Nonparametric rotations for sphere-sphere regression

December 12, 2017

Authors' Footnote:

Abstract

Regression of data represented as points on a hypersphere has traditionally been treated using parametric families of transformations that include the simple rigid rotation as an important, special case. On the other hand, nonparametric methods have generally focused on modelling a scalar response through a spherical predictor by representing the regression function as a polynomial, leading to component-wise estimation of a spherical response. We propose a very flexible, simple regression model where for each location of the manifold a specific rotation matrix is to be estimated. To make this approach tractable, we assume continuity of the regression function that, in turn, allows for approximations of rotation matrices based on a series expansion. It is seen that the non-rigidity of our technique motivates an iterative estimation within a Newton-Raphson learning scheme which exhibits bias reduction properties. Extensions to general shape matching are also outlined. Both simulations and real data are used to illustrate the results.

KEYWORDS: Bias Reduction, Fisher's Method of Scoring, Local Smoothing, Non-Rigid Rotation Estimation, Singular Value Decomposition, Skew-symmetric Matrices, Spherical Kernels, Wahba's Problem

1. INTRODUCTION

1.1 Motivation and literature

There are two main categories of spherical data: *directional* and *shape* data. Standard examples of directional phenomena are: animal movements, cosmic rays, winds, ocean currents, in which a direction can be represented as a point on the unit sphere. Recent fields of interest include genome sequence representations, text analysis and clustering, morphometrics, and computer vision. In shape analysis (Dryden and Mardia, 2016) the similarity between two objects, each being represented by a set of landmarks, is judged after superimposing them by translation, scaling and rotation. A further domain for spherical data is in the field of *compositional data analysis*, i.e. positive vectors whose components add to a given constant. If the latter is set to one, a square root transformation puts these data onto the unit hypersphere. This approach has been successfully used by Wang et al. (2007) as a model for forecasting a time series of compositional data.

Our objective is to obtain a general approach to relate paired spherical data using rotations. We start with some data examples from various fields in order to highlight the potential scientific interest. The first, which uses directional data, is to describe the location of magnetic poles using a type of autoregressive model. Specifically, the location of the magnetic North pole (Fig. 3 of supplementary material) changes each year, and our objective is to predict next year's location using previous data. Predicting this change, and the rate of change, is of interest to geophysicists and geologists. An example using shape data considers tectonic plate movement (Chang, 1986). The interest lies in the relative motion of a tectonic plate from another considered as fixed in its present location, which leads to superposition of two objects, here represented by 11 locations on the Earth's surface (Figure 1). The solution can be used in a residual analysis, prediction of new landmarks, and leads to a better understanding of the distribution of species and ecosystems in space and time. Our final example uses experimental data from vectorcardiograms, which capture aspects of the electrocardiogram in the form of "loops". Data are related to children of different ages and genders, using two different lead systems. Interest here naturally lies in any difference between gender, between the ages or between the two systems. Using summary of data in form of unit vectors, the systems can be compared by estimating a rotation which maps one system to the other (Downs, 1972; Rosenthal et al., 2014). Section 7 and supplementary material contain additional details. Further examples of finding an *optimal* rotation as a link between two scatters of spherical locations can be found in Mardia & Jupp (2000, p.259).

Given the unit hypersphere $\mathbb{S}^{d-1} := \{\boldsymbol{u} \in \mathbb{R}^d : ||\boldsymbol{u}|| = 1\}$, $d \geq 2$, consider pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, $i \in (1, \ldots, n)$. Letting $\mathrm{SO}(d)$ be the set of all *proper* rotations on $\mathbb{R}^d$ (orthogonal matrices with determinant 1), the optimal rotation, in the least squares sense, to superimpose the two sets of points is described by

$$\underset{\boldsymbol{R} \in \mathrm{SO}(d)}{\operatorname{argmin}} \sum_{i=1}^{n} ||\boldsymbol{y}_i - \boldsymbol{R}\boldsymbol{x}_i||^2, \tag{1}$$

that can be explained as follows. The matrix $\boldsymbol{R}$ rotates $\boldsymbol{x}_i$ in the d-dimensional Euclidean space through an angle about the origin of the Cartesian coordinate system. The classical way to solve this problem is to use the Singular Value Decomposition (SVD) of $\boldsymbol{Y}^T\boldsymbol{X}$, in which $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both $n \times d$ matrices with $\boldsymbol{x}_i^T$ and $\boldsymbol{y}_i^T$ as their respective ith rows. Then, to preclude those solutions which include a reflection, use $\hat{\boldsymbol{R}} = \boldsymbol{U}^T\boldsymbol{\Delta}\boldsymbol{V}$ in which $\boldsymbol{U}$ and $\boldsymbol{V}$ are obtained from the SVD: $\boldsymbol{Y}^T\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, and $\boldsymbol{\Delta}$ is a diagonal matrix of order d with entries $(1, \ldots, 1, |\boldsymbol{U}^T\boldsymbol{V}|)$, where $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$.

When we assume that the quantity $\boldsymbol{y}_i - \boldsymbol{R}\boldsymbol{x}_i$ is a realization of a random variable, we obtain an inferential version of problem (1), called *spherical regression.* Inference was first considered by Chang (1986), who used maximum likelihood with a rotationally symmetric error distribution. However, estimating a simple rotation, although widely used, is a very crude way to model regression, as it corresponds to simple data translation in the Euclidean setting. Even in the classic, favourable example used by Chang (1986) and discussed by Chang (1989) and Rivest (1989), a rigid fit seems hard to confirm, as can be seen in Figure 1. We can note a somewhat unsatisfactory fit – though the sample size is small – since these data are virtually noiseless (errors in plate tectonic data range from 2 to 20 *km* to be compared with the Earth's circumference). The reader is referred to Chang et al. (2000) for more sophisticated spherical regression techniques in plate tectonics.

A more general family of transformations, i.e. the Möbius group, has been introduced by Downs (2003) for the customary case $d = 3$. A further generalization has been recently proposed by Rosenthal et al. (2014), who considered a larger parametric family than the rotation group. Specifically, for a non-singular

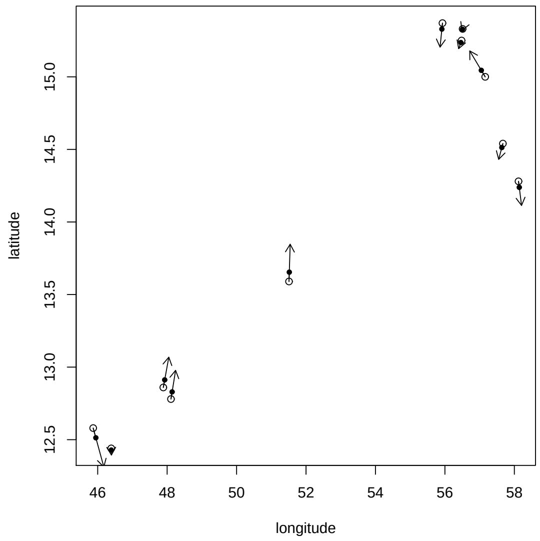

Figure 1: Open circles represent Somalian tectonic plate, black circles represent its reconstruction by rotating Arabian tectonic plate. The length of the arrows is proportional to the magnitude of the errors.

matrix $\boldsymbol{A}$ of order d they consider the transformations defined as $\boldsymbol{x} \to \boldsymbol{A}\boldsymbol{x}/||\boldsymbol{A}\boldsymbol{x}||$, and use a von Mises-Fisher distribution to model the noise, and to specify the likelihood function for parameter estimation. Despite these advances, the authors argue that the richest approach is the nonparametric one, but defer this as a topic for future research. A recent development of this idea is due to Rosenthal et al. (2017), where diffeomorphisms are considered to model spherical-spherical regression. Penalised maximum likelihood via gradient-based optimization is implemented for the three dimensional case. This strategy could be considered somewhat extreme whereas a very general model for the regression function is proposed, but it is still required that random errors are homoscedastic and follow a von Mises-Fisher distribution.

As opposed to the above parametric strategies, we could cite a number of nonparametric methods for regression or interpolation of spherical data. Their common strategy is similar to the Euclidean one; see, for example, Ruppert & Wand (1994), where a polynomial is used to locally (splines, needlets or Taylor-like ones) or globally (spherical harmonics) model the regression function. As a result, they generally work componentwise when used to predict spherical responses. A serious problem with these approaches obviously lies in explicitly modelling (or excluding) any correlation between dimensions, that, due to the spherical geometry, is customarily very relevant. For details, see Di Marzio et al. (2014), who proposed Taylor-like polynomials.

1.2 Main idea

Consider a pair of random variables $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$, both taking values on $\mathbb{S}^{d-1}$. Assume that the regression of $\boldsymbol{\mathcal{Y}}$ on $\boldsymbol{\mathcal{X}}$ exists for each $\boldsymbol{x} \in \mathbb{S}^{d-1}$. Since any two points on the sphere are related by a rotation describing the shortest arc joining them, we could always write such regression function by specifying a distinct rotation for each predictor value $\boldsymbol{x}$

$$\mathsf{E}[\boldsymbol{\mathcal{Y}} \,|\, \boldsymbol{\mathcal{X}} = \boldsymbol{x}] = \boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{x}\,. \tag{2}$$

As for the experimental error, it is naturally represented as a small random rotation of the true regression function. For more details on this see Rancourt et al. (2000) and Jupp (1988). To conveniently represent it, we need to preliminarily recall the formula of the matrix exponential, i.e., for a matrix $\boldsymbol{A}$ of order d, $\exp(\boldsymbol{A}) = \boldsymbol{I}_d + \boldsymbol{A} + \boldsymbol{A}^2/2 + \cdots$, with $\boldsymbol{I}_d$ denoting the identity matrix of order d. Any rotation matrix $\boldsymbol{R}$ has an exponential form $\boldsymbol{R} = \exp(\boldsymbol{S})$, where $\boldsymbol{S}$ is a skew-symmetric matrix, i.e. $\boldsymbol{S}^T = -\boldsymbol{S}$. Using more general concepts, we would say that the real skew-symmetric matrices, that constitute a *Lie algebra*, are mapped into the *Lie group* of orthogonal matrices by the matrix exponential. For a gentle introduction to computational results regarding exponentials of skew-symmetric matrices see Gallier & Xu (2003).

The above discussion motivates the regression for independent copies $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$ of $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ as

$$\boldsymbol{y}_i = \exp\left(\Phi(\boldsymbol{\varepsilon}_i)\right) \boldsymbol{R}_{\boldsymbol{x}_i}\boldsymbol{x}_i, \qquad i \in (1, \ldots, n), \tag{3}$$

where the function $\Phi(\boldsymbol{a})$ maps an $\mathbb{R}^d$ vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_{d(d-1)/2})^T$ into a skew-symmetric matrix; for example, for $d = 3$ we could use

$$\Phi(\boldsymbol{a}) = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix};$$

and the $\boldsymbol{\varepsilon}_i$s are random error terms satisfying $\mathsf{E}\left[\boldsymbol{\varepsilon}_i \mid \boldsymbol{x}_i\right] = \boldsymbol{0}_d$, where $\boldsymbol{0}_d$ stands for a d-dimensional vector of zeros, and $\mathsf{Var}\left[\boldsymbol{y}_i \mid \boldsymbol{x}_i\right] = \boldsymbol{\Sigma}_{\boldsymbol{x}_i}$, with $\boldsymbol{\Sigma}_{\boldsymbol{x}_i}$ being a matrix of order d with finite entries. Observe that, if $\boldsymbol{\varepsilon}_i$ has entries close to zero, then $\exp(\Phi(\boldsymbol{\varepsilon}_i)) \approx \boldsymbol{I}_d$. This is equivalent to assuming that the distribution of $\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mathcal{X}} = \boldsymbol{x}_i$ has expectation $\boldsymbol{R}_{\boldsymbol{x}_i}\boldsymbol{x}_i$. An advantage of this error characterization is that $\boldsymbol{\varepsilon}_i$ can be assumed to be a d-dimensional Euclidean random variable, making the model formulation more familiar.

Remark 1. *In our context, the usual Euclidean additive error undoubtedly appears to be somewhat more artificial than a rotational one because it cannot be a realization of a spherical density. In fact, sphere locations are not closed with respect to addition because the sphere is not a convex space. Specifically, it could even appear restrictive when we require that its distribution needs to be symmetric around the null direction. On the other hand, the multiplicative version used in model (3) makes things easy. For example, Downs (1972) requires that* $\boldsymbol{\varepsilon}_i$*s are independent copies of a normal random vector (which could be seen* *as a multivariate version of the wrapped normal distribution). Surely, in our nonparametric context the noise distribution does not need to belong to a known parametric family. Conversely, a maximum likelihood approach could be more appropriate.*

Model (3) is too general for inferential purposes because it requires the estimation of a distinct rotation matrix corresponding to each $\boldsymbol{x}_i$. However, it could be made more tractable by assuming that, within wide enough regions, the rotation matrices are "similar". In this paper we discuss model (3) under the assumption that the regression function (2) is continuous, i.e. $\lim_{\boldsymbol{x_i} \to \boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x_i}} = \boldsymbol{R}_{\boldsymbol{x}}$. This assumption motivates: (a) a local approximation of $\boldsymbol{R}_{\boldsymbol{x}_i}$ in terms of $\boldsymbol{R}_{\boldsymbol{x}}$ under suitable smoothness conditions, and (b) the use of the whole sample in the inference about $\boldsymbol{R}_{\boldsymbol{x}}$, provided that $\boldsymbol{x}_i$ contributes inversely to its distance from $\boldsymbol{x}$. In such a framework the shape of the scatter of predicted values is, in general, different from the shape of predictors. Such a transformation is usually referred to as a *non-rigid* rotation. Notice that, differently from *any* previous nonparametric technique, our scenario is set up as *simple regression* – although the observations are multidimensional – and the need to model correlation is removed.

The paper is organized as follows. Section 2 explores some mathematical tools, establishing that a specific estimator corresponds to a given order of a Taylor-like approximation of $\boldsymbol{R}_{\boldsymbol{x}_i}$. The simplest case, i.e. the one-term approximation estimator, is studied in Section 3. Section 4 contains some theory for the two-term approximation. Motivated by the fact that a smoothing process yields biased estimates, Section 5 introduces a Newton-Raphson algorithm that iterates the estimate using a progressive bias reduction. Section 6 investigates performance of our methods by means of simulations. Section 7 shows a real data experiment, also considering other methods for comparison. Extensions to shape matching is discussed in the concluding Section 8. Supplementary material contains additional simulations and real data case studies.

2. EXPANSIONS AND APPROXIMATIONS

Since the model associates a distinct rotation with each location on the sphere, it is natural to assume dependency of the entries of rotation matrices from the associated locations. Therefore, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$, we write $\boldsymbol{R}_{\boldsymbol{x}} = \exp(\boldsymbol{S}_{\boldsymbol{x}})$, where

$$\boldsymbol{S}_{\boldsymbol{x}} = \begin{bmatrix} 0 & s_{12}(\boldsymbol{x}) & \cdots & \cdots & s_{1d}(\boldsymbol{x}) \\ -s_{12}(\boldsymbol{x}) & 0 & \cdots & \cdots & s_{2d}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & s_{(d-1)\,d}(\boldsymbol{x}) \\ -s_{1d}(\boldsymbol{x}) & -s_{2d}(\boldsymbol{x}) & \cdots & -s_{(d-1)\,d}(\boldsymbol{x}) & 0 \end{bmatrix}.$$

Now, before introducing a local approximation for $\boldsymbol{R}_{\boldsymbol{x}_i}$, we need to recall some basic facts about parametrization and expansion of a function defined on $\mathbb{S}^{d-1}$. In particular, the ensuing discussion will provide tools aimed to nonparametrically model the entries of the skew-symmetrix matrices $\boldsymbol{S}_{\boldsymbol{x}_i}$, $i \in (1, \ldots, n)$, which, through matrix exponential, define the rotation matrices $\boldsymbol{R}_{\boldsymbol{x}_i}$. An obvious alternative, although not pursued in the literature, would be to consider parametric expressions for the entries of these matrices. In this latter case we could still preserve flexibility by using the weighting scheme as discussed in the sequel.

Coming to the parametrization, given $\boldsymbol{x} \in \mathbb{S}^{d-1}$, any vector $\boldsymbol{u} \in \mathbb{S}^{d-1}$ can be expressed as

$$\boldsymbol{u}(\boldsymbol{\xi}, \theta) := \boldsymbol{x}\cos(\theta) + \boldsymbol{\xi}\sin(\theta), \tag{4}$$

where $\theta = \arccos\left(\boldsymbol{u}^T\boldsymbol{x}\right)$, and $\boldsymbol{\xi}$ is a vector orthogonal to $\boldsymbol{x}$. Now, given a function $g : \mathbb{S}^{d-1} \to \mathbb{R}$, denoting as μ_d the Lebesgue measure of $\mathbb{S}^d$, and letting $\mathbb{T}_{\boldsymbol{x}} := \{\boldsymbol{\xi} \in \mathbb{S}^{d-1} : \boldsymbol{\xi} \perp \boldsymbol{x}\}$, the integration formula corresponding to parametrization (4) is

$$\int_{\mathbb{S}^{d-1}} g(\boldsymbol{u})\mathrm{d}\mu_{d-1}(\boldsymbol{u}) = \int_0^\pi \mathrm{d}\theta \sin^{d-2}(\theta) \int_{\mathbb{T}_{\boldsymbol{x}}} g(\boldsymbol{u}(\boldsymbol{\xi}, \theta))\mathrm{d}\mu_{d-2}(\boldsymbol{\xi}). \tag{5}$$

Moreover, letting $\bar{g}(\boldsymbol{x}) := g(\boldsymbol{x}/||\boldsymbol{x}||)$ be the homogeneous extension of g to $\mathbb{R}^d \setminus \{\boldsymbol{0}_d\}$, we have that

$$\left.\frac{\partial^\ell}{\partial\theta^\ell} g(\boldsymbol{u}(\boldsymbol{\xi}, \theta))\right|_{\theta=0} = \mathcal{D}_{\boldsymbol{\xi}}^{(\ell)}\bar{g}(\boldsymbol{x}),$$

where $\mathcal{D}_{\boldsymbol{\xi}}^{(\ell)}\bar{g}(\boldsymbol{x})$ is the directional derivative of order ℓ of $\bar{g}$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{\xi}$. Clearly $\mathcal{D}_{\boldsymbol{\xi}}^{(0)}\bar{g}(\boldsymbol{x}) = g(\boldsymbol{x})$, while, letting $\nabla\bar{g}(\boldsymbol{x})$ and $\nabla^2\bar{g}(\boldsymbol{x})$ respectively be the gradient and the Hessian matrix of $\bar{g}$ at $\boldsymbol{x}$, we have $\mathcal{D}_{\boldsymbol{\xi}}^{(1)}\bar{g}(\boldsymbol{x}) = \boldsymbol{\xi}^T\nabla\bar{g}(\boldsymbol{x})$, and $\mathcal{D}_{\boldsymbol{\xi}}^{(2)}\bar{g}(\boldsymbol{x}) = \boldsymbol{\xi}^T\nabla^2\bar{g}(\boldsymbol{x})\boldsymbol{\xi}$. Further, under continuity assumptions of the directional derivatives up to a suitable order, a pth-order Taylor series expansion of g around $\boldsymbol{x}$ yields

$$g(\boldsymbol{u}) \approx g(\boldsymbol{x}) + \sum_{\ell=1}^{p} \frac{\theta^\ell}{\ell!}\mathcal{D}_{\boldsymbol{\xi}}^{(\ell)}\bar{g}(\boldsymbol{x}). \tag{6}$$

Expansion (6) has been employed for deriving the asymptotic properties of kernel estimators for spherical densities by Hall et al. (1987) and Klemelä (2000), and to obtain a component-wise local approximation of spherical-spherical regression by Di Marzio et al. (2014).

Now, provided that the homogeneous extension of each non-zero entry of the skew-symmetric matrix $\boldsymbol{S}_{\boldsymbol{x}_i}$, say $s_{jk}(\boldsymbol{x}_i)$, with $(j,k) \in (1,\ldots,d) \times (1,\ldots,d)$ (with $j \neq k$), has p continuous derivatives in a neighbourhood of $\boldsymbol{x} \in \mathbb{S}^{d-1}$, expansion (6), for $s_{jk}(\boldsymbol{x}_i)$ around $\boldsymbol{x}$, yields

$$\boldsymbol{R}_{\boldsymbol{x}_i} = \exp(\boldsymbol{S}_{\boldsymbol{x}_i}) \approx \exp\left(\boldsymbol{S}_{\boldsymbol{x}} + \sum_{\ell=1}^{p} \mathsf{D}_{\boldsymbol{S}_{\boldsymbol{x}}}^{(\ell)}(\boldsymbol{x}_i, \boldsymbol{x})\right), \tag{7}$$

where $\mathsf{D}_{\boldsymbol{S}_{\boldsymbol{x}}}^{(\ell)}(\boldsymbol{x}_i, \boldsymbol{x})$ is the matrix of order d having $\theta_i^\ell/(\ell!)\mathcal{D}_{\boldsymbol{\xi}_i}^{(\ell)}\bar{s}_{jk}(\boldsymbol{x})$ as its (j,k)th entry. A further approximation, which uses the expansion of the matrix exponential, yields

$$\boldsymbol{R}_{\boldsymbol{x}_i} \approx \boldsymbol{R}_{\boldsymbol{x}}\left(\boldsymbol{I}_d + \sum_{\ell=1}^{p} \mathsf{D}_{\boldsymbol{S}_{\boldsymbol{x}}}^{(\ell)}(\boldsymbol{x}_i, \boldsymbol{x})\right). \tag{8}$$

Now, if in model (3) we approximate $\boldsymbol{R}_{\boldsymbol{x}_i}$ by (7), or, equivalently, by (8), we see that — due to the local character of the expansion — $\boldsymbol{R}_{\boldsymbol{x}}$ can be approximated by *all* the n rotations $\boldsymbol{R}_{\boldsymbol{x}_i}$, $i \in (1, \ldots, n)$, more accurately as the locations $\boldsymbol{x}_i$ are closer to $\boldsymbol{x}$. Consequently, we attain a tractable estimation problem, in the sense that each observation $\boldsymbol{x}_i$ can participate in the estimation of the rotation at $\boldsymbol{x}$ with the caveat that its contribution needs to be as smaller for larger $||\boldsymbol{x}_i - \boldsymbol{x}||$. Indeed, we have defined a class of estimators, depending on the number of terms, p, we choose in the expansion of $\boldsymbol{R}_{\boldsymbol{x}_i}$. In the next two sections we will examine the main cases.

3. ONE-TERM FIT

3.1 Estimator

A single term ($p = 0$) version of expansion (8) yields $\boldsymbol{R}_{\boldsymbol{x}_i} \approx \boldsymbol{R}_{\boldsymbol{x}}$, and so our estimator is given by the solution of the locally weighted least squares problem

$$\underset{\boldsymbol{R}_{\boldsymbol{x}} \in \mathrm{SO}(d)}{\operatorname{argmin}} \sum_{i=1}^{n} ||\boldsymbol{y}_i - \boldsymbol{R}_{\boldsymbol{x}} \boldsymbol{x}_i||^2 K_\kappa \left(\boldsymbol{x}_i^T \boldsymbol{x} \right), \tag{9}$$

where the weight function $K_\kappa(\boldsymbol{x}_i^T \boldsymbol{x})$ — often referred to as *kernel* — is chosen to reflect the geodesic distance from $\boldsymbol{x}_i$ to $\boldsymbol{x}$ rescaled by the *concentration* parameter $\kappa > 0$. To roughly understand the rôle of κ, suffice it to say that $1/\kappa$ is proportional to the width of the neighbourhood of $\boldsymbol{x}$ containing the observations effectively involved in the estimation process. As a weight function we could use $K_k(\boldsymbol{x}_i^T \boldsymbol{x}) \propto \exp\left(\kappa \boldsymbol{x}_i^T \boldsymbol{x}\right)$. This is a rotationally symmetric function with maximum at $\boldsymbol{x}$, and a parameter κ which governs how much the weight spreads around $\boldsymbol{x}$. As a result, $\boldsymbol{x}_i$ will receive a bigger weight the closer it is to $\boldsymbol{x}$, and for larger κ. Clearly, because model (3) assigns a rotation to each point, in order to reduce the bias we require κ to be sufficiently large, although this would involve a smaller effective sample size and therefore variance inflation. Additionally, given any finite dataset, we could always select a large enough κ that guarantees perfect superimposition of predictions and observed responses, and therefore complete adaptation. In the presence of experimental errors, this extreme scenario would perform very poorly on an out-of-sample test set of data. Taking all this into account, a natural strategy would set κ as increasing with n. Importantly, observe that such a rule will result in an arbitrarily accurate approximation (7), depending on n.

Rosenthal et al. (2017) use a roughness penalty as a bias-variance trade-off technique. This is equivalent to our smoothing parameter approach. As a *global* modelling strategy, their method has obviously a greater capacity to extrapolate to sparse regions, provided that the regression function belongs to the space of diffeomorphisms. *Local* methods, like ours, could in these cases still reasonably work if the smoothing parameter is not chosen by cross-validation and is not kept fixed for all locations. Surely, flexibility due to locality needs to be carefully managed, but has less chance of incurring model misspecification errors. In fact, many analysts would prefer not to carry out estimation in regions of no data. Our asymptotic theory, that we will present later to describe accuracy, clearly depends on the density of the observations, and requires this to be non-zero. Rosenthal et al. (2017) do not provide asymptotic properties.

Finally note that, whatever the value of d is, we have always a scalar concentration parameter κ. This contrasts with Euclidean theory where we may have a smoothing matrix of order d allowing specific smoothing degrees along the directions. Unfortunately, current distributions on the sphere do not allow such kind of flexibility when used as weights. Therefore directional smoothing on the sphere could be regarded an active research field. On the other hand, using a specific concentration for each observation (or estimation point), would constitute a straight extension of Euclidean theory.

Remark 2. *Optimization (9) could be considered a generalization of a problem posed by Wahba (1965). This latter was formulated for* $d = 3$ *in a non-stochastic framework, where* $\boldsymbol{x}_i$ *represents the known coordinates of a star,* $\boldsymbol{y}_i$ *refers to the position registered by a satellite, and the weight for the* i*th observation reproduces the precision of the experimental measurement* $\boldsymbol{x}_i$*. Importantly, in our case* κ *is not a feature of the observed phenomenon, but rather a parameter chosen by the statistician. The focus has traditionally been on the quality of superimposing algorithm, and SVD was seen to be one of the most robust, though not the fastest.*

Likewise for (1), we have $\hat{\boldsymbol{R}}_{\boldsymbol{x}} = \boldsymbol{U}^T \boldsymbol{\Delta} \boldsymbol{V}$, in which $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ is the SVD of $\boldsymbol{Y}^T \boldsymbol{W}_\kappa(\boldsymbol{x})\boldsymbol{X}$, with $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{\Delta}$ as before and, for $\boldsymbol{a} \in \mathbb{S}^{d-1}$, $\boldsymbol{W}_\kappa(\boldsymbol{a})$ is a diagonal matrix of order n having $K_k\left(\boldsymbol{x}_i^T \boldsymbol{a}\right)$ as its (i, i)th entry. Notice that, due to the non-linearity of the constraints required by a proper rotation, we do not have a closed-form estimator and this will make it difficult to establish statistical properties.

It is possible to compare this estimator with the approach of Di Marzio et al. (2014). They propose a nonparametric, component-wise spherical regression fit. Due to the localization of the estimator, spatial weights are also involved there. Exploring the limiting cases of κ, with sample size kept as fixed, indicates some differences and similarities. As $\kappa \to 0$, i.e. nearly zero concentration, the weights tend to have same value for all estimation points. Clearly the inference changes its nature into a parametric one because $\hat{\boldsymbol{R}}_{\boldsymbol{x}} \to \hat{\boldsymbol{R}}$, the rigid rotation solution. This is relevant because our method, although very flexible, still has the potential to include the basic parametric spherical regression model, i.e. the rigid rotation. On the other hand, using the same weight within Di Marzio et al. (2014) procedure for all observations produces the interpolating surface used for the expansion, a very poor result. Conversely, as $\kappa \to \infty$ for both methods the region containing the observations really participating to the estimation will become too small with the result of involving in the estimate one or no data depending on we are estimating either at $\boldsymbol{x}_i$ or at $\boldsymbol{x}$.

As for the smoothness degree selection, we see that, using a standard cross-validation (CV) procedure, the optimal choice of κ minimizes $-\sum_i \boldsymbol{y}_i^T \hat{\boldsymbol{y}}_i^{(-i)}$, where $\hat{\boldsymbol{y}}_i^{(-i)} = \hat{\boldsymbol{R}}_{\boldsymbol{x}_i}^{(-i)} \boldsymbol{x}_i$, and $\hat{\boldsymbol{R}}_{\boldsymbol{x}_i}^{(-i)}$ is estimated using the SVD but after removing the ith rows from the matrices $\boldsymbol{X}, \boldsymbol{Y}$, and the ith row and column from the matrix $\boldsymbol{W}_\kappa(\boldsymbol{x})$. Note that the CV function requires n SVD solutions to be found. Having obtained a suitable κ, this then defines function K_κ in Equation (9) with the proposed solution giving a predicted value $\hat{\boldsymbol{y}} = \hat{\boldsymbol{R}}_{\boldsymbol{x}}\boldsymbol{x}$.

3.2 Asymptotic properties

As observed, we do not have a closed-form estimator, and therefore we are unable to use a "direct" approach to derive asymptotic properties. Importantly, it will emerge that the estimator resulting from the search over $\mathbb{R}^{d\times d}$, rather than over $\mathrm{SO}(d)$ — which we refer to as the *unconstrained* estimator — is consistent, which means that constraints are asymptotically inactive within our rotation model. This *per se* would make it sufficient to simply explore the unconstrained solution. However, to be more detailed, we can still take into account, as an approximation, the solution of a least square problem with a linear constraint. In fact we could use the linear truncation $\boldsymbol{R}_{\boldsymbol{x}} = \exp(\boldsymbol{S}_{\boldsymbol{x}}) \approx \boldsymbol{I}_d + \boldsymbol{S}_{\boldsymbol{x}}$. Such a truncation is more precise for smaller rotation angles. However, this assumption is not very restrictive because our model requires that rotations are similar within a region centred on the estimation point. In fact, a smoothness hypothesis motivates a preliminary estimation of a pale rigid rotation in order to move the predictor values close to their respective responses.

Specifically, let $\mathrm{Skew}_q := \left\{\boldsymbol{M} \in \mathbb{R}^{q\times q} : \boldsymbol{M}^T = -\boldsymbol{M}\right\}$, and recall that the Frobenius norm of a matrix $\boldsymbol{A}$ is $||\boldsymbol{A}||_F := (\mathrm{trace}(\boldsymbol{A}^T\boldsymbol{A}))^{1/2}$. Then, letting $\boldsymbol{A}$ and $\boldsymbol{B}$ be $p \times q$ matrices, and

$$\tilde{\boldsymbol{C}} := \underset{\boldsymbol{C}\in\mathbb{R}^{q\times q}}{\mathrm{argmin}} ||\boldsymbol{A} - \boldsymbol{B}\boldsymbol{C}||_F^2,$$

we have that the solution of this least squares problem under a skew-symmetric constraint is

$$\underset{\boldsymbol{C}\in\,\mathrm{Skew}_q}{\mathrm{argmin}}\; ||\boldsymbol{A} - \boldsymbol{B}\boldsymbol{C}||_F^2 = \frac{1}{2}\left(\tilde{\boldsymbol{C}} - \tilde{\boldsymbol{C}}^T\right),$$

which is the *skew-symmetric part* of the unconstrained solution. Now, letting $\boldsymbol{X}$ and $\boldsymbol{Y}$ be defined as before, the solution of the least squares problem

$$\underset{\boldsymbol{S}_{\boldsymbol{x}}\in\,\mathrm{Skew}_d}{\mathrm{argmin}}\; ||\boldsymbol{W}_\kappa(\boldsymbol{x})^{1/2}\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{I}_d + \boldsymbol{S}_{\boldsymbol{x}})\}||_F^2$$

yields a *locally constant* estimator of $\boldsymbol{S}_{\boldsymbol{x}}$. Specifically, the above argument with $\boldsymbol{A} = \boldsymbol{W}_\kappa(\boldsymbol{x})^{1/2}(\boldsymbol{Y} - \boldsymbol{X})$, and $\boldsymbol{B} = \boldsymbol{W}_\kappa(\boldsymbol{x})^{1/2}\boldsymbol{X}$, yields

$$\hat{\boldsymbol{S}}_{\boldsymbol{x}} = \frac{1}{2}\left\{(\boldsymbol{X}^T\boldsymbol{W}_\kappa(\boldsymbol{x})\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_\kappa(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{W}_\kappa(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}_\kappa(\boldsymbol{x})\boldsymbol{X})^{-1}\right\}. \tag{10}$$

Now, for a given weight function K_κ, and $j \in \mathbb{N}$, set

$$b_j(\kappa) := \omega_{d-2}\int_0^\pi K_\kappa(\cos(\theta))\theta^j \sin^{d-2}(\theta)\mathrm{d}\theta, \quad \text{and} \quad \nu(\kappa) := \omega_{d-2}\int_0^\pi K_\kappa^2(\cos(\theta))\sin^{d-2}(\theta)\mathrm{d}\theta,$$

where $\omega_d := \mu_d\left(\mathbb{S}^d\right) = 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$ is the surface area of $\mathbb{S}^d$. Now, let u_ℓ be the ℓth entry of a d-dimensional vector $\boldsymbol{u}$, and for $j \in (1,\ldots,d)$, let $\boldsymbol{D}_j(\boldsymbol{x})$ be a matrix of order d having $\partial\bar{s}_{ik}(\boldsymbol{u})/\partial u_j\;|_{\boldsymbol{u}=\boldsymbol{x}}$

as its (i,k)th entry, and, for $(j,\ell) \in (1,\dots,d) \times (1,\dots,d)$, let $\boldsymbol{H}_{j\ell}(\boldsymbol{x})$ be the matrix of order d having $\partial^2 \bar{s}_{ik}(\boldsymbol{u})/(\partial u_j \partial u_\ell) \mid_{\boldsymbol{u}=\boldsymbol{x}}$ as its (i,k)th entry. Also, define the block matrices

$$\boldsymbol{P}(\boldsymbol{x}) := \begin{pmatrix} \boldsymbol{D}_1(\boldsymbol{x}) \\ \vdots \\ \boldsymbol{D}_d(\boldsymbol{x}) \end{pmatrix}, \quad \text{and} \quad \boldsymbol{Q}(\boldsymbol{x}) := \begin{pmatrix} \boldsymbol{H}_{11}(\boldsymbol{x}) \\ \vdots \\ \boldsymbol{H}_{1d}(\boldsymbol{x}) \\ \vdots \\ \boldsymbol{H}_{d1}(\boldsymbol{x}) \\ \vdots \\ \boldsymbol{H}_{dd}(\boldsymbol{x}) \end{pmatrix}.$$

Now, let $\text{vec}(\boldsymbol{A})$ denote the vectorization of matrix $\boldsymbol{A}$. If $\boldsymbol{A} \otimes \boldsymbol{B}$ indicates the Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and f is the common density of the $\boldsymbol{x}_i$s, we get

Theorem 1. *Given the $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$-valued random sample $(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$, assume model* (3). *If,*

i) $f(\boldsymbol{x}) > 0$, the derivatives of $\bar{f}$ up to order 1, and the derivatives of $\bar{s}_{ik}$ up to order 2, are continuous at $\boldsymbol{x} \in \mathbb{S}^{d-1}$, with $(i,k) \in (1,\dots,d) \times (1,\dots,d)$ and $i \neq k$,

ii) $\lim_{n\to\infty} b_2(\kappa) = 0$,

iii) $b_j(\kappa) = o(b_2(\kappa))$, for each $j > 2$,

then, for estimator (10), it holds that

$$\mathsf{E}[\hat{\boldsymbol{S}}_{\boldsymbol{x}} - \boldsymbol{S}_{\boldsymbol{x}} \,|\, \boldsymbol{X}] = \frac{b_2(\kappa)}{2f(\boldsymbol{x})} \left\{ 2(d-1)(\nabla^T_{\bar{f}}(\boldsymbol{x}) \otimes \boldsymbol{I}_d)\boldsymbol{P}(\boldsymbol{x}) + f(\boldsymbol{x}) \left(\int_{\mathbb{T}_{\boldsymbol{x}}} \text{vec}^T(\boldsymbol{\xi}\boldsymbol{\xi}^T) \mathrm{d}\mu_{d-2}(\boldsymbol{\xi}) \otimes \boldsymbol{I}_d \right) \boldsymbol{Q}(\boldsymbol{x}) \right\} + o(b_2(\kappa)\boldsymbol{I}_d).$$

Proof. See supplementary material. □

4. TWO-TERM FIT

4.1 Estimator

A two-term formulation could be appropriate in some cases. This assumes that, in a neighbourhood of $\boldsymbol{x}$, $\boldsymbol{R}_{\boldsymbol{x}_i}$ is adequately approximated by (8) with $p = 1$. This results in solving

$$\underset{\boldsymbol{R}_{\boldsymbol{x}} \in \mathrm{SO}(d), \mathsf{D}^{(1)}_{\boldsymbol{S}_{\boldsymbol{x}}}(\boldsymbol{x}_i, \boldsymbol{x}) \in \mathrm{Skew}_d}{\text{argmin}} \sum_{i=1}^{n} \left|\left| \boldsymbol{y}_i - \boldsymbol{R}_{\boldsymbol{x}} \left\{ \boldsymbol{I}_d + \mathsf{D}^{(1)}_{\boldsymbol{S}_{\boldsymbol{x}}}(\boldsymbol{x}_i, \boldsymbol{x}) \right\} \boldsymbol{x}_i \right|\right|^2 K_\kappa\left(\boldsymbol{x}_i^T \boldsymbol{x}\right), \tag{11}$$

for which — distinctly from the case $p = 0$ where the SVD is used — we are not aware of any closed form solution. Some computational comments for numerical methods are considered in Section 6.1.

4.2 Asymptotic properties

Extending the reasoning of Section 3.2 to $p = 1$, we consider the approximation $\boldsymbol{R}_{\boldsymbol{x}_i} \approx \boldsymbol{I}_d + \boldsymbol{S}_{\boldsymbol{x}} + \mathsf{D}^{(1)}_{\boldsymbol{S}_{\boldsymbol{x}}}(\boldsymbol{x}_i, \boldsymbol{x})$. Then, letting $\mathbb{X}$ and $\boldsymbol{\beta}$ respectively be $nd \times (d+1)d$ and $(d+1)d \times d$ matrices defined as

$$\mathbb{X} = \begin{pmatrix} 1 & \theta_1 \boldsymbol{\xi}_1^T \\ \vdots & \vdots \\ 1 & \theta_n \boldsymbol{\xi}_n^T \end{pmatrix} \otimes \boldsymbol{I}_d, \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{S}_{\boldsymbol{x}} \\ \boldsymbol{D}_1(\boldsymbol{x}) \\ \vdots \\ \boldsymbol{D}_d(\boldsymbol{x}) \end{pmatrix},$$

and setting $\tilde{\mathbb{X}} = \text{diag}(\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)\mathbb{X}$, according to the above expansion, we have $\mathsf{E}[\boldsymbol{Y} \mid \boldsymbol{X}] \approx \boldsymbol{X} + \tilde{\mathbb{X}}\boldsymbol{\beta}$, and the solution for $\boldsymbol{S}_{\boldsymbol{x}}$ of the locally weighted least squares problem

$$\underset{\boldsymbol{S}_{\boldsymbol{x}} \in \, \mathrm{Skew_d}, \boldsymbol{D}_1(\boldsymbol{x}) \in \, \mathrm{Skew_d}, \cdots, \boldsymbol{D}_d(\boldsymbol{x}) \in \, \mathrm{Skew_d}}{\operatorname{argmin}} \quad ||\boldsymbol{W}_\kappa(\boldsymbol{x})^{1/2}(\boldsymbol{Y} - \boldsymbol{X} - \tilde{\mathbb{X}}\boldsymbol{\beta})||_F^2, \tag{12}$$

defines a two-term estimator of $\boldsymbol{S}_{\boldsymbol{x}}$. In particular, reasoning as before, we can write the solutions of problem (12) as the skew-symmetric parts of the unconstrained solutions. Hence, the solution for $\boldsymbol{S}_{\boldsymbol{x}}$ is

$$\hat{\boldsymbol{S}}_{\boldsymbol{x}} = \frac{1}{2}\left\{ \boldsymbol{E}_1^T \left(\tilde{\mathbb{X}}^T \boldsymbol{W}_\kappa(\boldsymbol{x})\tilde{\mathbb{X}}\right)^{-1} \tilde{\mathbb{X}}^T \boldsymbol{W}_\kappa(\boldsymbol{x})(\boldsymbol{Y} - \boldsymbol{X}) - (\boldsymbol{Y} - \boldsymbol{X})^T \boldsymbol{W}_\kappa(\boldsymbol{x})\tilde{\mathbb{X}} \left(\tilde{\mathbb{X}}^T \boldsymbol{W}_\kappa(\boldsymbol{x})\tilde{\mathbb{X}}\right)^{-1} \boldsymbol{E}_1 \right\}, \tag{13}$$

where $\boldsymbol{E}_1$ is a $(d+1)d \times d$ block matrix having $\boldsymbol{I}_d$ as its first block and null matrices otherwise. Concerning asymptotic properties, we get the following

Theorem 2. *Given the $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$-valued random sample $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$, under assumptions i)−iii) of Theorem 1, for estimator (13), it holds that*

$$\mathsf{E}[\hat{\boldsymbol{S}}_{\boldsymbol{x}} - \boldsymbol{S}_{\boldsymbol{x}} \,|\, \boldsymbol{X}] = \frac{b_2(K_\kappa)}{2} \left(\int_{\mathbb{T}_{\boldsymbol{x}}} \text{vec}^T(\boldsymbol{\xi}\boldsymbol{\xi}^T) \mathrm{d}\mu_{d-2}(\boldsymbol{\xi}) \otimes \boldsymbol{I}_d \right) \boldsymbol{Q}(\boldsymbol{x}) + o\left(b_2(\kappa)\boldsymbol{I}_d\right)$$

Proof. See supplementary material. □

Comparing this result with Theorem 1, it emerges that the asymptotic bias of the two-term version of $\hat{\boldsymbol{S}}_{\boldsymbol{x}}$, differently from the one-term one, does not depend on the design density f, nor on the first order derivatives of the homogeneous extensions of the entries of $\boldsymbol{S}_{\boldsymbol{x}}$. However, the estimators share the order of the asymptotic bias. As for the asymptotic variance, for both the one and two-term versions of $\hat{\boldsymbol{S}}_{\boldsymbol{x}}$, we get

Theorem 3. *Given the $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$-valued random sample $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$, if*

i) $f(\boldsymbol{x}) > 0$, and all entries of $\boldsymbol{\Sigma}_{\boldsymbol{x}_i}$, $i \in (1, \ldots, n)$, are continuous at $\boldsymbol{x} \in \mathbb{S}^{d-1}$,

ii) the weight function K_κ is such that $\lim_{n\to\infty} n^{-1}\nu(\kappa) = 0$,

then the asymptotic variances of the vectorizations of estimators (10) *and* (13) *are both $O(n^{-1}\nu(\kappa)\boldsymbol{I}_{d^2})$.*

Proof. See supplementary material. □

As a consequence of Theorems 1–3, the asymptotic biases of the vectorizations of estimators (10) and (13) are both $O(b_2(\kappa)\mathbf{1}_{d^2})$, with $\mathbf{1}_d$ being a d-dimensional vector of ones, while the asymptotic variances are both $O(n^{-1}\nu(\kappa)\boldsymbol{I}_{d^2})$. Now, consider the special case of a von Mises-Fisher kernel, which can be regarded as the spherical counterpart of the Gaussian kernel, and is defined, on $\mathbb{S}^{d-1}$, as

$$K_\kappa(\boldsymbol{x}^T\boldsymbol{\mu}) = \frac{\kappa^{d/2-1}e^{\kappa\boldsymbol{x}^T\boldsymbol{\mu}}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa)},$$

where $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is the mean direction, $\kappa > 0$ is the concentration parameter, and $\mathcal{I}_u(\cdot)$ stands for the modified Bessel function of the first kind and order u. For such a kernel, and for κ big enough, it holds that

$$b_j(\kappa) \sim \frac{2^{j/2}\Gamma\left((d+j-1)/2\right)}{\kappa^{j/2}\Gamma\left((d-1)/2\right)}, \qquad \text{and} \qquad \nu(\kappa) \sim \frac{\kappa^{(d-1)/2}}{2^{d-1}\pi^{(d-1)/2}},$$

hence it satisfies condition *iii)* of Theorem 1, whereas assumption *ii)* of Theorem 1 and assumption *ii)* of Theorem 3 respectively imply that, as $n \to \infty$, $\kappa \to \infty$, and $n^{-1}\kappa^{(d-1)/2} \to 0$. Then, using the above kernel as the weight function K_κ, for both (10) and (13), $\text{vec}(\hat{\boldsymbol{S}}_{\boldsymbol{x}})$ has conditional asymptotic bias of order $O(\kappa^{-1}\mathbf{1}_{d^2})$, and conditional asymptotic variance of order $O(n^{-1}\kappa^{(d-1)/2}\boldsymbol{I}_{d^2})$. Thus the *optimal* value of κ, which minimizes the conditional *asymptotic mean squared error* of $\text{vec}(\hat{\boldsymbol{S}}_{\boldsymbol{x}})$, being the leading part of

$$\mathsf{E}[||\text{vec}(\hat{\boldsymbol{S}}_{\boldsymbol{x}}) - \text{vec}(\boldsymbol{S}_{\boldsymbol{x}})||^2 \mid \boldsymbol{X}] = \text{trace}(\mathsf{Var}[\hat{\boldsymbol{S}}_{\boldsymbol{x}} \mid \boldsymbol{X}]) + ||\mathsf{E}[\text{vec}(\hat{\boldsymbol{S}}_{\boldsymbol{x}}) - \text{vec}(\boldsymbol{S}_{\boldsymbol{x}}) \mid \boldsymbol{X}]||^2,$$

is $O(n^{2/(d+3)})$. The resulting convergence rate of $\text{vec}(\hat{\boldsymbol{S}}_{\boldsymbol{x}})$, given by the order of the asymptotic mean squared error using the optimal κ, is $O(n^{-4/(d+3)})$.

5. AN ITERATIVE ALGORITHM FOR BIAS REDUCTION

5.1 Motivation

After obtaining $\hat{\boldsymbol{y}}_i$s, a natural question is whether a further rotation of the predicted values could improve the fit. As a motivation, observe that in such an iteration we could use different weights. Also note, since our method gives non-rigid rotations, then the interpoint distances of the fitted values will differ from those of the original covariates. In general, we could perform several iterations such that, at each step, the current estimate of $\boldsymbol{y}_i$ is used in place of $\boldsymbol{x}_i$. As a sort of residuals fitting technique, this idea clearly has a relationship with L_2-*boosting*, which has been used with splines, and kernel methods in regression; see Bühlmann and Yu (2003) and Di Marzio and Taylor (2008), respectively.

Algorithm 1, which can easily be adjusted to make predictions for several new design points, implements the proposed method. Moreover, note that, if $\kappa = 0$ (line 1), we would obtain the rigid solution, and taking further iterations ($M > 1$) will have no effect. Further, several variants of this algorithm could be investigated. For example, κ could be allowed to change along iterations, or the weight matrix could be

```
Require: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $n \times d$ data matrices
    input:
 1: $\kappa \leftarrow$ smoothing parameter
 2: $M \leftarrow$ number of iterations
 3: given $\boldsymbol{x} \in \mathbb{S}^{d-1}$
    initialize:
 4: $\hat{\boldsymbol{R}}_{\boldsymbol{x}} \leftarrow \boldsymbol{I}_d$
 5: $\boldsymbol{W} \leftarrow \boldsymbol{W}_\kappa(\boldsymbol{x})$
 6: $\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{X}$
    loop:
 7: for $m \leftarrow 1$ to $M$ do
 8:     factorize: $\boldsymbol{Y}^T\boldsymbol{W}\hat{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$
 9:     $\boldsymbol{R} \leftarrow \boldsymbol{U}^T\boldsymbol{V}$
10:     $\hat{\boldsymbol{Y}} \leftarrow \hat{\boldsymbol{Y}}\boldsymbol{R}^T$
11:     $\hat{\boldsymbol{R}}_{\boldsymbol{x}} \leftarrow \boldsymbol{R}\hat{\boldsymbol{R}}_{\boldsymbol{x}}$
    output:
12: $\hat{\boldsymbol{y}} \leftarrow \hat{\boldsymbol{R}}_{\boldsymbol{x}}\boldsymbol{x}$
```

algorithm 1: Newton-Raphson iterative rotation fitting.

updated (after line 10) to correspond to the current $\hat{\boldsymbol{Y}}$, or — as in the case of some boosting algorithms — to depend on the current residuals. In any case, κ and M will need to be jointly chosen, with larger values of M corresponding to smaller κ in order to obtain perfect superimposition for very large values of M.

Noting that Algorithm 1 is computationally intensive — particularly if leave-one-out cross-validation is used to select κ — and considering analogous situations for boosting kernel regression in the Euclidean case, we also investigate an approximation in which the weight function is adjusted to mimic the solution for further iterations. So, for $M = 2$, simple algebra leads to define the ith kernel as

$$2K_\kappa\left(\boldsymbol{x}_i^T\boldsymbol{x}\right) - K_{\kappa/4}\left(\boldsymbol{x}_i^T\boldsymbol{x}\right), \tag{14}$$

without any iterations. Such kernel structure is reminiscent of the *twicing* estimator proposed by Stuetzle & Mittal (1979) in the context of robust, nonparametric estimation. It should be noted that this weight is not necessarily positive, and that, in this case, the normalizing constant for the kernel will be required.

We now show that Algorithm 1 iteratively estimates the bias, and then corrects for it. We will also see that it has an optimal (Newton-Raphson) convergence rate to the local minimum.

5.2 Some properties

To keep things simple consider a one-term version of (8) for $\boldsymbol{R}_{\boldsymbol{x}_i}$, i.e. $\boldsymbol{R}_{\boldsymbol{x}_i} \approx \boldsymbol{R}_{\boldsymbol{x}}$. The one-step problem requires minimizing the function

$$\mathsf{F}(\boldsymbol{R}, \boldsymbol{L}, \lambda) := ||\boldsymbol{W}_{\kappa}(\boldsymbol{x})^{1/2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{R})||_F^2 - \text{trace}\left(\frac{1}{2}\boldsymbol{L}(\boldsymbol{R}^T\boldsymbol{R} - \boldsymbol{I})\right) - \lambda(|\boldsymbol{R}| - 1),$$

where $\boldsymbol{L}$ is a diagonal matrix of order d of Lagrange multipliers and λ is also a Lagrange multiplier. Partial differentiation of F with respect to $\boldsymbol{R}$, $\boldsymbol{L}$ and λ leads to the following system of equations

$$\begin{aligned}
\frac{\partial \mathsf{F}(\boldsymbol{R}, \boldsymbol{L}, \lambda)}{\partial \boldsymbol{R}} &= -2\boldsymbol{Y}^T\boldsymbol{W}_{\kappa}(\boldsymbol{x})\boldsymbol{X} + 2\boldsymbol{R}\boldsymbol{X}^T\boldsymbol{W}_{\kappa}(\boldsymbol{x})\boldsymbol{X} - \boldsymbol{R}\boldsymbol{L} - \lambda\boldsymbol{R} = \mathbf{0}_{d\times d}\\
\frac{\partial \mathsf{F}(\boldsymbol{R}, \boldsymbol{L}, \lambda)}{\partial \boldsymbol{L}} &= \frac{1}{2}(\boldsymbol{R}^T\boldsymbol{R} - \boldsymbol{I}) = \mathbf{0}_{d\times d}\\
\frac{\partial \mathsf{F}(\boldsymbol{R}, \boldsymbol{L}, \lambda)}{\partial \lambda} &= |\boldsymbol{R}| - 1 = 0,
\end{aligned}$$

where $\mathbf{0}_{d\times d}$ is the null matrix of order d. In order to obtain some properties, according to the Fisher's scoring idea (see, for example, Osborne (1992)), we expand the gradient of the objective function taken at the SVD solution $\hat{\boldsymbol{R}}_{\boldsymbol{x}}$ around the unknown $\boldsymbol{R}_{\boldsymbol{x}}$. This will lead to an expression for the unknown estimation error $\hat{\boldsymbol{R}}_{\boldsymbol{x}} - \boldsymbol{R}_{\boldsymbol{x}}$. Specifically, letting $\mathsf{f}(\boldsymbol{R}) := \partial\mathsf{F}(\boldsymbol{R}, \boldsymbol{L}, \lambda)/\partial\boldsymbol{R}$, by expanding $\mathsf{f}(\hat{\boldsymbol{R}}_{\boldsymbol{x}})$ around $\boldsymbol{R}_{\boldsymbol{x}}$, we obtain $\mathsf{f}(\hat{\boldsymbol{R}}_{\boldsymbol{x}}) = \mathsf{f}(\boldsymbol{R}_{\boldsymbol{x}}) + \mathsf{f}'(\boldsymbol{R}_{\boldsymbol{x}})(\hat{\boldsymbol{R}}_{\boldsymbol{x}} - \boldsymbol{R}_{\boldsymbol{x}}) = \mathbf{0}_{d\times d}$, where $\mathsf{f}'(\boldsymbol{R}_{\boldsymbol{x}}) := \partial\mathsf{f}(\boldsymbol{R})/\partial\boldsymbol{R}|_{\boldsymbol{R}=\boldsymbol{R}_{\boldsymbol{x}}}$. Notice that this equality is due to the linearity of first derivative. Consequently, we get

$$\hat{\boldsymbol{R}}_{\boldsymbol{x}} - \boldsymbol{R}_{\boldsymbol{x}} = -\mathsf{f}'(\boldsymbol{R}_{\boldsymbol{x}})^{-1}\mathsf{f}(\boldsymbol{R}_{\boldsymbol{x}}). \tag{15}$$

Now, the objective function, at step $m \geq 1$, is

$$\mathsf{F}_m(\boldsymbol{R}, \boldsymbol{L}, \lambda) := \left|\left|\boldsymbol{W}_{\kappa,m}(\boldsymbol{x})^{1/2}\left\{\boldsymbol{Y} - \left(\boldsymbol{X}\prod_{\ell=0}^{m-1}\hat{\boldsymbol{R}}_{\boldsymbol{x},\ell}\right)\boldsymbol{R}\right\}\right|\right|_F^2 - \text{trace}\left(\frac{1}{2}\boldsymbol{L}\left(\boldsymbol{R}^T\boldsymbol{R} - \boldsymbol{I}\right)\right) - \lambda(|\boldsymbol{R}| - 1),$$

where $\hat{\boldsymbol{R}}_{\boldsymbol{x},0} = \boldsymbol{I}_d$, $\hat{\boldsymbol{R}}_{\boldsymbol{x},\ell}$ is the rotation matrix estimated at the ℓth step, and $\boldsymbol{W}_{\kappa,m}(\boldsymbol{x})$ is the weight matrix used at the mth step. The resulting rotation matrix, using obvious notation, is

$$\hat{\boldsymbol{R}}_{\boldsymbol{x},m} = \hat{\boldsymbol{R}}_{\boldsymbol{x},m-1} - \mathsf{f}'_m(\hat{\boldsymbol{R}}_{\boldsymbol{x},m-1})^{-1}\mathsf{f}_m(\hat{\boldsymbol{R}}_{\boldsymbol{x},m-1}).$$

Such a solution resembles an iteration of a Newton-Raphson minimization algorithm. Indeed, it also coincides with an iteration in the gradient descent algorithm, where the step length is equal to $\mathsf{f}'_m(\hat{\boldsymbol{R}}_{\boldsymbol{x},m-1})^{-1}$. A statistical interpretation of the algorithm is possible in connection with Equation (15). In fact, it says that $\mathsf{f}'_m(\hat{\boldsymbol{R}}_{\boldsymbol{x},m-1})^{-1}\mathsf{f}_m(\hat{\boldsymbol{R}}_{\boldsymbol{x},m-1})$ is an estimate of the bias matrix, therefore the algorithm iteratively estimates the bias and corrects for it.

Regarding the asymptotic accuracy, we observe that theory contained in Section 3.2 (or 4.2 if we use the two-term fit as the base learner) provides an improving description of it at each step because the algorithm will estimate progressively smaller rotations.

6. SIMULATIONS

6.1 Computational Comments

To simulate data, a model of the form (2) can be specified either by providing $\boldsymbol{R}_{\boldsymbol{x}}$ or by specifying a skew-symmetric matrix. In the latter case, given $\boldsymbol{S}_{\boldsymbol{x}}$, we obtained $\boldsymbol{R}_{\boldsymbol{x}} = \exp(\boldsymbol{S}_{\boldsymbol{x}})$ from a truncated (after the second term) summation of the exponential series such that computed value satisfies the condition $\max(||\boldsymbol{R}_{\boldsymbol{x}}| - 1|, ||\boldsymbol{R}_{\boldsymbol{x}}^T - \boldsymbol{R}_{\boldsymbol{x}}^{-1}||_F^2) < 10^{-7}$.

The one-term solution is straightforward to obtain, since the rotation matrix is available from the SVD. This was used as a starting value in the two-term solution of (11), for which we resorted to a nonlinear optimization in the components of the 4 skew-symmetric matrices (i.e. 12 parameters) for 3-d data. Use of the skew-symmetric matrix seemed well-behaved in the Newton-Raphson optimization, and no constraints were required — as they would have been for a direct rotation solution. For $d = 3$ it is then easy to obtain $\hat{\boldsymbol{R}}_{\boldsymbol{x}}$ from $\hat{\boldsymbol{S}}_{\boldsymbol{x}}$ using the Rodrigues rotation formula. Although we found the two-term solution to be well-behaved, even for small n, it was very slow.

For the two-term fit, and the one-term fit (any M in Algorithm 1) the kernel function used for the weights was the von-Mises Fisher density. However, since normalization of the weights makes no difference to the SVD solution, we simply used $K_\kappa(\boldsymbol{x}_i^T\boldsymbol{x}) = \exp(\kappa(\boldsymbol{x}_i^T\boldsymbol{x} - 1))$, which is numerically stable for all κ.

6.2 Models

Fixing $d = 3$ we consider models in which the explanatory variable $\boldsymbol{\mathcal{X}}$ is uniformly distributed, as well as limited support. The response variable $\boldsymbol{\mathcal{Y}}$ is simulated according to model (3), in which the errors $\boldsymbol{\varepsilon}_i$ come from a normal distribution with independent components all having zero mean and variance 0.10, although other error densities, symmetric around the null direction, would surely lead to similar results. The distribution of the random errors is less important due to the nonparametric character of our estimators. There are many choices for $\boldsymbol{R}_{\boldsymbol{x}_i}$, but those reported in Table 1 cover the situations in which: the two-term model should be optimal (Model 1); the rotation is far from rigid (Model 2); and a reflection-rotation is permitted (Model 3). A non-smooth model is discussed in the supplementary material as a counter-example. The first two models are specified by a skew-symmetric matrix $\boldsymbol{S}_{\boldsymbol{x}}$, for $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{S}^2$. For each

Model 1	Model 2	Model 3
$\boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{2}\begin{pmatrix} 0 & -x_1 & -x_2 \\ x_1 & 0 & -x_3 \\ x_2 & x_3 & 0 \end{pmatrix}$,	$\boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{2}\begin{pmatrix} 0 & -\mathrm{e}^{2x_1} & -\mathrm{e}^{2x_2} \\ \mathrm{e}^{2x_1} & 0 & -\mathrm{e}^{2x_3} \\ \mathrm{e}^{2x_2} & \mathrm{e}^{2x_3} & 0 \end{pmatrix}$,	$\boldsymbol{R} = \begin{pmatrix} -0.36 & 0.48 & -0.8 \\ 0.8 & 0.48 & 0 \\ -0.48 & 0.64 & 0.6 \end{pmatrix}$.

Table 1: Regression models used for simulations.

training dataset of size n, we simulate an independent test set of size N from the same model. Unless otherwise stated, the smoothing parameter is selected by cross-validation. The estimates are obtained using the training data, with a goodness of fit computed from the test set using $E = (Nd)^{-1} \sum_i ||\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i||^2$. When multiple simulations are performed, the average over all datasets of this error is reported. The methods investigated include (a) rigid rotation (Chang, 1986), which is a special case ($\kappa = 0$) of the (b) one-term fit; (c) an iterated solution, using Algorithm 1; (d) two-term fit; (e) one-term fit using weight (14); and (f) the spherical regression methods (both local constant and local linear) of Di Marzio et al. (2014).

6.3 Mean squared error and κ

In the first experiment we show the dependence on smoothing parameter; cross-validation is not used here, see the supplementary material for examples on this. We simulate 50 samples of size $n = 100$ from each model, with an independent test set of size $N = 100$. Figure 2 shows the average value of E for the standard weighted solution, and up to three further iterations, together with that of the two-term fit and one-term fit with weight function (14). Notice that using an error variance equal to 0.10 corresponds to $E = 0.066$ with perfect knowledge of the model, so this would be a lower bound on any method. As can be seen in Figure 2, the rigid solution (Chang, 1986), which corresponds to $\kappa = 0$, is clearly suboptimal. The method of Rosenthal et al. (2014) — who consider non-rotation transformations followed by a projection onto the sphere — does somewhat better on these models, with errors of $0.101, 0.367$ and 0.226 for Models 1–3, respectively. However, for suitable choice of smoothing parameter, the weighted solutions perform much better.

In common with other iterative schemes, the optimal smoothing parameter depends on the number of iterations M, and decreases with it. Also, to higher values of M correspond reduced errors, though by increasingly smaller amounts. This is in accordance with theory, where we saw that at each step the algorithm fits "residuals"of the previous one. The slightly odd behaviour for one-term fit using kernel (14) is due to the fact that negative weights result if κ is chosen too large, which causes some instability. The two-term solution performs very well on Model 1 — as may be expected — since, in this case, the first derivative of $\boldsymbol{S}_{\boldsymbol{x}}$ is constant, so this will capture all of the signal in the model. Finally, comparison of the first and fourth panels indicates a slight deterioration in the performance when the design points in the training set are restricted — particularly when using weight (14), but the average error is increased more for the rigid solution ($\kappa = 0$) than for the standard kernel solutions, both local constant and local linear. This last point is investigated further in the next subsection.

6.4 Bias-squared and variance

A residual analysis was also carried out in order to examine the bias and variance of the various solutions. In this case, we used a random design (over the whole sphere) for the training data (100 observations), but chose a regular grid of 104 test data points ($\boldsymbol{x}$) at which to predict. The box-plots in Figure 3 show how the

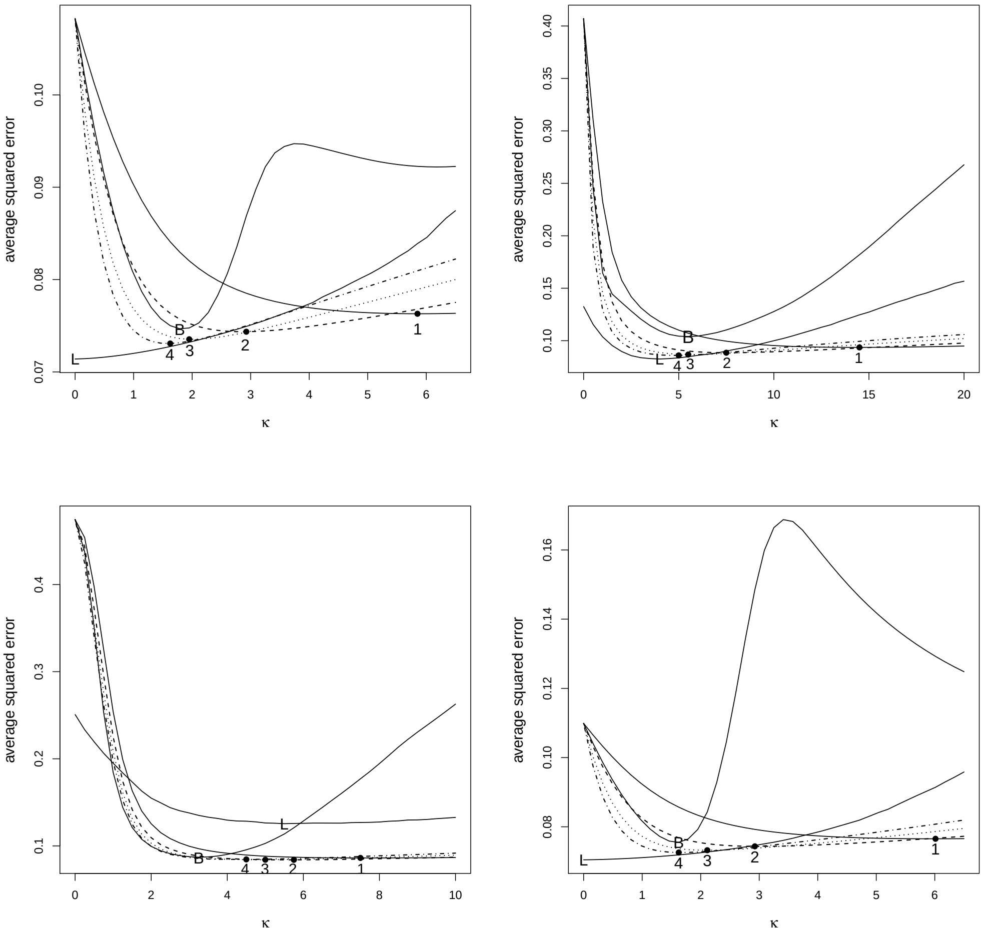

Figure 2: Average (over 50 simulations) of average squared error (E) of $N = 100$ test data, with transformations estimated (for each test data point) from $n = 100$ training data. The data are simulated from Model 1 (top left), Model 2 (top right), Model 3 (bottom left) — each with train and test $\boldsymbol{x}$ values which are uniformly i.i.d. over the sphere — and Model 1 (bottom right), with (only) the training set $\boldsymbol{x}$ restricted to $\prod_j x_j > 0$, with x_j being the jth entry of $\boldsymbol{x}$. Lines drawn show values corresponding to smoothing parameter κ ($\kappa = 0$ is the rigid solution), with local minima shown by points for iterations $M = 1, 2, 3$ and $M = 4$, together with B: weights obtained from (14), and L: two-term solution.

bias and variance change in the iterative method. As expected, with fixed κ the bias-squared is decreased with iteration, but with smaller corresponding increase in variance. Further analysis of this decomposition is made by comparing the one-term ($M = 1$ iteration) with the two-term solution. At each test data point we obtained an average of the predictions, which leads to a "bias vector", and $\sqrt{}$(variance). Specifically, for each test point $\boldsymbol{x}$ we obtained the predicted $\hat{\boldsymbol{y}}$, and then computed (over all simulations) the spherical average and variance corresponding to the known $\boldsymbol{y}$. Figure 3 also shows results of the two-term, and one-term solutions for Model 1 with training data which is uniform over the sphere, or restricted as above.

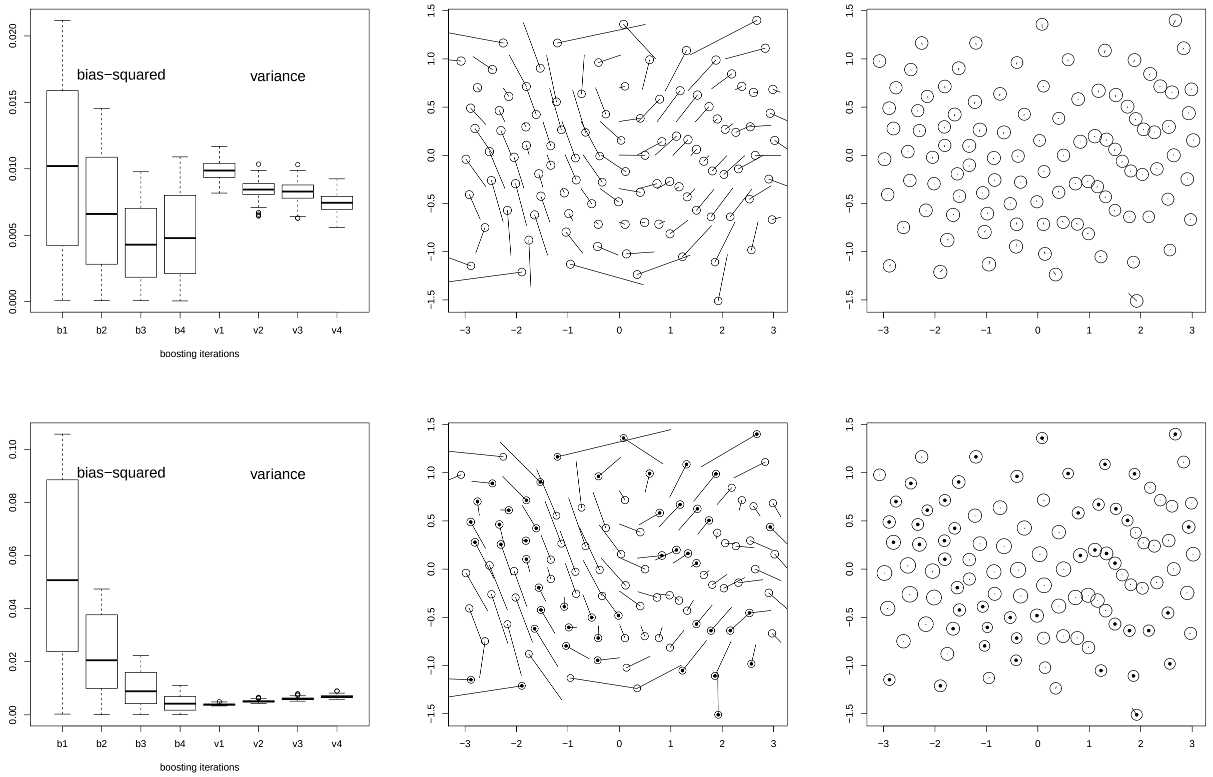

Figure 3: Over 100 datasets, with $n = 100$ training pairs from Model 1, the bias and variance for 104 equally spaced testing design points are computed. In the left column the box-plots show the effect of iteration, in which κ is chosen according to M from Figure 2 (top) and $\kappa = 1.35$ is fixed (bottom). We plot a "bias line" from the mean of the predicted values to the true $\boldsymbol{y}$, with circles which have radius equal to $\sqrt{}$(variance) for a one-term solution (middle panels) and two-term solution (right) with κ chosen according to the results in Figure 2. Top right panels have design training data uniformly distributed on the sphere, and the lower right panels have restricted design (such that $\prod x_j > 0$, with x_j being the jth entry of $\boldsymbol{x}$) in which we indicate the (transformed equally-spaced) test data associated with the design condition with filled points.

In the plot of data we have projected the sphere, so the vectors will appear longer at the poles, and no attempt has been made to project the "circles" corresponding to the standard deviations. We can see that the two-term solution has almost eliminated the bias, which is much more evident in the one-term solution, (with an appearance of a "vector-field" of errors) which, in turn, has a smaller variance.

A comparison of the lower panels with the upper panels shows that the test points which are not in the restricted training set region are not predicted so well, as expected. Comparing the average of the mean-squared error (MSE) of the test points which were "out" *vs* "in" the restricted region, showed an increase of

34%, 111% and 20% for the rigid, one-term, and two-term solutions respectively. However, for those points out of the restricted region, again the average of the MSE of the two-term solution (0.0085) was better than the one-term solution (0.0363), which, in turn, was much better than the rigid solution (0.1615). The supplementary material contains an additional simulation study on the bias of our methods.

7. EXPERIMENTS WITH REAL DATA

Using the data displayed in Figure 1, which are taken from Table 1 of Chang (1986), we consider three methods to estimate the corresponding points of the Arabian tectonic plate, using the latitude and longitude of the Somalian plate. In this experiment we use leave-one-out estimation, in which parameters are estimated using all pairs except the ith observation, $i \in (1, \ldots, n)$. For the one-term estimator, we select the smoothing parameter for each of the n datasets, using leave-one-out cross-validation amongst the $n-1$ observations. For the rigid rotation, we simply estimate $\boldsymbol{R}$ using $\boldsymbol{X}^{(-i)}$. We also consider the projective general linear (PGL) model of Rosenthal et al. (2014). The average squared errors ($E \times 10^7$) over the 11 observations are: 5.44 (rigid rotation), 3.45 (one-term solution) and 5.84 (PGL). This is a very small dataset — too small to consider the two-term solution — but this encouraging result provides a useful comparison with a previous analysis. Additional examples are contained in the accompanying supplementary material.

8. GENERALIZATION TO SHAPE MATCHING

A generalization of our method to non-spherical data is to similarly consider weights in *ordinary Procrustes analysis* (Dryden and Mardia, 2016). In the standard (non-weighted) setting, this seeks to match two sets of landmarks by a suitable rotation, translation and scaling. So, in this more general *shape* context, the conditional expectation (2) would be extended to

$$\mathsf{E}[\boldsymbol{\mathcal{Y}} \,|\, \boldsymbol{\mathcal{X}} = \boldsymbol{x}] = \rho_{\boldsymbol{x}} \boldsymbol{R}_{\boldsymbol{x}} \boldsymbol{x} + \boldsymbol{\tau}_{\boldsymbol{x}}\,,$$

where $\rho_{\boldsymbol{x}} > 0$ is a scale parameter, $\boldsymbol{\tau}_{\boldsymbol{x}}$ is a location (or translation) vector, and $\boldsymbol{R}_{\boldsymbol{x}}$ is a rotation matrix depending on $\boldsymbol{x}$. Here the data points are not i.i.d. since they are generally obtained as landmarks belonging to a single "object". Using our core hypothesis of smoothness, we could say that transformations associated to close locations are similar. Consequently, if we are given a location $\boldsymbol{x}$ which is related to the landmarks of object $\boldsymbol{X}$ we could seek a weighted (with respect to $\boldsymbol{x}$) solution giving local estimates $\hat{\boldsymbol{\tau}}_{\boldsymbol{x}}, \hat{\boldsymbol{R}}_{\boldsymbol{x}}$ and $\hat{\rho}_{\boldsymbol{x}}$. However, since shape analysis is intended to eliminate different choices of origin (which may have no meaning), it seems unnecessary to have a centring (or translation) transformation which is weighted in any way. This would lead to the estimates: $\hat{\boldsymbol{R}}_{\boldsymbol{x}} = \boldsymbol{U}^T \boldsymbol{V}$, in which $\boldsymbol{U}$ and $\boldsymbol{V}$ are obtained from the SVD of $||\boldsymbol{X}||_F^{-1} ||\boldsymbol{Y}||_F^{-1} \boldsymbol{Y}^T \boldsymbol{W}_\kappa(\boldsymbol{x}) \boldsymbol{X}$, and

$$\hat{\rho}_{\boldsymbol{x}} = \frac{\operatorname{trace}(\boldsymbol{Y}^T \boldsymbol{W}_\kappa(\boldsymbol{x}) \boldsymbol{X} \hat{\boldsymbol{R}}_{\boldsymbol{x}})}{\operatorname{trace}(\boldsymbol{X}^T \boldsymbol{W}_\kappa(\boldsymbol{x}) \boldsymbol{X})}.$$

Depending on the smoothing degree and the distance function used, these nonparametric transformations will give a continuous, smooth deformation, but may lead to folding in the case that κ is chosen too large. It is also equivariant under location, scale and rotation of the objects. The flexible solutions may be similar to the thin-plate spline deformations which have previously been used for aligning objects (Bookstein, 1989).

REFERENCES

Bookstein, F.L. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 567–585.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324–339.

Chang, T. (1986). Spherical regression. *The Annals of Statistics*, **14**, 907–924.

Chang, T. (1989). Spherical Regression with error in variables. *The Annals of Statistics*, **17**, 293–306.

Chang, T., Ko, D., Royer, J.-Y. and Lu, J. (2000). Regression techniques in plate tectonics. *Statistical Science*, **15**, 342–356.

Di Marzio, M., Panzera, A. and Taylor, C.C. (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, **141**, 2156–2173.

Di Marzio, M., Panzera, A. and Taylor, C.C. (2013). Nonparametric regression for circular responses. *Scandinavian Journal of Statistics*, **40**, 238–255.

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

Di Marzio, M. and Taylor, C.C. (2008). On boosting kernel regression. *Journal of Statistical Planning and Inference*, **138**, 2483–2498.

Downs, T.D. (1972). Orientation statistics. *Biometrika*, **59**, 665–676.

Downs, T.D. (2003). Spherical regression. *Biometrika*, **90**, 655–668.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: with applications in R, 2nd edition.* John Wiley & Sons, New York.

Gallier, J. and Xu, D. (2003). Computing exponentials of skew-symmetric matrices and logarithm of orthogonal matrices. *International Journal of Robotics and Automation*, **18**, 10–20.

Gower, J.C. (1975). Generalized procrustes analysis. *Psychometrika*, **40**, 33–51.

Hall, P., Watson, G.S. and Cabrera, J.(1987). Kernel density estimation with spherical data. *Biometrika*, **74**, 751–762.

Jupp, P.E.(1988). Residuals for directional data. *Journal of Applied Statistics*, **15**, 137–147.

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

Klemelä, J.(2000). Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis*, **73**, 18–40.

Mackenzie, J.K. (1957). The estimation of an orientation relationship. *Acta Crystallographica*, **10**, 61–62.

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

Osborne, M.R.(1992). Fisher's Method of Scoring. *International Statistical Review / Revue Internationale de Statistique*, **60**, 99–117.

Rancourt, D., Rivest, L.-P. and Asselin, J. (2000). Using orientation statistics to investigate variations in human kinematics. *The Journal of the Royal Statistical Society, Series C*, **49**, 81–94.

Rivest, L.-P. (1989). Spherical Regression for Concentrated Fisher-Von Mises Distributions. *The Annals of Statistics*, **17**, 307–317.

Rosenthal, M. and Wu, W. and Klassen, E. and Srivastava A. (2014). Spherical Regression Models Using Projective Linear Transformations. *Journal of the American Statistical Association*, **109**, 1615–1624.

Rosenthal, M. and Wu, W. and Klassen, E. and Srivastava A. (2017). Nonparametric Spherical Regression Using Diffeomorphic Mappings, https://arxiv.org/abs/1702.00823.

Ruppert, D. & Wand, M.P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics* **22**, 1346–1370.

Stuetzle, W. & Mittal, Y. (1979). Some comments on the asymptotic behavior of robust smoothers. In: Gasser T., Rosenblatt M. (eds) Smoothing Techniques for Curve Estimation. *Lecture Notes in Mathematics* **757**, Springer, Berlin, Heidelberg. 191–195.

Wahba, G. (1965). Problem 651: A Least Squares Estimate of Spacecraft Attitude. *SIAM Review*, **7**, 409.

Wang H., Liu, Q., Mok, M.K., Fu, L. and Tse, W.M. (2007). A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, **179**, 459–468.