribution based on the transformed composition data; but some local features of certain components indicated by the density estimate based on the transformed raw data are not captured by the marginal distributions indicated by the estimated ESAG, such as the bimodality of Na^+ composition. Such mismatch between the two estimated marginal densities does not necessarily suggest that the marginal distribution induced from the estimated ESAG fits the data for that component poorly. After all, the exhibited local features may be due to the kernel density estimate overfitting the transformed composition data, especially when the sample size is as small as 43 for the second set of locations. On the other hand, because the square-root transformed compositional data only lie on the non-negative hyperoctant of the support of ESAG, fitting an ESAG to such data is practically more adequate when the mean direction is further away from the boundary of the non-negative hyperoctant, but can raise

concern otherwise as one may perceive from the estimated marginal densities. Most of the estimated densities in Figure 9, although supported on $\mathbb{R}$, do tail off quickly at the lower tails stretching towards zero from above, and thus using a marginal distribution supported on the entire real line to model non-negative data is practically acceptable in most cases in this particular application. However, for K^+ from Anoia tributaries that has the lowest composition among all considered components and locations, the estimated densities have lower tails relatively heavy and extending further into the negative half of the real line. A band-aid solution that may alleviate the concern is to transform the direction data again so that, after a second transformation, the mean direction is further away from the boundary of the non-negative hyperoctant of the unit hypersphere. An example of such transformation is outlined in equation (1) in [27], which involves re-scaling then re-normalizing $\mathbf{Y}$. To address this concern more formally from the modelling point of view, one may consider using a truncated ESAG supported on the non-negative hyperoctant of the unit hypersphere to model the square-root transformed compositional data. A complication in this solution is the normalization constant defined by a d-dimensional integral in the density of a truncated ESAG. Indeed, modelling compositional data using the directional data point of view deserves a systematic investigation to adequately address unique data features such as nearly zero or zero-inflated compositions that can often arise especially when d is large. We relegate this investigation to a separate study considering the length of the current manuscript.

Between the two collections of locations, the estimated marginal distributions of certain component appear to be substantially different, e.g., for Ca^{2+}. In fact, when we fit the ESAG model to the 110 records across these two sets of locations, we obtain an estimated p-value of 0.02 from the GOF test, with the corresponding QQ plot clearly deviating from a straight line (see the right panel in Figure 8). We thus conclude that an ESAG distribution is inadequate for modeling the data that mix compositional data from Anoia tributaries and those from tributaries of the lower Llobregat course. This lack of fit is not surprising because Anoia mostly passes through vineyards and industrialized zones, whereas the Llobergat lower course also flows through densely populated areas with high demands of water besides agricultural and industrial areas. This explains the vastly different patterns and sources of anthorpogenic and geological pollution between Anoia and the lower Llobregat course [10], which create substantial heterogeneity in the mixed compositional data that an ESAG model is unlikely to capture.

7. Discussion

Given the wide range of applications where directional data are of scientific interest and typically of dimension higher than three, an important first step towards sound statistical analysis of such data is the formulation of a directional distribution of arbitrary dimension. We adopt the initial formulation of the ESAG

distribution proposed by [20], and take it to the next level via a sequence of reparameterizations leading to a distribution family indexed by unconstrained parameters. The resultant parametric family for directional data avoids pitfalls that many existing directional distributions suffer so that, unlike the Kent distribution for instance, there is no hard-to-compute normalization constant in the density function, and it is easy to simulate data from an ESAG of any dimension. More importantly, the proposed parameterization of ESAG lends itself to straightforward maximum likelihood inference procedures that are numerically stable and less dependent on "good" starting values for parameter estimation. New parameters introduced along the way of reparameterization have statistically meaningful interpretations, which facilitate formulating hypothesis testing where one compares a reduced ESAG model, such as an isotropic or a partially isotropic model, with a saturated ESAG model. In summary, the proposed ESAG family of arbitrary dimension sets the stage for carrying out a full range of likelihood-based inference for directional data, including parameter estimation, uncertainty assessment, and hypothesis testing.

To ease the concerns of model misspecification when assuming a parametric family in a given application, we develop graphical and quantitative diagnostics methods that utilize directional residuals. Maximum likelihood estimation and the proposed diagnostics methods for ESAG can be easily implemented using the R code developed and maintained by the first author, available at `https://github.com/Zehaoyu217/ESAG/blob/main/ESAG.R`.

An immediate follow-up step is to consider regression models for directional data, which is well motivated by the lack of fit of a marginal ESAG distribution for the mixed compositional data entertained in Section 6. We conjecture that, conditioning on covariates relating to geological features of considered tributaries and covariates reflecting human activities developed in regions these tributaries running through, the mixed compositioinal data can be better modelled by an ESAG distribution with covariate-dependent $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$. With $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ ranging over the entire real space of adequate dimensions, the proposed ESAG family prepares itself well for regression analysis of directional data without using complicated link functions to introduce dependence of model parameters on covariates $\mathbf{W}$. For example, one may consider a fully parametric regression model as simple as $\mathbf{Y}|\mathbf{W} \sim \text{ESAG}(\boldsymbol{\mu}(\mathbf{W}), \mathbf{V}(\mathbf{W}))$, where $\boldsymbol{\mu}(\mathbf{W})$ is a linear function of covariates $\mathbf{W}$, and $\mathbf{V}(\mathbf{W})$ is determined by $\boldsymbol{\mu}(\mathbf{W})$ and $\boldsymbol{\gamma}(\mathbf{W})$, with the latter also a linear function of covariates. This is similar to but generalizes the second type of regression models considered in [21] for arbitrary $d \geq 3$. More flexible dependence structures of $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ on covariates are also worthy of consideration in the follow-up research along the line of regression analysis. Once we enter the realm of regression models, the dimension of the parameter space grows more quickly as d increases than before considering regression analysis for directional data. Upon completion of the study presented in this article, we have embarked on the exciting journey of developing scalable inference procedures suitable for settings with high dimensional parameter space following the strategies of frequentist penalized maximum likelihood estimation and Bayesian shrinkage estimation via hierarchical modeling.

Appendix A: Implication of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ being equivalent

Under the proposed parameterization of $\mathrm{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{V}$ is determined by $\boldsymbol{\gamma}$ after $\boldsymbol{\mu}$ is specified. We thus write $\mathbf{V}$ as $\mathbf{V}(\boldsymbol{\gamma})$ in this appendix, and view quantities related to $\mathbf{V}$ as functions of $\boldsymbol{\gamma}$, such as the eigenvalues of $\mathbf{V}$ and the radial parameters in (2.5). If $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are equivalent, then $\mathbf{V}(\boldsymbol{\gamma}) = \mathbf{V}(\boldsymbol{\gamma}')$, and thus $\mathbf{V}(\boldsymbol{\gamma})$ and $\mathbf{V}(\boldsymbol{\gamma}')$ share the same eigenvalues. By (2.5), $\{\lambda_j(\boldsymbol{\gamma}) = \lambda_j(\boldsymbol{\gamma}')\}_{j=1}^{d-1}$ implies that $\{r_j(\boldsymbol{\gamma}) = r_j(\boldsymbol{\gamma}')\}_{j=1}^{d-2}$. Lastly, from Section 2.4, $r_j = \|\tilde{\boldsymbol{\gamma}}_j\|$, for $j = 1, \ldots, d-2$. Therefore, if $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are equivalent, $\|\tilde{\boldsymbol{\gamma}}_j\| = r_j(\boldsymbol{\gamma}) = r_j(\boldsymbol{\gamma}') = \|\tilde{\boldsymbol{\gamma}}'_j\|$, for $j = 1, \ldots, d-2$.

Appendix B: Identifiability of $\widetilde{\boldsymbol{\Omega}}$

We provide a detailed discussion on the identifiability of parameters in $\widetilde{\boldsymbol{\Omega}}$ in this appendix. Before considering $d > 3$ in general, we first focus on the case with $d = 5$ as in Table 1 with $\widetilde{\boldsymbol{\Omega}} = (r_1, r_2, r_3, \theta_1, \theta_2, \theta_3, \tilde{\phi}_{2,1}, \tilde{\phi}_{3,1}, \tilde{\phi}_{3,2})^{\mathrm{T}}$ for ease of exposition. Under our proposed parameterization of ESAG, inference for the third group of parameters in $\widetilde{\boldsymbol{\Omega}}$, $(r_3, \theta_3, \tilde{\phi}_{3,1}, \tilde{\phi}_{3,2})$, directly relates to inference for $\tilde{\boldsymbol{\gamma}}_3 = (\gamma_{3,1}, \gamma_{3,2}, \gamma_{3,3}, \gamma_{3,4})^{\mathrm{T}}$. But, by the third group of spherical-to-Cartesian coordinates transformations in Table 1, if $\tilde{\phi}_{3,1} = 0$ (or π), then $\tilde{\boldsymbol{\gamma}}_3 = (r_3, 0, 0, 0)^{\mathrm{T}}$ (or $(-r_3, 0, 0, 0)^{\mathrm{T}}$), for all $(\theta_3, \tilde{\phi}_{3,2}) \in [-\pi, \pi) \times [0, \pi]$. Hence, $(\theta_3, \tilde{\phi}_{3,2})$ are not identifiable when $\tilde{\phi}_{3,1}$ is on the boundary. In (2.9), we set $\theta_3 = 0$ when $\gamma_{3,3}^2 + \gamma_{3,4}^2 = 0$, and set $\tilde{\phi}_{3,2} = 0$ when $\gamma_{3,2}^2 + \gamma_{3,3}^2 + \gamma_{3,4}^2 = 0$ (i.e., when $\tilde{\phi}_{3,1} = 0$ or π) for simplicity. In effect, when $\tilde{\gamma}_{3,1} > 0$, we map $\tilde{\boldsymbol{\gamma}}_3 = (\tilde{\gamma}_{3,1}, 0, 0, 0)^{\mathrm{T}}$ to $(r_3, \theta_3, \tilde{\phi}_{3,1}, \tilde{\phi}_{3,2}) = (\tilde{\gamma}_{3,1}, 0, 0, 0)$ following (2.9) despite the fact that, under the spherical coordinate system, all points in $\{(r_3, \theta_3, \tilde{\phi}_{3,1}, \tilde{\phi}_{3,2}) : r_3 = \tilde{\gamma}_{3,1}, \theta_3 \in [-\pi, \pi), \tilde{\phi}_{3,1} = 0, \tilde{\phi}_{3,2} \in [0, \pi]\}$ map to $\tilde{\boldsymbol{\gamma}}_3 = (\tilde{\gamma}_{3,1}, 0, 0, 0)^{\mathrm{T}}$ according to (2.8).

Generalizing to $d > 3$, one can see from (2.8) that $(\theta_j, \tilde{\phi}_{j,2}, \ldots, \tilde{\phi}_{j,j-1})$ are non-identifiable when $\tilde{\phi}_{j,1} = 0$ or π, for $j \in \{2, \ldots, d-2\}$. More generally, for $j \in \{3, \ldots, d-2\}$ when $d > 4$, $(\theta_j, \tilde{\phi}_{j,k+1}, \ldots, \tilde{\phi}_{j,j-1})$ are non-identifiable when $\tilde{\phi}_{j,k} = 0$ or π, for $k \in \{2, \ldots, j-2\}$. By convention (https://en.wikipedia.org/wiki/N-sphere), when $\tilde{\phi}_{j,k}$ lies on the boundary for some $j \in \{2, \ldots, d-2\}$ and some $k \le j-2$, we set the corresponding non-identifiable angles at zero in (2.9). By using (2.8) and (2.9) to relate $\widetilde{\boldsymbol{\Omega}}$ and $\boldsymbol{\gamma}$, we create certain ambiguity in the mappings that connect these two sets of parameters. Such ambiguity can be better apprehended using a partition of the parameter space associated with $\widetilde{\boldsymbol{\Omega}} = (r_1, \ldots, r_{d-2}, \theta_1, \ldots, \theta_{d-2}, \tilde{\boldsymbol{\phi}}_1^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\phi}}_{d-2}^{\mathrm{T}})^{\mathrm{T}}$. Denoted by $\mathcal{S}$ this parameter space, that is, $\mathcal{S} = (\mathbb{R}^+ \cup \{0\})^{d-2} \times [-\pi, \pi)^{d-2} \times [0, \pi]^{(d-2)(d-3)/2}$.

Define $\mathcal{S}_B$ as a subspace of $\mathcal{S}$ that includes all points with $\tilde{\phi}_{j,k}$ lying on the boundary for some $j \in \{2, \ldots, d-2\}$ and some $k \le j-2$. That is,

$$\mathcal{S}_B = \{\widetilde{\boldsymbol{\Omega}} : \tilde{\phi}_{j,k} = 0 \text{ or } \pi, \text{ for some } j \in \{2, \ldots, d-2\} \text{ and some } k \le j-2\}.$$

We further define a subset of $\mathcal{S}_B$,

$$\mathcal{S}_{B1} = \{\widetilde{\boldsymbol{\Omega}} : \tilde{\phi}_{j,k} = 0 \text{ or } \pi, \text{ for some } j \in \{2, \ldots, d-2\} \text{ and some } k \le j-2,$$

and $\theta_j = \tilde{\phi}_{j,k+1} = \cdots = \tilde{\phi}_{j,j-1} = 0\}$,

then define $\mathcal{S}_{B2} = \mathcal{S}_B \setminus \mathcal{S}_{B1}$, where $\setminus$ is the set subtraction operator. Lastly, let $\mathcal{S}_I = \mathcal{S} \setminus \mathcal{S}_B$. Now we have a partition of $\mathcal{S}$, $\mathcal{S}_I \cup \mathcal{S}_{B1} \cup \mathcal{S}_{B2}$. A partition of the parameter space associated with $\boldsymbol{\gamma}$ is $\mathcal{G}_I \cup \mathcal{G}_B$, where $\mathcal{G}_I = \{\mathcal{T}(\widetilde{\boldsymbol{\Omega}}) : \widetilde{\boldsymbol{\Omega}} \in \mathcal{S}_I\}$, $\mathcal{G}_B = \mathbb{R}^{(d-2)(d+1)/2} \setminus \mathcal{G}_I$, and the mapping $\mathcal{T} : \mathcal{S} \to \mathbb{R}^{(d-2)(d+1)/2}$ is specified by the transformations in (2.8). There is no ambiguity when limiting to the mapping $\mathcal{T} : \mathcal{S}_I \to \mathcal{G}_I$, which is bijective, and thus the inverse transformation $\mathcal{T}^{-1}(\boldsymbol{\gamma})$ is well-defined. Hence, when $\widetilde{\boldsymbol{\Omega}} \in \mathcal{S}_I$, all orientation parameters are identifiable. But the mapping $\mathcal{T} : \mathcal{S}_B \to \mathcal{G}_B$ is surjective (i.e., many-to-one) as suggested by our earlier remarks about non-identifiable angles when $\widetilde{\boldsymbol{\Omega}} \in \mathcal{S}_B$, and this is where the ambiguity in terms of inverse mapping arises. We circumvent this ambiguity in (2.9) by letting $\mathcal{T}^{-1}$ map from $\mathcal{G}_B$ to (only) $\mathcal{S}_{B1}$.

In the context of inferring ESAG$(\boldsymbol{\mu}, \mathbf{V})$, a direct consequence of the treatment in (2.9) to avoid ambiguity of $\mathcal{T}^{-1}$ is that, if $\mathbf{V} = \mathbf{V}_0$ is specified by $\boldsymbol{\mu}_0$ and $\widetilde{\boldsymbol{\Omega}}_0 \in \mathcal{S}_{B2}$, then there exists no point in the parameter space of $\boldsymbol{\gamma}$ such that $\mathbf{V}_0$ can be formulated by this point of $\boldsymbol{\gamma}$ along with $\boldsymbol{\mu}_0$. However, one can show that one can formulate a sequence of points in $\mathcal{G}_I$, $\{\boldsymbol{\gamma}_t : t = 1, 2, \ldots\}$, such that $\lim_{t\to\infty} \mathcal{T}^{-1}(\boldsymbol{\gamma}_t) = \widetilde{\boldsymbol{\Omega}}_0$. This hints at the possibility of achieving consistent estimation of $\mathbf{V}$ even when there is no point in the parameter space $\mathbb{R}^p$ associated with $\boldsymbol{\Omega} = (\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ that leads to $\mathbf{V}_0$. A simulation study presented next provides some empirical confirmation of this possibility.

In the simulation study, random samples of size n from ESAG$_3(\boldsymbol{\mu}, \mathbf{V})$ are generated, with $\boldsymbol{\mu} = (2, -5, 3, 5)^{\mathrm{T}}$ and $\mathbf{V}$ specified by this chosen $\boldsymbol{\mu}$ and three choices of $\widetilde{\boldsymbol{\Omega}}$: $(r_1, r_2, \theta_1, \theta_2, \tilde{\phi}_{2,1}) = (5.8, 5.4, \pi/3, 6\pi/7, \pi/3) \in \mathcal{S}_I$, $(3, 3, 0, 0, 0) \in \mathcal{S}_{B1}$, and $(3, 3, 0, \pi/4, 0) \in \mathcal{S}_{B2}$. The last setting of $\widetilde{\boldsymbol{\Omega}}$ is where θ_2 is non-identifiable and thus the corresponding $\mathbf{V}$ cannot be formulated by any point in the parameter space of $\boldsymbol{\Omega}$. Regardless, Figure B.1 provides empirical evidence, based on 1000 Monte Carlo replicates for each simulation setting as n varies, suggesting that the non-identifiabilty issue with the orientation parameter does not affect finite sample performance of the MLE for $\mathbf{V}$ when comparing with settings where all parameters in $\widetilde{\boldsymbol{\Omega}}$ are identifiable.

Appendix C: Expectations of compositions

Here, we show that, if $\mathbf{Y} \sim$ ESAG$_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, then

$$E(\mathbf{Y}^2) = \boldsymbol{\Xi}^2 E(\mathbf{K}^2), \tag{C.1}$$

where $\mathbf{K} = \boldsymbol{\Xi}^{\mathrm{T}}\mathbf{Y}$, and $\boldsymbol{\Xi} = [\boldsymbol{\xi}_d \mid \boldsymbol{\xi}_{d-1} \mid \ldots \mid \boldsymbol{\xi}_1]$, that is, the columns of $\boldsymbol{\Xi}$ are the eigenvectors of $\mathbf{V}$, with $\boldsymbol{\xi}_d = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$, corresponding to eigenvalues in the diagonal matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_d, \lambda_{d-1}, \ldots, \lambda_1)$, with $\lambda_d = 1$ and $0 < \lambda_1 \leq \cdots \leq \lambda_{d-1}$. This result is directly deduced from Proposition 1 in [25], which is applicable once the following three properties of $\mathbf{K} = (K_1, \ldots, K_d)^{\mathrm{T}}$ are established:

(i) $E(K_1) \geq 0$;

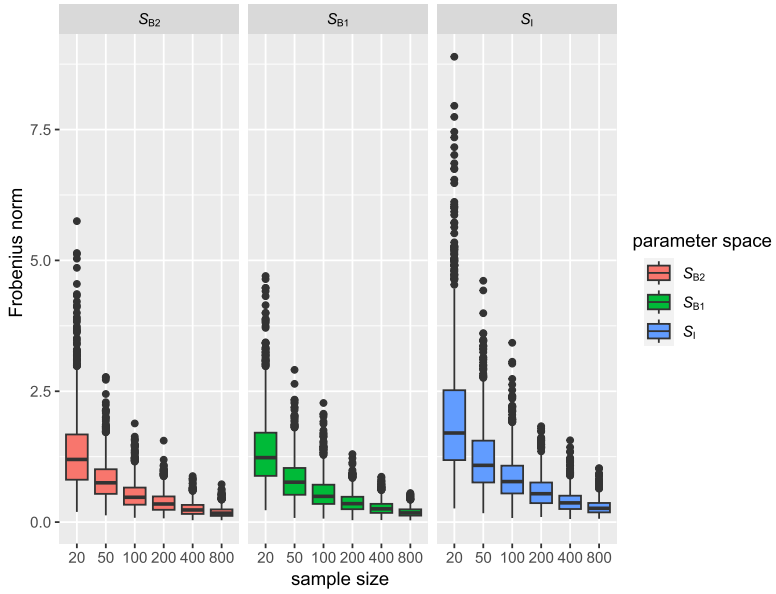

FIG B.1. *Boxplots of the Frobenius norm of* $\mathbf{V}-\hat{\mathbf{V}}$ *as sample size* n *varies when the truth of* $\widetilde{\boldsymbol{\Omega}}$ *is in* $\mathcal{S}_{B2}$ *(left panel, the case with a non-identifiable angle),* $\mathcal{S}_{B1}$ *(middle panel), and* $\mathcal{S}_I$ *(right panel), respectively.*

(ii) $E(K_2^2) \geq \cdots \geq E(K_d^2)$;
(iii) $g(\mathbf{K}) = g(\mathbf{HK})$, where $g(\cdot)$ is the density of $\mathbf{K}$, and $\mathbf{H} = \text{diag}(1, \pm 1, \ldots, \pm 1)$.

In other words, to prove (C.1), it suffices to show that (i)–(iii) hold for $\mathbf{K} = \boldsymbol{\Xi}^{\mathrm{T}}\mathbf{Y}$.

With $\mathbf{Y} \sim \text{ESAG}_{d-1}(\boldsymbol{\mu}, \mathbf{V})$, we may rewrite $\mathbf{Y} = \mathbf{X}/\|\mathbf{X}\|$, where $\mathbf{X} = \mathbf{V}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}$ and $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$. By the spectral decomposition theorem, $\mathbf{V} = \boldsymbol{\Xi}\boldsymbol{\Lambda}\boldsymbol{\Xi}^{\mathrm{T}}$, hence

$$\begin{aligned}\mathbf{K} &= \boldsymbol{\Xi}^{\mathrm{T}}\mathbf{Y} \\ &= (\boldsymbol{\Xi}^{\mathrm{T}}\mathbf{V}^{1/2}\mathbf{Z} + \boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\mu})/\|\mathbf{X}\| \\ &= (\boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Xi}^{\mathrm{T}}\mathbf{Z} + \boldsymbol{\Xi}^{\mathrm{T}}\boldsymbol{\mu})/\|\mathbf{X}\| \\ &= (\boldsymbol{\Lambda}^{1/2}\mathbf{U} + \|\boldsymbol{\mu}\|\mathbf{e}_1)/\|\mathbf{X}\|, \text{ where } \mathbf{U} = (U_1, \ldots, U_d)^{\mathrm{T}} = \boldsymbol{\Xi}^{\mathrm{T}}\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d),\end{aligned} \tag{C.2}$$

and therefore

$$K_1 = (U_1 + \|\boldsymbol{\mu}\|)/\|\mathbf{X}\|, \tag{C.3}$$

$$K_j = \sqrt{\lambda_{d-j+1}}U_j/\|\mathbf{X}\|, \text{ for } j = 2, \ldots, d. \tag{C.4}$$

Similarly, using $\mathbf{V}^{1/2} = \boldsymbol{\Xi}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Xi}^{\mathrm{T}}$, one can show that

$$\|\mathbf{X}\| = \sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} U_j^2 + (U_1 + \|\boldsymbol{\mu}\|)^2}. \tag{C.5}$$

Now we are ready to prove property (i). For a random variable A that can be multivariate in general, we use $F_A(a)$ to denote the cumulative distribution function of A evaluated at a. By (C.3) and (C.5),

$$\begin{aligned}
&E(K_1) \\
&= E\left\{ \frac{U_1 + \|\boldsymbol{\mu}\|}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} U_j^2 + (U_1 + \|\boldsymbol{\mu}\|)^2}} \right\} \\
&= \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} \frac{u_1 + \|\boldsymbol{\mu}\|}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + (u_1 + \|\boldsymbol{\mu}\|)^2}} \, dF_{U_1}(u_1) dF_{U_2,\ldots,U_d}(u_2, \ldots, u_d) \\
&= \int_{\mathbb{R}^{d-1}} \int_{-\infty}^{-\|\boldsymbol{\mu}\|} \frac{u_1 + \|\boldsymbol{\mu}\|}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + (u_1 + \|\boldsymbol{\mu}\|)^2}} \, dF_{U_1}(u_1) dF_{U_2,\ldots,U_d}(u_2, \ldots, u_d) \\
&\quad + \int_{\mathbb{R}^{d-1}} \int_{-\|\boldsymbol{\mu}\|}^{\infty} \frac{u_1 + \|\boldsymbol{\mu}\|}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + (u_1 + \|\boldsymbol{\mu}\|)^2}} \, dF_{U_1}(u_1) dF_{U_2,\ldots,U_d}(u_2, \ldots, u_d) \\
&= \int_{\mathbb{R}^{d-1}} \int_{-\infty}^{0} \frac{a}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + a^2}} \, dF_A(a) dF_{U_2,\ldots,U_d}(u_2, \ldots, u_d) \qquad \text{(C.6)} \\
&\quad + \int_{\mathbb{R}^{d-1}} \int_{0}^{\infty} \frac{a}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + a^2}} \, dF_A(a) dF_{U_2,\ldots,U_d}(u_2, \ldots, u_d), \qquad \text{(C.7)}
\end{aligned}$$

where we apply change of variable in (C.6) and (C.7) by letting $a = u_1 + \|\boldsymbol{\mu}\|$. Because $A = U_1 + \|\boldsymbol{\mu}\| \sim N(\|\boldsymbol{\mu}\|, 1)$, the inner integral in (C.6) is

$$\begin{aligned}
&\int_{-\infty}^{0} \frac{a}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + a^2}} \, dF_A(a) \\
&= \int_{-\infty}^{0} \frac{a}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + a^2}} \times \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(a - \|\boldsymbol{\mu}\|)^2}{2} \right\} da \\
&= -\int_{0}^{\infty} \frac{a}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + a^2}} \times \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(a + \|\boldsymbol{\mu}\|)^2}{2} \right\} da.
\end{aligned}$$

Combining this result for (C.6) with a similar elaboration of (C.7) gives

$$E(K_1) = \int_{\mathbb{R}^{d-1}} \int_{0}^{\infty} \frac{a}{\sqrt{\sum_{j=2}^{d} \lambda_{d-j+1} u_j^2 + a^2}} \frac{1}{2\pi} \left[\exp\left\{ -\frac{(a - \|\boldsymbol{\mu}\|)^2}{2} \right\} \right.$$

$$- \exp\left\{-\frac{(a+\|\boldsymbol{\mu}\|)^2}{2}\right\}\Bigg] \, da \, dF_{U_2,\ldots,U_d}(u_2,\ldots,u_d),$$

which is non-negative because $\exp\{-(a-\|\boldsymbol{\mu}\|)^2/2\} - \exp\{-(a+\|\boldsymbol{\mu}\|)^2/2\} \geq 0$, $\forall a > 0$. This completes the proof of (i).

Next we prove property (ii) that states $E(K_j^2 - K_\ell^2) \geq 0$, for $2 \leq j < \ell \leq d$. By (C.4) and (C.5),

$$\begin{aligned} E(K_j^2 - K_\ell^2) &= E\left\{\frac{\lambda_{d-j+1}U_j^2 - \lambda_{d-\ell+1}U_\ell^2}{\sum_{k=2}^d \lambda_{d-k+1}U_k^2 + (U_1+\|\boldsymbol{\mu}\|)^2}\right\} \\ &= \int_{\mathbb{R}^+}\int_{\mathbb{R}^2} \frac{a^2-b^2}{a^2+b^2+c}\, dF_{A,B}(a,b) dF_C(c), \end{aligned} \tag{C.8}$$

where we view $A = \lambda_{d-j+1}^{1/2}U_j$, $B = \lambda_{d-\ell+1}^{1/2}U_\ell$, $C = \sum_{k\neq 1,j,\ell}\lambda_{d-k+1}U_k^2 + (U_1 + \|\boldsymbol{\mu}\|)^2$, and C is a non-negative random variable by construction. Because $A \sim N(0,\lambda_{d-j+1})$, $B \sim N(0,\lambda_{d-\ell+1})$, and $A \perp B$, the inner integral in (C.8) is equal to

$$\begin{aligned} &\int_{\mathbb{R}^2} \frac{a^2-b^2}{a^2+b^2+c}\, dF_A(a) dF_B(b) \\ &= \int_{\mathbb{R}^2} \frac{a^2-b^2}{a^2+b^2+c} \times \frac{1}{2\pi\sqrt{\lambda_{d-j+1}\lambda_{d-\ell+1}}} \exp\left(-\frac{\lambda_{d-\ell+1}a^2+\lambda_{d-j+1}b^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}}\right) dadb \\ &= \int_{\{(a,b)\in\mathbb{R}^2:\, a^2<b^2\}} \frac{a^2-b^2}{a^2+b^2+c} \times \frac{1}{2\pi\sqrt{\lambda_{d-j+1}\lambda_{d-\ell+1}}} \\ &\quad \times \exp\left(-\frac{\lambda_{d-\ell+1}a^2+\lambda_{d-j+1}b^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}}\right) dadb \\ &\quad + \int_{\{(a,b)\in\mathbb{R}^2:\, a^2\geq b^2\}} \frac{a^2-b^2}{a^2+b^2+c} \times \frac{1}{2\pi\sqrt{\lambda_{d-j+1}\lambda_{d-\ell+1}}} \\ &\quad \times \exp\left(-\frac{\lambda_{d-\ell+1}a^2+\lambda_{d-j+1}b^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}}\right) dadb. \end{aligned}$$

For the first integral above, we apply change of variables by letting $(s,t) = (b,a)$ to re-express the first integral as

$$\begin{aligned} &\int_{\{(s,t)\in\mathbb{R}^2:\, s^2>t^2\}} \frac{t^2-s^2}{t^2+s^2+c} \times \frac{1}{2\pi\sqrt{\lambda_{d-j+1}\lambda_{d-\ell+1}}} \\ &\quad \times \exp\left(-\frac{\lambda_{d-\ell+1}t^2+\lambda_{d-j+1}s^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}}\right) dsdt \\ &= -\int_{\{(a,b)\in\mathbb{R}^2:\, a^2>b^2\}} \frac{a^2-b^2}{a^2+b^2+c} \times \frac{1}{2\pi\sqrt{\lambda_{d-j+1}\lambda_{d-\ell+1}}} \\ &\quad \times \exp\left(-\frac{\lambda_{d-j+1}a^2+\lambda_{d-\ell+1}b^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}}\right) dadb. \end{aligned}$$

Combining this expression for the first integral with the second integral yields

$$\begin{aligned}
&\int_{\mathbb{R}^2} \frac{a^2-b^2}{a^2+b^2+c}\, dF_A(a) dF_B(b) \\
&= \int_{\{(a,b)\in\mathbb{R}^2:\, a^2\geq b^2\}} \frac{a^2-b^2}{a^2+b^2+c} \times \frac{1}{2\pi\sqrt{\lambda_{d-j+1}\lambda_{d-\ell+1}}} \\
&\quad \times \left\{ \exp\left(-\frac{\lambda_{d-\ell+1}a^2+\lambda_{d-j+1}b^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}} \right) - \exp\left(-\frac{\lambda_{d-j+1}a^2+\lambda_{d-\ell+1}b^2}{2\lambda_{d-j+1}\lambda_{d-\ell+1}} \right) \right\} da db,
\end{aligned}$$

which is non-negative because $(\lambda_{d-\ell+1}a^2+\lambda_{d-j+1}b^2)-(\lambda_{d-j+1}a^2+\lambda_{d-\ell+1}b^2) = (\lambda_{d-\ell+1}-\lambda_{d-j+1})(a^2-b^2) \leq 0$ when $a^2 \geq b^2$ since $\lambda_{d-j+1} \geq \lambda_{d-\ell+1}$, for $2 \leq j < \ell \leq d$. Using this result for the inner integral in (C.8) shows that $E(K_j^2 - K_\ell^2) \geq 0$, for $2 \leq j < \ell \leq d$. This completes the proof of (ii).

Lastly, we show property (iii) stating that $\mathbf{K} \overset{\mathcal{L}}{=} \mathbf{HK}$ for $\mathbf{H} = \mathrm{diag}(1, \pm 1, \ldots, \pm 1)$. By (C.2),

$$\begin{aligned}
\mathbf{HK} &= (\mathbf{H}\boldsymbol{\Lambda}^{1/2}\mathbf{U} + \|\boldsymbol{\mu}\|\mathbf{H}\mathbf{e}_1)/\|\mathbf{X}\| \\
&= (\mathbf{H}\boldsymbol{\Lambda}^{1/2}\mathbf{U} + \|\boldsymbol{\mu}\|\mathbf{e}_1)/\|\mathbf{X}\| \\
&\overset{\mathcal{L}}{=} (\boldsymbol{\Lambda}^{1/2}\mathbf{U} + \|\boldsymbol{\mu}\|\mathbf{e}_1)/\|\mathbf{X}\| \\
&= \mathbf{K},
\end{aligned}$$

where the second to last equation is by the fact that $\boldsymbol{\Lambda}^{1/2}\mathbf{U} \sim N_d(\mathbf{0}, \boldsymbol{\Lambda})$ and thus $\mathbf{H}\boldsymbol{\Lambda}^{1/2}\mathbf{U} \sim N_d(\mathbf{0}, \boldsymbol{\Lambda})$ since $\mathbf{H}\boldsymbol{\Lambda}\mathbf{H} = \boldsymbol{\Lambda}$. This completes the proof of (iii).

With properties (i)–(iii) established for $\mathbf{K}$, by Proposition 1 in [25], we now have $E(\mathbf{YY}^\mathrm{T}) = \boldsymbol{\Xi}\,\mathrm{diag}(K_1^2, \ldots, K_d^2)\,\boldsymbol{\Xi}^\mathrm{T}$, which implies (C.1).

Appendix D: Proof of equation (4.3)

By the spectral decomposition theorem, $\mathbf{V}^\alpha = \mathbf{P}\mathbf{D}^\alpha\mathbf{P}^\mathrm{T}$, where $\mathbf{D}^\alpha = \mathrm{diag}(\lambda_1^\alpha, \ldots, \lambda_d^\alpha)$ and $\mathbf{P} = [\boldsymbol{\xi}_1 \mid \ldots \mid \boldsymbol{\xi}_d]$. Let $\mathbf{P}_{-d} = [\boldsymbol{\xi}_1 \mid \ldots \mid \boldsymbol{\xi}_{d-1}]$. Using this decomposition with $\alpha = -1$ and $1/2$, we have

$$\begin{aligned}
Q &= \mathbf{r}^\mathrm{T}\mathbf{V}^{-1}\mathbf{r} \\
&= \frac{\mathbf{Z}^\mathrm{T}}{\|\mathbf{X}\|}\mathbf{V}^{1/2}\mathbf{P}_{-d}\mathbf{P}_{-d}^\mathrm{T} \times \mathbf{V}^{-1} \times \mathbf{P}_{-d}\mathbf{P}_{-d}^\mathrm{T}\mathbf{V}^{1/2}\frac{\mathbf{Z}}{\|\mathbf{X}\|} \\
&= \frac{\mathbf{Z}^\mathrm{T}}{\|\mathbf{X}\|^2}\mathbf{P}\mathbf{D}^{1/2}\mathbf{P}^\mathrm{T}\mathbf{P}_{-d}\mathbf{P}_{-d}^\mathrm{T} \times \mathbf{P}\mathbf{D}^{-1}\mathbf{P}^\mathrm{T} \times \mathbf{P}_{-d}\mathbf{P}_{-d}^\mathrm{T}\mathbf{P}\mathbf{D}^{1/2}\mathbf{P}^\mathrm{T}\mathbf{Z},
\end{aligned}$$

where

$$\mathbf{P}^\mathrm{T}\mathbf{P}_{-d} = \begin{bmatrix} \mathbf{P}_{-d}^\mathrm{T} \\ \boldsymbol{\xi}_d^\mathrm{T} \end{bmatrix} \mathbf{P}_{-d} = \begin{bmatrix} \mathbf{I}_{d-1} \\ \mathbf{0}^\mathrm{T} \end{bmatrix},$$

and thus

$$\mathbf{P}^\mathrm{T}\mathbf{P}_{-d}\mathbf{P}_{-d}^\mathrm{T}\mathbf{P} = \begin{bmatrix} \mathbf{I}_{d-1} & \mathbf{0} \\ \mathbf{0}^\mathrm{T} & 0 \end{bmatrix} \triangleq \tilde{\mathbf{I}}_d.$$

It follows that

$$\begin{aligned} Q &= \frac{1}{\|\mathbf{X}\|^2}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\mathbf{D}^{1/2}\tilde{\mathbf{I}}_d\mathbf{D}^{-1}\tilde{\mathbf{I}}_d\mathbf{D}^{1/2}\mathbf{P}^{\mathrm{T}}\mathbf{Z} \\ &= \frac{1}{\|\mathbf{X}\|^2}\mathbf{Z}^{\mathrm{T}}\mathbf{P}\tilde{\mathbf{I}}_d\mathbf{D}^{1/2}\mathbf{D}^{-1}\mathbf{D}^{1/2}\tilde{\mathbf{I}}_d\mathbf{P}^{\mathrm{T}}\mathbf{Z} \\ &= \frac{1}{\|\mathbf{X}\|^2}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{I}}_d\tilde{\mathbf{I}}_d\mathbf{U}, \text{ where } \mathbf{U} = \mathbf{P}^{\mathrm{T}}\mathbf{Z}, \\ &= \frac{1}{\|\mathbf{X}\|^2}\mathbf{U}_{-d}^{\mathrm{T}}\mathbf{U}_{-d}, \text{ where } \mathbf{U}_{-d} = \tilde{\mathbf{I}}_d\mathbf{U}, \end{aligned}$$

which gives (4.3).

Acknowledgments

We are grateful to the two anonymous referees for their detailed suggestions and thought-provoking comments on earlier versions this manuscript. Their suggestions greatly improved the manuscript, and their comments inspired new ideas for our follow-up research.

References

[1] Al Yammahi, A., Marpu, P. R. and Ouarda, T. B. (2021). Modeling directional distributions of wind data in the United Arab Emirates at different elevations. *Arabian Journal of Geosciences* **14** 774.

[2] Alenazi, A. (2021). A review of compositional data analysis and recent advances. *Communications in Statistics – Theory and Methods* 1–33. MR4608904

[3] Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics* 1148–1159. MR0145607

[4] Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S. and Ridgeway, G. (2005). Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research* **6**. MR2249858

[5] Blumenson, L. (1960). A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly* **67** 63–66. MR1530579

[6] Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning. MR1051420

[7] Chakravarti, I. M., Laha, R. G. and Roy, J. (1967). Handbook of methods of applied statistics. *Wiley Series in Probability and Mathematical Statistics (USA) eng*. MR0225405

[8] Dortet-Bernadet, J.-L. and Wicker, N. (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. *Biostatistics* **9** 66–80.

[9] Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons. MR0633058

[10] González, S., López-Roldán, R. and Cortina, J.-L. (2012). Presence and biological effects of emerging contaminants in Llobregat River basin: a review. *Environmental Pollution* **161** 83–92.
[11] Hernandez-Stumpfhauser, D., Breidt, F. J. and van der Woerd, M. J. (2017). The general projected normal distribution of arbitrary dimension: Modeling and Bayesian inference. *Bayesian Analysis* **12** 113–133. MR3597569
[12] Jupp, P. (1988). Residuals for directional data. *Journal of Applied Statistics* **15** 137–147.
[13] Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)* **44** 71–80. MR0655376
[14] Lee, A. (2010). Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics* **2** 477–486.
[15] Mardia, K. V. (2014). *Statistics of directional data*. Academic press. MR0336854
[16] Mardia, K. V., Foldager, J. I. and Frellsen, J. (2018). Directional statistics in protein bioinformatics. In *Applied Directional Statistics: Modern Methods and Case Studies* 1–24. CRC Press.
[17] Miller, K. (1964). Distributions involving norms of correlated Gaussian vectors. *Quarterly of Applied Mathematics* **22** 235–243.
[18] Murnaghan, F. D. (1952). The element of volume of the rotation group. *Proceedings of the National Academy of Sciences of the United States of America* **38** 69. MR0047049
[19] Otero, N., Tolosana-Delgado, R., Soler, A., Pawlowsky-Glahn, V. and Canals, A. (2005). Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean river. *Water Research* **39** 1404–1414.
[20] Paine, P., Preston, S. P., Tsagris, M. and Wood, A. T. (2018). An elliptically symmetric angular Gaussian distribution. *Statistics and Computing* **28** 689–697. MR3761349
[21] Paine, P. J., Preston, S., Tsagris, M. and Wood, A. T. (2020). Spherical regression models with general covariates and anisotropic errors. *Statistics and Computing* **30** 153–165. MR4057477
[22] Pawlowsky-Glahn, V. and Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications* **264** 1–10.
[23] Presnell, B., Morrison, S. P. and Littell, R. C. (1998). Projected multivariate linear models for directional data. *Journal of the American Statistical Association* **93** 1068–1077. MR1649201
[24] Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons. MR2401650
[25] Rivest, L.-P. (1984). On the information matrix for symmetric distributions on the hypersphere. *The Annals of Statistics* **12** 1085–1089. MR0751295
[26] Scealy, J. and Welsh, A. (2011). Regression for compositional data by

using distributions defined on the hypersphere. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 351–375. MR2815780

[27] Scealy, J. and Wood, A. T. (2019). Scaled von Mises–Fisher distributions and regression models for paleomagnetic directional data. *Journal of the American Statistical Association*. MR4047280

[28] Shi, P., Zhang, A. and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* **10** 1019–1040. MR3528370

[29] Soler, A., Canals, A., Goldstein, S., Otero, N., Antich, N. and Spangenberg, J. (2002). Sulfur and strontium isotope composition of the Llobregat River (NE Spain): tracers of natural and anthropogenic chemicals in stream waters. *Water, Air, and Soil Pollution* **136** 207–224.

[30] Soukissian, T. H. and Karathanasi, F. E. (2021). Joint modelling of wave energy flux and wave direction. *Processes* **9** 460.

[31] Van den Boogaart, K. G. and Tolosana-Delgado, R. (2008). "Compositions": a unified R package to analyze compositional data. *Computers & Geosciences* **34** 320–338.

[32] Wang, F. and Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology* **10** 113–127. MR2974815

[33] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society* 1–25. MR0640163